PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO

NIELSEN LUIZ RECHIA MACHADO

UM FRAMEWORK PARA IDENTIFICAÇÃO E MONITORAMENTO DE PERFIS E COMPORTAMENTOS DE CONSUMIDORES BASEADO NO USO DE APLICATIVOS EM DISPOSITIVOS MÓVEIS

Porto Alegre
2019

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

UM *FRAMEWORK* PARA IDENTIFICAÇÃO E MONITORAMENTO DE PERFIS E COMPORTAMENTOS DE CONSUMIDORES BASEADO NO USO DE APLICATIVOS EM DISPOSITIVOS MÓVEIS.

NIELSEN LUIZ RECHIA MACHADO

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz

Porto Alegre
2019

Ficha Catalográfica

M149f Machado, Nielsen Luiz Rechia

Um Framework para Identificação e Monitoramento de Perfis e Comportamentos de Consumidores Baseado no Uso de Aplicativos em Dispositivos Móveis / Nielsen Luiz Rechia Machado . – 2019.

205.

Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz.

1. Identificação de Perfis. 2. Monitoramento de Perfis. 3. Monitoramento de Comportamentos. 4. Aprendizado de Máquina. 5. Dispositivos Móveis. I. Ruiz, Duncan Dubugras Alcoba. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Salete Maria Sartori CRB-10/1363

Nielsen Luiz Rechia Machado

UM FRAMEWORK PARA IDENTIFICAÇÃO E MONITORAMENTO DE PERFIS E COMPORTAMENTOS DE CONSUMIDORES BASEADO NO USO DE APLICATIVOS EM DISPOSITIVOS MÓVEIS

Tese/Dissertação apresentada como requisito parcial para obtenção do grau de Doutor/Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciencia da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado em 27 de Março de 2019.

BANCA EXAMINADORA:

Prof. Dra. Karin Becker (PPGC/UFRGS)

Prof. Dr. Silvio César Cazella (Programa de Pós-Graduação em Ensino da Saúde/UFCSPA)

Profa. Dr. Rodrigo Coelho Barros(PPGCC/PUCRS)

Prof. Dr. Duncan Dubugras Alcoba Ruiz (PPGCC/PUCRS - Orientador)

DEDICATÓRIA

Dedico este trabalho a minha família. Principalmente à minha esposa Lenise Thies, minha filha Nicole Thies Rechia e minha mãe Marijane Rechia.

"Comece fazendo o que é necessário, depois o que é possível, e de repente você estará fazendo o impossível."
(São Francisco de Assis)

AGRADECIMENTOS

Gostaria de registrar os meus agradecimentos a todos que me apoiaram na realização de mais uma importante etapa da minha vida.

Agradeço inicialmente a Deus por me dar saúde e força durante todos estes anos de estudo e aprendizado.

Agradeço a minha esposa Lenise Thies, por estar ao meu lado me apoiando e me incentivando em todos os momentos. Seu companheirismo e dedicação foram imprescindíveis ao longo dessa jornada.

A minha filha Nicole Thies Rechia que é um dos meus maiores motivos na busca de ser uma pessoa melhor. Ela sempre me fez sorrir e esquecer os problemas que enfrentei nesta caminhada.

A minha mãe Marijane Rechia que me educou e mostrou a ser um ser humano melhor durante este período. Ela sempre me incentivou a pensar positivamente e me mostrou que obstáculos podem ser sempre superados com amor e com tranquilidade.

Ao meu orientado Duncan Dubugras Alcoba Ruiz que me norteou ao longo do desenvolvimento da minha pesquisa. Com sua ajuda e sabedoria tive a oportunidade de executar uma grande trabalho. Com seu ensinamento, sua experiência, suas dicas, sua motivação e sua dedicação possibilitaram que tudo fosse realizado no seu devido tempo.

A todos os colegas e amigos do PPGCC e GPIN, pela oportunidade de pesquisa em conjunto, pela troca de experiências, convivência, amizade e de apoio durante este trabalho. Principalmente ao Henry Cagnini que esteve sempre me ouvindo e me aconselhando.

A todos os demais familiares e também aos amigos que sempre estiveram na torcida, me incentivaram e me deram palavras de apoio ao longo destes anos.

Ao CNPq, a Motorola e a Dell que proveram recursos financeiros para que esta pesquisa fosse concluída.

UM *FRAMEWORK* PARA IDENTIFICAÇÃO E MONITORAMENTO DE PERFIS E COMPORTAMENTOS DE CONSUMIDORES BASEADO NO USO DE APLICATIVOS EM DISPOSITIVOS MÓVEIS.

RESUMO

É possível observar um crescimento significativo no uso de dispositivos móveis, bem como na utilização de aplicativos nestes dispositivos ao longo dos último anos. Além disso, a inovação tecnológica e a disputa acirrada na conquista do mercado faz com que empresas fabricantes de tais dispositivos aumentem suas atenções para interesses de seus clientes. Estes clientes realizam diariamente muitas atividades por meio do uso de aplicativos, o que gera, em tempo real, uma grande quantidade de eventos. Diante disso, é importante para estas empresas entender como seus clientes utilizam aplicativos em seus dispositivos. Neste sentido, mecanismos automáticos capazes de ajudar na identificação e no monitoramento de perfis e comportamento de tais clientes, podem contribuir na tomada de decisões das partes interessadas. Assim, esta pesquisa propõe um *framework* para identificação e monitoramento de perfis e comportamentos de uso de aplicativos em dispositivos móveis. Para alcançar este objetivo, técnicas de Mineração de dados como, Transformação e Discretização, tarefas de Aprendizado de Máquina como, Regras de Associação e Agrupamento, e técnicas de Detecção de Novidade como, Mudança e Evolução de Conceito são utilizadas. Com o objetivo de fazer uma análise comparativa, foram avaliados abordagens relatadas na literatura, considerando para tanto, um fluxo contínuio de dados de uso de aplicativos real. Resultados da análise experimental mostram que o *framework* proposto apresenta melhores resultados ao cenário abordado apontando perfis e comportamentos que evoluem conforme o fluxo contínuo de dados.

Palavras-Chave: Identificação de Perfis, Monitoramento de Perfis, Monitoramento de Comportamentos, Aprendizado de Máquina, Aplicativos Móveis.

A FRAMEWORK FOR IDENTIFICATION AND MONITORING OF CONSUMER PROFILES AND BEHAVIORS BASED ON APPLICATION USAGE.

ABSTRACT

It is possible to observe a significant growth in the use of mobile devices as well as the use of applications on such devices over the last years. In addition, the technological innovation and fierce dispute to conquer the market make mobile device manufacturers companies increase their attention to the interests of their clients. These clients perform daily many activities through the use of applications, which generates, in real time, a large number of events. Therefore, it is important for aforementioned companies to understand how their customers use applications on their devices. In this sense, automatic mechanisms, capable of assisting in the identification and monitoring of profiles and behavior of such clients, can contribute to the decision making of the stackholders. Based on this, this study proposes a framework for the identification and monitoring of the profiles and behaviors of app usage on mobile devices. To achieve this goal, Data Mining techniques such as Transformation and Discretization, Machine Learning tasks such as Association Rules and Clustering, and Novelty Detection techniques such as Concept Drift and Concept Evolution, are used to explore the app usage, identify app usage patterns, pinpoint profiles, and monitor customer behaviors over time. In order to make a comparative analysis, we have evaluated the approaches adopted by the literature, considering a real app usage data stream. Results of the experimental analysis show that the proposed framework presents better results to the addressed scenario pointing to profiles and behaviors that evolve according to the data stream.

Keywords: Profile Identification, Profile Monitoring, Behavior Monitoring, Machine Learning, Mobile Apps.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS

GPS. – Sistema de Posicionamento Global - *Global Positioning System*

FCD. – Fluxo Contínuo de Dados - *Data Stream*

CDR. – Registro de Chamadas - *Call Data Records*

IMEI. – Identificação Internacional de Equipamento Móvel - *International Mobile Station Equipment Identity*

Pkg. – Pacote - *Package*

CPF. – Cadastro de Pessoas Físicas

RG. – Registro Geral

IP. – Particionamento Intuitivo - *Intuitive Partitioning*

LHS. – Lado Antecedente - *Left Hand Side*

RHS. – Lado Consequênte - *Right Hand Side*

SWC. – Largura da Silhueta - *Silhouette Width Criterion*

DBI. – *Davis-Bouldin Index*

DUNN. – *Dunn Index*

CH. – *Calinski-H Criterion*

SSE. – Soma do Erro Quadrático - *Sum of Squared Error*

GAP. – *Gap Statistic*

PDA. – Assistente Pessoal Digital - *Personal Digital Assistent*

FIFO. – *First In First Out*

StArt. – *State of the Art through Systematic Review*

SMS. – Serviço de Mensagem Curta - *Short Message Service*

EM. – *Expectation Maximization*

BIC. – *Bayesian Information Criterion*

MOA. – *Massive Online Analysis*

MONIC. – *Modeling and Monitoring Cluster Transitions*

MEC. – *Monitoring the Evolution of Clusters*

TF–IDF. – *Term Frequency–Inverse Document Frequency*

f-DOPE. – *Framework for iDentification and mOnitoring of Profiles and bEhaviors*

IWCMC. – *International Wireless Communications and Mobile Computing Conference*

SBBD. – Brazilian Symposium on Databases

WTDBD. – Thesis and Dissertations Workshop

CIKM. – International Conference on Information and Knowledge Management

ARI. – Artificial Intelligence Review

TMC. – Transactions on Mobile Computing

SO. – Sistema Operacional - *Operating System*

API. – Interface de Programação de Aplicação - *Application program interface*

IIQ. – Intervalo Inter Quartil - *Inter Quartile Range*

Q3. – Quartil Superior - *Upper Quartile*

Q1. – Quartil Inferior - *Lower Quartile*

LS. – Limite Superior - *Upper Whisker*

LI. – Limite Inferior - *Lower Whisker*

LISTA DE SÍMBOLOS

SUMÁRIO

1. INTRODUÇÃO

Dispositivos móveis sem fio como relógios inteligentes (Mohammad et al., 2017) e *smartphones* (Do et al., 2011) podem ser considerados os principais fatores da inclusão de plataformas de computação e comunicação na vida das pessoas. Tais dispositivos estão equipados com um número crescente de sensores, como acelerômetro, compasso digital, giroscópio e GPS (Sistema de Posicionamento Global - *Global Positioning System*). Estes sensores, os quais são amplamente utilizados para diferentes propósitos em computação móvel e ubíqua, são capazes de obter dados em larga escala de diferentes formatos (Fan e Bifet, 2013). Ao longo dos últimos anos, estes dispositivos evoluíram de um meio de comunicação simples para ferramentas dinâmicas que ajudam consumidores em suas diferentes atividades diárias (Hamka et al., 2014).

Uma das principais ferramentas dos dispositivos móveis são os aplicativos. Aplicativos são programas desenvolvidos para que consumidores realizem atividades diárias específicas. Neste caso, aplicativos podem ser abertos, utilizados e fechados por diversas vezes em um único dia. Tais atividades englobam, por exemplo, chamadas por vídeo, ouvir música, assistir filmes e escrever e-mails ou relatórios (Blondel et al., 2015). Em 2015, aproximadamente 4.4 bilhões de pessoas possuíam dispositivos móveis em todo mundo (Atlas, 2016). De acordo com Portal (2016, 2019), nas principais lojas de aplicativos *online*, chamadas de *marketplaces*, como *Google Play Store*[1] e *Apple store iOS*[2], o número de aplicativos disponíveis entre 2011 e 2018 cresceu cinco vezes (ex: 675.000 para 3.300.000).

Uma pessoa, ao utilizar um aplicativo em seu dispositivo móvel, faz com que um evento seja gerado com diferentes informações, como o nome do aplicativo, o tempo total de utilização e dados gerados por sensores. Dessa forma, muitos eventos são gerados dadas as atividades realizadas por uma pessoa em seu dispositivo no seu dia a dia. Assim, é possível capturar algumas das informações que são geradas, as quais podem ser exploradas de diferentes formas, possibilitando a descoberta de padrões sobre tais atividades (Böhmer et al., 2011). Nesse cenário, para várias partes interessadas, é importante entender como os aplicativos são utilizados e como os comportamentos de uso de tais aplicativos mudam com o tempo (Xu et al., 2011).

O conjunto de eventos de utilização de aplicativos emitidos por dispositivos móveis pode ser classificado como um FCD (Fluxo Contínuo de Dados - *Data Stream*). Gama (2010) define um FCD como um processo estocástico em que eventos ocorrem continuamente e independentemente uns dos outros. Este tipo de dado geralmente possui um grande Volume (ex: *Big Data*), uma grande Variedade (ex: diferentes características) e é gerado em alta Velocidade (ex: gerado em tempo real). Além disso, em cenários de FCDs, os dados evoluem ao longo do tempo, sendo necessário investigar mudanças na distribuição destes dados, o que é chamado de Mudança de Conceito, do inglês *Concept Drift*. Em cenários estacionários (*batch*), não é necessária tal investigação, visto que os dados disponíveis são conhecidos a priori e não existe uma evolução temporal dos mesmos. Contudo,

[1] Maior loja de aplicativos para dispositivos *Android* - https://goo.gl/WpO8h2
[2] Loja de aplicativos para dispositivos *iOS* - https://goo.gl/po7mvG

em FCD, novos conceitos podem surgir e conceitos conhecidos podem evoluir ou desaparecer. Tais variações podem ser descobertas com a aplicação da tarefa de Detecção de Novidade (Gama, 2010). Diante disso, uma ferramenta eficaz, capaz de realizar análises de FCD de uso de aplicativos em dispositivos móveis auxiliando pesquisadores e empresas a extrair conhecimento sobre tais FCD, é quase que obrigatória.

1.1 Fidelização de Clientes

Os avanços na indústria de dispositivos móveis, como novos serviços e tecnologias, bem como os avanços nas áreas de Mineração de Dados e FCD, ampliaram a competição deste mercado. Empresas como fabricantes de tais dispositivos e operadoras de telecomunicação (telecom) visam manter seus clientes leais e engajados. Tal intenção está relacionada ao custo para atrair novos clientes, que chega a ser seis vezes maior do que o valor gasto na manutenção de clientes existentes. Como resultado, estas empresas estão sobre intensa pressão para identificar e monitorar perfis e comportamentos de seus clientes. Como estes clientes são a principal fonte de recursos para tais empresas e o mercado encontra-se saturado, novos mecanismos para o gerenciamento destes consumidores são vitais para a sobrevivência e o desenvolvimento de tais empresas (Almana et al., 2014). Neste sentido, empresas fabricantes de dispositivos móveis buscam obter regras/padrões de uso de aplicativos e um número limitado de perfis com o objetivo de entender as diferentes formas de utilização de aplicativos em tais dispositivos (Fan e Bifet, 2013).

Dado o atual crescimento no número de clientes, assim como na quantidade de aplicativos o monitoramento deste cenário visa o entendimento da variação dos perfis existentes bem como dos comportamentos de consumidores, os quais podem indicar diferentes possibilidades aos negócios. Por exemplo, perfis específicos podem representar situações de tendência no mercado. Neste caso, como os perfis representam padrões das atividades mais realizadas por centenas/milhares de consumidores, é possível entender as necessidades de tais clientes em relação aos dispositivo utilizados. Por outro lado, existem situações de risco, como clientes que abandonam um determinado produto/serviço. Tal situação é oriunda de um comportamento produzido por um grupo infrequente de clientes. Contudo, este tipo de situação difere de situações apresentadas por consumidores com comportamentos anômalos (Wei e Chiu, 2002). Wei e Chiu (2002) afirmam que uma taxa de rotatividade é considerada alta, quando é identificado o abandono de mais de $1,5\%$ dos consumidores de uma determinada marca (período mensal). No entanto, as empresas fabricantes de dispositivos móveis, ao contrário das operadoras de telecom, desconhecem exatamente quando ocorre o abandono de seus consumidores. Um dos motivos é a ausência de contratos entre os consumidores e as fabricantes dos dispositivos. Por exemplo, a empresa patrocinadora desta pesquisa desconhece a quantidade exata de usuários que deixam de usar seus dispositivos móveis. Sendo assim, tanto para pesquisadores quanto para empresas identificar e monitorar perfis e comportamentos de consumidores são tarefas importantes e complexas.

1.2 Identificação e Monitoramento de Perfis e Comportamentos

Ainda é difícil ter acesso a conjuntos com dados de uso de dispositivos móveis, principalmente de dados de uso de aplicativos. Mesmo com o atual crescimento, tanto de pesquisas e da inovação na indústria de dispositivos móveis, os primeiros conjuntos de dados deste tipo foram apresentados somente nos últimos anos (Wagner et al., 2013; de Montjoye et al., 2014; Blondel et al., 2012). Entretanto, tais conjuntos ainda não possuem informações persistentes sobre a utilização de aplicativos. Por exemplo, alguns dispositivos podem ou não enviar eventos de uso de aplicativos de acordo com sua configuração de privacidade. Além disso, a anonimização realizada não permite o mapeamento dos eventos entre dispositivos. Mais ainda, por serem considerados os principais conjuntos de dados para extração de informações de consumidores, estes dados são geralmente privados ou não estão disponíveis devido a regulamentações que preservam a privacidade dos clientes e/ou organizações (Blondel et al., 2015).

Nos últimos anos, estudos baseados na utilização de dispositivos móveis foram elaborados para diferentes fins. Alguns trabalhos visaram a identificação e a análise dos aplicativos mais utilizados em diferentes contextos (Verkasalo, 2009; Xu et al., 2011; Li et al., 2015). Contudo, tais trabalhos não buscam a identificação de perfis de uso e, consequentemente, não realizam o monitoramento de perfis e de comportamentos dos consumidores ao longo do tempo. Outros trabalhos buscaram a predição de clientes com potencial chance de abandonar uma marca (Wei e Chiu, 2002; Chu et al., 2007; Zhang, 2007; Dasgupta et al., 2008; Sohn e Kim, 2008; Li e Deng, 2012; Bahmani et al., 2013; Ballea et al., 2013; Meireles, 2014; Rehman e Raza Ali, 2015; Backiel et al., 2016). Entretanto, estes trabalhos consideram tal predição, em sua maioria, como um problema de Classificação e utilizam conjuntos de dados estacionários considerando os objetos como indivíduos independentes. Desta forma, perfis não são buscados e o monitoramento de perfis também não é aplicado. Além disso, tais trabalhos utilizam diferentes tipo de dados, como CDR (Registros de Dados de Ligações - *Call Data Records*) que possuem informações de ligações realizadas pelos consumidores, e informações de faturamento ou pessoais (ex: pagamentos, sexo, idade), sem abordar dados de uso de aplicativos. Por outro lado, alguns estudos propuseram a identificação de perfis, sendo tal tarefa realizada em diferentes áreas de aplicação, como por exemplo, na Alimentação (Hajiha et al., 2011), na Comunicação (Pyo et al., 2015) e em Bibliotecas (Hsu et al., 2012). A identificação de perfis também é pesquisada em áreas que investigam dispositivos móveis. Dentre tais pesquisas, algumas não utilizaram dados de uso de aplicativos (Sohn e Kim, 2008; Rehman e Raza Ali, 2015), enquanto outra complementou seu conjunto de dados com com este tipo de informação (Hamka et al., 2014). Contudo, Hamka et al. (2014) investigaram somente a quantidade de aplicativos utilizados pelos consumidores, sem levar outras informações, como tempo de uso dos aplicativos, em consideração.

Na última década, algumas abordagens foram propostas em cenários de FCDs com o objetivo de identificar e acompanhar mudanças em conceitos aprendidos (Spiliopoulou et al., 2006; Oliveira e Gama, 2010c; Ntoutsi et al., 2012). Estes trabalhos visam a identificação e o moni-

toramento de perfis ao longo do tempo e também foram abordados em diferentes áreas, como Economia (Oliveira e Gama, 2010c; Siddiqui et al., 2012) e Medicina (Siddiqui et al., 2015). Entretanto, tais abordagens destinam-se somente a analisar as variações que ocorrem ao longo do tempo entre os diferentes conceitos detectados e não investigam os comportamentos dos objetos pertencente a tais perfis. Neste cenário, um único trabalho buscou o monitoramento de perfis de uso em FCDs de dispositivos móveis (Pereira e Mendes-Moreira, 2016). Todavia, tal trabalho não utilizou dados de uso de aplicativos, somente investigando *CDR*.

1.3 Caracterização do Problema

A empresa patrocinadora desta pesquisa, uma fabricante multinacional de dispositivos móveis, monitorou, ao longo de 2017, cerca de 60 milhões de *smartphones*. Tais dispositivos produziram, a cada dia, aproximadamente 1 bilhão de eventos (atividades) de uso de aplicativos. Estes eventos foram gerados pelo uso de mais de $1,5$ milhão de aplicativos distintos. Logo, é possível observar que a quantidade de consumidores, bem como a quantidade de aplicativos para este tipo de dispositivo é muito alta. Além disso, estes números vêm crescendo aceleradamente causando uma grande disputa entre fabricantes de dispositivos móveis na manutenção e conquista de clientes.

De acordo com Rehman e Raza Ali (2015) a persistência de dados de dispositivos móveis é um grande desafio. Muitas empresas mantém somente uma porção representativa dos dados capturados por um longo período de tempo, respeitando a privacidade dos dados e as preferências dos clientes em relação ao tipo de informação a ser armazenada. Este tipo de imposição é necessária dado os bilhões de eventos gerados que representam centenas de *terabytes* que devem ser armazenados em memória (Fan e Bifet, 2013). Ao longo desta pesquisa, o FCD de uso de aplicativos permaneceu armazenado em sua totalidade por um período de três meses. Apenas 1% deste FCD foi mantido salvo por um período máximo de seis meses em um banco de dados em nuvem[3].

Apesar do FCD de uso de aplicativos possuir um enorme conjunto de eventos rico em informações, tais eventos precisam ser pré-processados pois não estão em um formato adequado para executar por completo o estudo desenvolvido nesta pequisa. Em um cenário real, o qual é exemplificado pela Figura 1.1, consumidores realizam diversas atividades por meio do uso de aplicativos em dispositivos móveis. Na Figura 1.1 (a) é demonstrado o uso de aplicativos por três diferentes clientes (dispositivos) ao longo de um determinado tempo. Para cada dispositivo (1, 2 e 3), cada ícone (aplicativo) representa uma atividade realizada pelo cliente por meio do uso de tal aplicativo, o qual deve estar em *foreground*. Assim, um FCD contém bilhões de atividades, as quais são executadas em milhões de dispositivos por meio de um dos milhões de aplicativos disponíveis. Além disso, é importante notar que uma determinada atividade, por exemplo, com o uso do aplicativo *WhatsApp*, pode ser realizada repetidamente ao longo do tempo. Mais ainda, tal

[3]https://cloud.google.com/bigquery/

atividade pode ser executada por muitos consumidores, bem como por diferentes quantidades de tempo (ex: 30 segundos, 2 minutos, etc).

Figura 1.1: Representação do FCD de uso de aplicativos em dispositivos móveis. Diferentes consumidores realizam diferentes atividades (a). As atividades são capturadas em Janelas de Eventos (b).

Em geral, as atividades realizadas por um proprietário de dispositivo móvel tendem a se repetir conforme a rotina de tal consumidor. Neste caso, a mudança no uso de aplicativos não é frequente indicando que consumidores tendem a ter um mesmo costume no seu dia-a-dia. Esta rotina normalmente é alterada quando alguma mudança, seja no dispositivo, nos hábitos, ou lançamento/suspensão de aplicativos, transforme significativamente as atividades realizadas por tais clientes em seus dispositivos.

Uma explicação detalhada de cada atributo que compõe um evento do FCD de uso de aplicativos é apresentado a seguir. Formalmente, define-se um evento ε como uma tupla (i, p, et, d):

- ***IMEI*** (Identificação Internacional de Equipamento Móvel - *International Mobile Station Equipment Identity*) - o número de série do dispositivo anonimizado que identifica um único dispositivo - (i).
- ***pkg*** (Pacote - *Package*) - o nome dado ao aplicativo utilizado em uma atividade - (p).
- ***end_timestamp*** - o tempo final em que o aplicativo foi fechado ou colocado em *background* pelo consumidor[4] - (et).

[4]Aplicativos abertos que estão em *background* não geram eventos pois não estão sendo efetivamente utilizados pelos consumidores.

- ***duration_sec*** - o tempo em segundos em que o aplicativo foi utilizado pelo consumidor - (d).

É importante citar que o FCD de uso de aplicativos não possui informações dos consumidores proprietários de cada dispositivo, como por exemplo, CPF (número do Cadastro de Pessoas Físicas do Brasil) ou RG (número do Registro Geral da Carteira de Identidade do Brasil). Entretanto, assume-se que cada consumidor possui um único dispositivo, sendo este utilizado em grande parte do tempo pelo seu proprietário. Além disso, os IMEIs são anonimizados para proteger a privacidade dos consumidores. Contudo, o fato de ser realizada a anonimização dos IMEIs não afeta a utilidade deste estudo. Ademais, não é possível a reversão da anonimização para a re-identificação do dispositivos.

Todas as atividades podem ou não ser realizadas em uma mesma Janela de Eventos (ver Seção 3.2.2) (ex: dia, semana ou mês), como mostra a Figura 1.1 (b). As Janelas de Eventos são a maneira mais simples de manter uma quantidade possível de dados na memória física. Tais janelas podem ajudar na transição entre dados do passado próximo e dados de um passado distante (Gama, 2010). De acordo com o tipo de janela utilizada (Babcock et al., 2002b), mais ou menos eventos são capturados, para serem pré-processados e/ou resumidos. Devido ao grande número de eventos gerados e a pequena quantidade de tempo de disponibilidade do FCD é necessário definir um período de tempo (ex: dia, semana, mês, etc) ou uma quantidade de eventos (ex: 1 milhão, 1 bilhão) permitindo o armazenamento e a análise de parte do FCD. Além disso, é necessário definir um número de janelas que permita a identificação e o monitoramento de perfis e comportamentos de consumidores. Mais ainda, é necessário investigar, em cada janela, quão importantes são estes aplicativos e como tais aplicativos são utilizados (Li et al., 2015). Uma vez que a geração de padrões de uso de aplicativos e a descoberta de perfis podem ser baseadas no tempo de utilização destes aplicativos, é necessário investigá-los visando identificar suas relevâncias. Dessa forma, fica evidente que cada evento não pode ser considerado como uma representação completa de um objeto independente (dispositivo), o que é comumente realizado por muitos algoritmos tradicionais de FCDs (Gama, 2010).

1.4 Motivação

Dispositivos móveis são considerados os principais equipamentos para a aquisição de dados sobre consumidores. Estes dados são gerados em tempo real e em diferentes aspectos. Neste sentido, a quantidade de dados que empresas podem potencialmente manipular, aumentou gradualmente e de forma significativa, excedendo a quantidade de dados do nosso passado recente. Em um passado não tão distante, existiam apenas registros de chamadas e processamentos que eram utilizados para fins de faturamento, ao contrário dos dados que são atualmente possíveis de se obter. Nesta direção, FCD de dispositivos móveis possuem muitas propriedades, visto a quantidade de inovações tecnológicas aplicadas a este tipo de equipamento. Com a evolução de tais dispositivos, bem como

dos aplicativos para estes dispositivos, em conjunto com o aumento no número de consumidores, é possível capturar grandes e complexos FCDs que devem ser processados e analisados visando diferentes resultados (Fan e Bifet, 2013).

Neste cenário, para diversas partes interessadas, é importante entender como estes aplicativos são utilizados pelos consumidores e como os comportamentos destes consumidores, relacionados ao uso de tais aplicativos, mudam ao longo do tempo (Xu et al., 2011). Neste contexto, fabricantes de dispositivos móveis e operadoras de telecom desejam identificar aplicativos populares ou problemáticos para fornecer, por exemplo, sistemas de recomendação de aplicativos (Böhmer et al., 2011). Por outro lado, desenvolvedores de aplicativos, visam entender por que seus aplicativos estão ou não sendo apreciados pelos consumidores, o que pode ajudar na melhoria do design de tais aplicativos. Consumidores finais, por sua vez, desejam obter um conhecimento abrangente de como seus aplicativos consomem, por exemplo, a bateria e o tráfego de dados visando otimizar suas decisões na hora de definir quais aplicativos devem ser escolhidos. Desta forma, dada a falta de uma análise considerável em FCD de uso de aplicativos em dispositivos móveis, os quais podem refletir perfis e comportamentos de clientes, a maioria destas importantes questões permanecem sem resposta (Li et al., 2015).

1.5 Objetivo

Esta tese visa responder a seguinte questão de pesquisa – *Dado um fluxo contínuo de dados de uso de aplicativos em dispositivos móveis por consumidores, como segmentá-los, identificando e monitorando perfis e comportamentos?*

1.5.1 Objetivo Principal

Dado os problemas elencados anteriormente e as lacunas existentes nos trabalhos relacionados, esta pesquisa tem como objetivo principal:

Desenvolver um framework para identificação e monitoramento de perfis e comportamentos de uso de aplicativos em dispositivos móveis por consumidores.

Com este objetivo visa-se a segmentação de tais consumidores em perfis de uso, tendo em vista a identificação e o monitoramento de diferentes comportamentos ao longo do tempo. Tal objetivo foi consolidado com base no desenvolvimento de um *framework*, chamado *f-DOPE*, onde eventos reais de uso de aplicativos em dispositivos móveis são explorados por técnicas de Aprendizado de Máquina não supervisionadas (conforme Capítulo 2) e técnicas aplicadas a FCD (conforme Capítulo 3). É possível obter padrões de uso de aplicativos por meio da tarefa de Regras de Associação (conforme Seção 2.2.1), que por sua vez servem como base para a identificação de perfis de uso por meio da tarefa de Agrupamento (conforme Seção 2.2.1), sendo este processo realizado ao longo de várias janelas de eventos (conforme Seção 3.2.2). Os perfis identificados são analisados

ao longo de tais janelas por técnicas de Detecção de Mudança e Evolução de Conceitos (conforme Seção 3.3.2), utilizados para monitorar e detectar diferentes comportamentos de consumidores ao longo do tempo e reagindo de forma efetiva às mudanças em FCD.

Como decorrência, é demonstrada a qualidade dos resultados obtidos com a aplicação do *f-DOPE* em FCD reais, onde seus resultados são comparados com resultados oriundos de uma abordagem da literatura que mais se assemelha ao *f-DOPE*, visando avaliar seus resultados no referido cenário.

1.5.2 Objetivos Específicos

Além do objetivo principal, os objetivos específicos desta tese são:

- A definição de um procedimento cíclico e contínuo de coleta, análise e pré-processamento de eventos de uso de aplicativos em dispositivos móveis;
- A elaboração de um algoritmo capaz de detectar padrões na utilização de aplicativos por meio de associações entre diferentes formas de uso dos vários aplicativos existentes;
- O desenvolvimento de um algoritmo para identificar e caracterizar um conjunto limitado de perfis de uso com base nos padrões de utilização de aplicativos em dispositivos móveis;
- O desenvolvimento de um algoritmo capaz de monitorar e analisar os perfis de uso visando a detecção de mudanças e evoluções de conceitos que podem ocorrer ao longo do FCD, assim como capaz de monitorar comportamentos de consumidores baseado em tais evoluções;
- Criação de um *framework* que agregue todo o conhecimento do problema, aprendido em diferentes etapas, aplicando metodologias de avaliação em tais etapas visando medir sua qualidade em comparação à outras abordagens existentes.

1.6 Desenho da Pesquisa

Esta tese foi desenvolvida com base em pesquisas de métodos para a concepção do objetivo buscado. Inicialmente, foi realizada uma revisão da literatura sobre o processo identificação e monitoramento de perfis e comportamentos, visando selecionar etapas a serem atendidas por tal tese. Após, foi inciado um levantamento do estado da arte das abordagens computacionais (algoritmos, métodos, técnicas, etc), e posteriormente foram realizadas implementações e testes de cada abordagem. Algumas abordagens fora utilizadas como base para uma primeira versão do *framework*. Em continuidade, foi realizada uma revisão sistemática da literatura, visando identificar os trabalhos relacionados a esta tese. Tais trabalhos foram identificados e seus pontos fortes e fracos foram destacados, bem como alguns foram selecionados para a avaliação final do *framework*, como forma

de comparação. Posteriormente, foram selecionados diferentes conjuntos de dados representativos de FCD de uso de aplicativos em dispositivos móveis. Estes FCD foram testados e validados visando experimentos para o aperfeiçoamento dos algoritmos do *framework*. Assim, estudos de casos foram realizados, onde o *framework* foi implementado com diferentes abordagens pesquisadas e aplicado nos FCD visando comparar tais abordagens em busca dos melhores resultados para o cenário abordado. Por fim, foi realizado uma análise comparativa entre a versão final do *framework* e os trabalhos relacionados selecionados, os quais foram identificados pela revisão sistemática da literatura.

1.7 Organização

Esta tese é constituída por sete Capítulos que estão organizados da seguinte maneira:

- O Capítulo 2 introduz e expõe técnicas de Mineração de Dados, como pré-processamento de dados, transformação de dados e discretização. Além disso, apresenta tarefas de Aprendizado de Máquina, com foco em tarefa de Aprendizado de Máquina não supervisionado que são utilizadas no *framework* desenvolvido.
- No Capítulo 3 são apresentados princípios de FCD considerados indispensáveis para o entendimento do cenário abordado por esta pesquisa. Também são apresentadas técnicas voltadas à FCD que são aplicadas no *framework* desenvolvido.
- O Capítulo 4 apresenta e discute os trabalhos relacionados a esta proposta, detalhando a revisão sistemática da literatura realizada em busca de trabalhos sobre identificação e monitoramento de perfis.
- No Capítulo 5 é apresentado e discutido o *framework f-DOPE* que foi desenvolvido durante esta pesquisa, incluindo uma visão esquemática de tal *framework* e a descrição das etapas e fases elaboradas para o mesmo.
- O Capitulo 6 apresenta o conjunto de dados e os resultados obtidos por meio de experimentos realizados visando a validação estatística do presente trabalho.
- No Capítulo 7 são apresentadas as considerações finais da tese, bem como as contribuições, as limitações, as direções para trabalhos futuros e a descrição de publicações referente a presente pesquisa.

2. MINERAÇÃO DE DADOS

Mineração de Dados é um campo que envolve diferentes áreas, tais como: Estatística, Aprendizado de Máquina, Inteligência Artificial e Banco de Dados (Tan et al., 2006). É possível considerar a Mineração de Dados como um processo onde são realizadas a exploração e a análise de conjunto de dados, por meios automáticos ou semi-automáticos, visando favorecer a obtenção de padrões em tais dados (Han et al., 2011). Desta forma, a Mineração de Dados possui dois principais objetivos: (i) explicar o passado e (ii) prever o futuro (Tan et al., 2006; Han et al., 2011).

Ambos objetivos, quando alcançados, se tornam valiosos. Principalmente para empresas que vêm armazenando grandes quantidades de dados, como por exemplo, fabricantes de dispositivos móveis e operadoras de telecom. A partir do objetivo alcançado, tais empresas buscam utilizar o conhecimento obtido para diferentes propósitos como, por exemplo, melhorar seus serviços, fidelizar seus clientes e aprimorar suas práticas de vendas. Portanto, é possível dizer que a Mineração de Dados é um processo que permite encontrar e descrever padrões em conjunto de dados, com o uso de aprendizado, visando explicar o comportamento de tais dados e permitindo a realização de predições (Witten e Frank, 2011).

No presente capítulo são apresentadas: técnicas de Mineração de Dados (ver Seção 2.1), que buscam pré-processar dados e são utilizadas em diversas áreas de aplicação, e também, tarefas de Aprendizado de Máquina (ver Seção 2.2), com enfoque em tarefas de aprendizado não supervisionado, as quais visam extrair conhecimento de tais dados. Por fim são realizadas as considerações finais do Capítulo (ver Seção 2.3).

2.1 Pré-processamento de Dados

Nos últimos anos, grandes empresas nascionais e multinacionais, vêm armazenando enormes quantidades de dados (por exemplo, centenas de *Terabytes* ou mais). Estes grande quantidade de dados, os quais são capturados de uma ou mais fontes distintas, é chamada de *Big Data* (Jagadish, 2015). *Big Data* é fortemente suscetível a ruídos, dados ausentes e dados inconsistentes (dados esparsos), o que motiva a necessidade de pré-processamento dos dados. Atualmente, existem diversas técnicas de pré-processamento, as quais são aplicadas antes do processo de aprendizado, que visam melhorar a qualidade dos dados capturados. Tais técnicas podem ser utilizadas separadamente ou em conjunto, dependendo da aplicação. Nesta pesquisa, foram utilizadas algumas das técnicas de Transformação de Dados e Discretização de Atributos, as quais são descritas nas próximas seções.

2.1.1 Transformação de Dados

Visando melhorar o resultado da Mineração de Dados, muita vezes é necessário que os dados originais sejam transformados ou consolidados em diferentes formatos. Esta transformação busca exibir as propriedades comuns contidas em tais dados. Neste sentido, a transformação de dados pode envolver diferentes técnicas (Tan et al., 2006), como por exemplo:

- *Normalização*: faz com que os dados originais de um atributo sejam individualmente reescalados em um pequeno intervalo especificado de valores ([0, 1] ou [−1, +1]). A *Normalização*, também conhecida como *min-max*, ocorre de acordo com a Equação 2.1, para cada valor x de um atributo. Sendo min_x o valor mínimo e max_x o valor máximo de tal atributo. Desta forma, os valores originais são reescalados entre 0 e 1.

$$x' = \frac{(x - min_x)}{(max_x - min_x)} \tag{2.1}$$

- *Padronização*: permite que os dados originais de um atributo sejam individualmente padronizados com base na média e no desvio-padrão de tal atributo. A *Padronização*, também chamada de *z-score*, é realizada de acordo com a Equação 2.2, para cada valor x de um atributo, sendo μ_x a média dos valores deste atributo e σ_x o desvio padrão de tal atributo.

$$x' = \frac{(x - \mu_x)}{(\sigma_x)} \tag{2.2}$$

- *Transformação Logarítmica*: faz com que os dados originais de um atributo sejam individualmente transformados por meio da aplicação de logaritmo. Este tipo de transformação visa tornar distribuições altamente distorcidas, menos distorcidas. Esta transformação ocorre de acordo com a Equação 2.3, para cada valores x de um atributo.

$$x' = \log(x) \tag{2.3}$$

Outras técnicas, como Agregação e Generalização, também são consideradas técnicas de transformação dos dados. Em geral, qualquer função decrescente pode ser empregada para transformação de atributos em diferentes escalas (Han et al., 2011).

2.1.2 Discretização de Dados

Em Mineração de Dados, muitas vezes é necessário transformar um atributo contínuo em um atributo categórico (Tan et al., 2006). Alguns algoritmos de Mineração de Dados, especialmente

aqueles para a tarefa de Mineração de Regras de Associação (ver Seção 2.2.1), requerem dados categóricos. Em geral, os resultados obtidos por meio de indução de regras, utilizando atributos discretos, são geralmente mais compactos, curtos e precisos do que com uso de atributos contínuos (Tan et al., 2006). Além disso, por vezes é fundamental uma redução do número de valores presentes em um atributo contínuo. Neste sentido, técnicas de Discretização podem ser aplicadas tanto na redução quanto na transformação de atributos contínuos. Tal transformação ocorre por meio da divisão dos valores presentes em um atributo em diferentes intervalos. Como resultado, ao invés de utilizar os valores originais de tal atributo, são utilizados os rótulos dos intervalos definidos.

Atributos categóricos estão mais perto de uma representação de nível de conhecimento do que atributos contínuos. Para ambos, usuários e especialistas, características discretas são mais fáceis de compreender, usar e explicar. Dessa forma, o sucesso da discretização pode prolongar significativamente as fronteiras de muitos algoritmos de Aprendizado de Máquina, pois torna o aprendizado mais preciso e rápido (Liu et al., 2002). Contudo, a melhor discretização geralmente depende do algoritmo e também dos atributos a serem transformados. Em resumo, a Discretização envolve duas etapas (Tan et al., 2006):

- *Etapa 1*: Decidir quantas categorias deseja-se obter;
- *Etapa 2*: Determinar como mapear os valores dos atributos contínuos para tais categorias.

Na primeira etapa, depois de serem ordenados, os valores contínuos do atributo são divididos em diferentes intervalos. Na segunda etapa, todos os valores em um intervalo são mapeados para o mesmo valor categórico. Entretanto, decidir quantos intervalos devem ser escolhidos e como mapear os valores para tais intervalos é um problema da discretização (Han et al., 2011). Em geral, os dados podem ser supervisionados ou não supervisionados dependendo se tais dados possuem (supervisionado) ou não (não supervisionado) informações de classe. Dessa forma, a discretização supervisionada considera a informação da classe, enquanto a discretização não supervisionada não considera tal informação (Tan et al., 2006). Nesta pesquisa o conjunto de classes é desconhecido. Com isso, somente técnicas de discretização não supervisionadas são investigadas.

Algumas técnicas de discretização que são utilizadas quando a informação de classe não existe, são *largura igual* e *frequência igual*. A técnica de largura igual, divide a faixa de valores do atributo em um número de intervalos especificado, onde cada intervalo contém o mesmo tamanho (mesma faixa de valores). Entretanto, esta abordagem é sensível a valores discrepantes (*outliers*). Neste caso, é possível utilizar a abordagem de frequência igual, a qual visa colocar o mesmo número de valores em cada intervalo escolhido. Além destas técnicas, também é possível aplicar algoritmos de tarefa de Agrupamento (Ver Seção 2.2.1). Neste caso, o algoritmo *K-Means* pode ser aplicado em busca de grupos que representem os intervalos (Tan et al., 2006). Na Figura 2.1 é possível observar a aplicação de técnicas de Discretização em um conjunto de valores de um atributo exemplo, os quais pertencem a 4 diferentes classes. Na Figura 2.1 (*a*) estão os valores originais do atributo, sem discretização. O eixo horizontal do gráfico indica os valores deste atributo, os quais devem

ser discretizados. Os valores referente ao eixo vertical foram escolhidos randomicamente buscando melhorar a visualização dos valores de tal atributo.

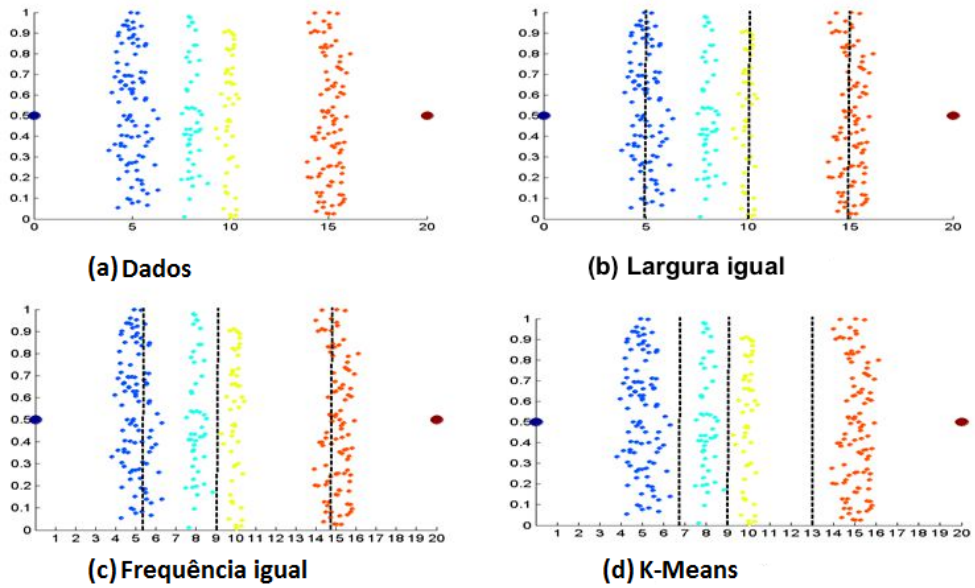


Figura 2.1: Diferentes técnicas de discretização não supervisionada. Adaptado de Tan et al. (2006).

O número de intervalos escolhido para a discretização é 4 e estão representados pelas linhas pontilhadas em preto nas na Figura 2.1 (b), (c) e (d). Assim, a Figura 2.1 (b) mostra a aplicação da discretização por *largura igual*, enquanto a Figura 2.1 (c) mostra a utilização da técnica de *frequência igual*. Por último, a Figura 2.1 (d) apresenta o algoritmo de Agrupamento *K-Means* aplicado como técnica de discretização não supervisionada. Neste exemplo, a aplicação do algoritmo *K-Means* resultou na melhor categorização dos dados do atributo, seguido pela técnica de *frequência igual* e depois pela técnica de *largura igual*.

Apesar da existência das técnicas de discretização mencionadas acima e dos bons resultados que tais técnicas apresentam em diversos cenários, muitas vezes é necessário obter um número mínimo de intervalos ou conseguir obter intervalos considerados naturais. Por exemplo, dividir salários em intervalos de $R\$2.000$ até $R\$3.000$, ao invés de $R\$2.153,97$ até $R\$3.034,75$. Neste caso, a discretização por *IP*(Particionamento Intuitivo - *Intuitive Partitioning*) pode ser utilizada (Han et al., 2011). Nesta forma de discretização a regra 3-4-5 é aplicada. Tal regra divide o conjunto de dados do atributo em 3, 4 ou 5 intervalos da seguinte maneira:

1. Encontrar os valores max_x, min_x, low_x (5º percentil) e $high_x$ (95º percentil).

2. Encontrar o digito mais significativo (*msd*) com base nos valores low_x e $high_x$ e definir os valores low'_x, arredondando o valor de low_x para baixo, e $high'_x$, arredondando o valor de $high_x$ para cima, ambos de acordo com o *msd* encontrado.

3. Realizar o cálculo da regra 3-4-5 buscando encontrar número de valores distintos presente no *msd*, o qual definirá o número de intervalos em que o conjunto será dividido. Assim, o número de valores distintos é obtido pela Equação abaixo,

$$distincts = (high'_x - (-low'_x))/msd \tag{2.4}$$

onde *distincts* é o número de valores distintos encontrados. A partir deste resultado, a definição do número de intervalos da discretização ocorre da seguinte maneira:

- Quando *distincts* for igual a 3, 6, 7 ou 9, a discretização será em 3 intervalos, sendo 3 intervalos de tamanhos iguais quando *distincts* for 3, 6, ou 9, e intervalos de tamanho 2, 3, 2 quando *distincts* for 7.
- Quando *distincts* for igual a 2, 4 ou 8, a discretização será em 4 intervalos de tamanhos iguais.
- Quando *distincts* for igual a 1, 5 ou 10, a discretização será em 5 intervalos de tamanhos iguais.
- Ao final é necessário investigar se os valores min_x e max_x foram cobertos pelos intervalos criados dado o resultado de *distincts*. Caso algum desses valores não esteja coberto, um intervalo adicional é criado para cobrir tal valor.
- Se for necessário dividir cada intervalo obtido em mais intervalos, o processo deve ser repetido em uma segunda execução, a qual será realizada em cada um dos intervalos obtidos pela primeira execução.

Apesar da grande quantidade de técnicas de transformação de dados e de discretização de atributos, a área de *Pré-processamento* continua sendo amplamente explorada, principalmente pela grande quantidade de dados inconsistentes e pela complexidade dos problemas mais atuais.

2.2 Aprendizado de Máquina

De acordo com Simon (1996) aprendizado são mudanças que capacitam um sistema a fazer a mesma tarefa, ou tarefas similares, de maneira mais efetiva em uma próxima vez. Neste contexto, Mitchell (1997) e Alpaydin (2014) definem o Aprendizado de Máquina como um processo onde algoritmos e modelos aprendem automaticamente por meio de experiências anteriores visando otimizar um critério de desempenho.

Em geral, algoritmos e modelos de Aprendizado de Máquina podem ser classificados em três diferentes tipos, Aprendizado de Máquina não supervisionado, Aprendizado de Máquina semi-supervisionado e Aprendizado de Máquina supervisionado, os quais são caraterizados da seguinte forma (Tan et al., 2006):

- *Aprendizado de Máquina não supervisionado*: onde algoritmos aprendem por conta própria sem nenhuma informação de rótulo dos dados. Para tarefas deste tipo os algoritmos visam aprender padrões, regras, ou categorias nos dados.

- *Aprendizado de Máquina supervisionado*: onde algoritmos aprendem descrições gerais de conceitos com base em conjuntos de dados de entrada que são previamente rotulados por um supervisor. Dependendo do tipo do rótulo informado o problema pode ser de *Classificação* (rótulos discretos) ou de *Regressão* (rótulos contínuos).
- *Aprendizado de Máquina semi-supervisionado*: onde algoritmos aprendem com base em conjuntos de dados que são parcialmente rotulados enquanto outra parte de tais conjuntos não é rotulada.

A Seção a seguir apresenta uma visão geral de Aprendizado de Máquina não supervisionado, demonstrando algumas das tarefas que foram utilizadas nesta pesquisa.

2.2.1 Aprendizado de Máquina não supervisionado

Em algoritmos de Aprendizado de Máquina não supervisionado são utilizados procedimentos que visam encontrar padrões ou partições nos dados originais. Contudo, é necessário encontrar rótulos (classes ou conceitos) implícitos nestes dados, sem qualquer forma de supervisão (Nilsson, 1996; Tan et al., 2006). Em geral, existem três principais tarefas para o Aprendizado de Máquina não supervisionado, Regras de Associação, Agrupamento e Redução de Dimensionalidade, as quais possuem os seguintes objetivos (Tan et al., 2006):

- *Regras de Associação*: encontrar regras/padrões frequentes em conjunto de dados por meio da identificação de elementos que implicam na presença de outros elementos em uma mesma transação (evento).
- *Agrupamento*: organizar dados em grupos visando uma alta similaridade entre os objetos de um mesmo grupo e baixa similaridade entre instâncias de grupos distintos.
- *Redução de Dimensionalidade*: reduzir a dimensionalidade dos dados criando novos atributos que representem uma combinação dos antigos atributos buscando eliminar dados irrelevantes e diminuir a quantidade de ruídos.

A seguir são apresentadas visões gerais sobre as tarefas de Regras de Associação e Agrupamento, as quais são as principais tarefas abordadas nesta pesquisa.

Regras de Associação

Entre as diversas aplicações de Mineração de Dados, algumas estão ligadas a um típico e importante problema chamado Mineração de Regras de Associação. Esta tarefa, a partir de FCDs representa uma das mais importantes direções na comunidade de Mineração de Dados (Giannella et al., 2003). Devido ao grande crescimento no volume de dados, a mineração de padrões frequentes

em FCDs precisa ser executada a partir de grandes volumes de dados contínuos e em uma memória limitada. Este tipo de problema foi inicialmente introduzido por Agrawal et al. (1993) no início da década de 1990. Dado um conjunto de dados de transações, onde cada transação consiste em uma lista de itens (por exemplo, aplicativos utilizados em um *smartphone* ao longo de um período de tempo), a tarefa de Regras de Associação tem como objetivo localizar, de maneira eficiente, um conjunto de regras/padrões frequentes e significantes. Tais regras/padrões irão prever a ocorrência de um conjunto de itens com base na ocorrência de outros itens presentes nas transações. De acordo com Agrawal et al. (1993), esta tarefa pode ser descrita da seguinte maneira. Considere um conjunto de itens denominado I, onde $I = \{\mathbf{i}_1, \mathbf{i}_2, ..., \mathbf{i}_m\}$ e um conjunto de transações denominado T, onde $T = \{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_n\}$. Cada transação **t** em T contém um conjunto de itens **i** pertencente ao conjunto I. Se um conjunto de itens possui **k** itens, este é chamado de **k**-conjunto de itens. Uma regra de associação é a implicação $X \Rightarrow Y$, onde X e Y são itens (ou conjunto de itens) que satisfazem $X, Y \subseteq I$ e $X \cap Y = \varnothing$. X e Y são chamados, respectivamente, de *antecedente* (LHS) e *consequente* (RHS) de uma regra de associação. Em resumo, a geração de regras de associação possui dois passos principais: (a) geração de conjunto de itens frequentes e (b) geração de regras de associação (Agrawal et al., 1993). Para este fim, existem diferentes algoritmos que tem demonstrado sua efetividade. Neste caso, um dos algoritmos mais utilizados é o algoritmo *Apriori* proposto por Agrawal et al. (1994).

A força de uma regra pode ser medida em termos de *suporte* e *confiança*. Para selecionar regras interessantes valores de *suporte* e *confiança* são aplicados visando mensurar a significância das regras geradas (Agrawal et al., 1994). A medida de *suporte* (Equação 2.5), indica uma fração de transações que contém X e também Y, determinando quão frequente uma regra é aplicada para um dado conjunto de dados. Um grande valor de $suporte(X \Rightarrow Y)$ indica que uma grande porcentagem de transações contêm X e também Y. O *suporte* é uma importante medida pois uma regra que contém um valor muito baixo de *suporte* pode ocorrer por simples acaso (Tan et al., 2006).

$$suporte(X \Rightarrow Y) = \frac{número\ de\ transações\ contendo\ X\ e\ Y}{número\ total\ de\ transações\ em\ T} \tag{2.5}$$

Por outro lado, a medida de confiança (Equação 2.6), determina quão frequente os itens em Y aparecem em transações que contém itens de X. Um grande valor de $confiança(X \Rightarrow Y)$ significa que muitas transações contendo X também contém Y. A *confiança* também fornece uma estimativa da probabilidade condicional de Y dado X.

$$confiança(X \Rightarrow Y) = \frac{suporte(X \Rightarrow Y)}{suporte(X)} \tag{2.6}$$

Assim, a especificação de limiares de *suporte mínimo* e de *confiança miníma*, para a geração de regras de associação, permite que somente regras com valores acima de tais limiares possam ser geradas. Neste caso, resultados obtidos a partir de regras de associação precisam ser interpretados com cuidado. A inferência que uma regra de associação não necessariamente implica em causalidade. Em vez disso, tais regras sugerem um forte relação de co-ocorrência entre os itens

em X e os itens em Y. Muitos problemas desta tarefa confiam somente nas medidas de *suporte* e *confiança* para eliminar padrões desinteressantes. Entretanto, em alguns casos, regras que possuem altos valores de *confiança*, são algumas vezes desconsideradas. Isto ocorre, pelo fato de que a medida de *confiança* ignora o *suporte* referente a *consequência* da regra (Y). Uma das maneiras de enfrentar esse problema é aplicando um outra métrica chamada *lift* (Equação 2.7) (Tan et al., 2006; Han et al., 2011).

$$lift(X \Rightarrow Y) = \frac{suporte(X \Rightarrow Y)}{suporte(X) * suporte(Y)} \tag{2.7}$$

A métrica *lift* mede a razão entre o *suporte* e a *confiança* de uma regra. Tal métrica indica a probabilidade de ocorrência entre X e Y independentemente um do outro. Neste caso, é possível interpretar o *lift* da seguinte forma:

$$lift(X \Rightarrow Y) \begin{cases} = 1, \text{se } X \text{ e } Y \text{ são } independentes \\ > 1, \text{se } X \text{ e } Y \text{ são correlacionados } positivamente \\ < 1, \text{se } X \text{ e } Y \text{ são correlacionados } negativamente \end{cases} \tag{2.8}$$

Na literatura é possível encontrar muitas métricas que podem ser utilizadas para avaliar as regras de associação. Tew et al. (2014) apresentam um estudo onde mais de 100 métricas com nomes distintos foram encontradas. Tal estudo busca investigar a real diferença entre estas métricas. Apesar de existirem nomes diferentes, algumas das métricas encontradas eram iguais quando considerado suas equações. Ao final, tal estudo detectou 61 métricas distintas (em termos de nome e equações). Contudo, Tew et al. (2014) afirmam que muitas métricas são similares entre si, sendo possível agrupá-las em 21 tipos de métricas. Neste cenário, algumas métricas parecem ser similares ao *lift*. Mesmo assim, o *lift* continua sendo uma das métricas mais exploradas, principalmente por ser computacionalmente mais simples (Tew et al., 2014; Tan et al., 2006).

Algumas regras obtidas podem ser consideradas *redundantes*. Isso ocorre devido a relacionamentos entre um conjunto de itens existentes em regras anteriormente obtidas. Neste caso, uma regra pode ser considerada redundante se o seu *lift* é igual ou menor em relação a regras geradas anteriormente ao longo do processo (Tan et al., 2006). Por exemplo, pode-se obter duas regras: (a) $X \Rightarrow Y$, $lift = 3.019$ e (b) $Y \Rightarrow X$, $lift = 3.019$. Sendo assim, a regra (b) pode ser considerada uma regra *redundante* em relação a regra (a), pois esta regra não apresenta conhecimento extra ao conhecimento obtido anteriormente, podendo ser então desconsiderada (Han et al., 2011).

Além disso, os conjuntos de dados utilizados afetam diretamente o desempenho de muitos algoritmos de mineração de regras de associação. Muitos conjuntos reais possuem uma distribuição de *suporte* distorcida, onde a maioria dos itens possuem frequências relativamente baixa ou moderada, mas um pequeno número dos itens possuem frequências elevadas. Neste sentido, se o limiar de *suporte mínimo* escolhido for um valor mais alto em relação a frequência média ou mediana, a probabilidade de perder muitos padrões interessantes envolvendo itens de baixo suporte é alta.

Por outro lado, quando tal limiar é muito baixo, existe o risco de serem gerados padrões falsos, que relacionam um item de alta frequência com um item de baixa frequência, os chamados *padrões de suporte cruzados* (Tan et al., 2006). Padrões de suporte cruzados são conjuntos de itens cuja a relação de *suporte* é menor que um limiar especificado, onde itens de alta frequência (ex: *suporte* de $0,7$) são relacionados a itens de baixa frequência (ex: *suporte* de $0,03$). Tais padrões são susceptíveis de serem falsos pois as correlações entre os itens tendem a ser fracas, principalmente quando o *suporte mínimo* é suficientemente baixo (Tan et al., 2006). No entanto, como a detecção de padrões de suporte cruzados ocorre na geração de conjuntos de itens frequentes (passo *a*), é possível a utilização da métrica chamada *all-confidence*, algumas vezes chamada de *h-confidence*, que foi introduzida por Xiong et al. (2003). Com esta métrica, padrões de suporte cruzados podem ser detectados pela verificação da regra com menor *confiança* que pode ser gerada a partir de um conjunto de itens frequentes. Tal métrica (Equação 2.9) mede a correlação entre os itens dos conjunto de itens frequentes (Han et al., 2011). Dessa forma, um limiar de *all-confidence mínimo* deve ser indicado com o objetivo de eliminar tais padrões.

$$all\text{-}confidence(X) = \frac{suporte(X)}{max\_suporte\_item(X)} = \frac{suporte\,(X)}{max\,\{suporte(\mathbf{i}_m \mid \forall \mathbf{i}_m \in X)\}} \tag{2.9}$$

Por fim, é importante lembrar que não existe uma maneira pré-definida para a definição dos limiares de, *suporte*, *confiança*, *lift* e *all-confidence* (Tan et al., 2006; Han et al., 2011). Além disso, os resultados obtidos com Regras de Associação podem ajudar em outras tarefas de Aprendizado de Máquina como, por exemplo, Agrupamento e Classificação (Han et al., 2011). Em geral, a tarefa de Regras de Associação desempenha um importante papel na identificação de associações, correlações e outras relações interessantes entre dados, as quais podem ser empregadas em outras tarefas.

Agrupamento

Organizar dados em grupos é uma das mais importantes maneiras de entender e aprender (Jain e Dubes, 1988). Agrupamento é uma tarefa que visa agrupar um conjunto de objetos (instâncias), em diferentes classes (grupos), de acordo com suas similaridades. O objetivo desta tarefa é determinar um conjunto finito de grupos que descrevam de forma apropriada o conjunto de objetos. Neste sentido, objetos pertencentes a um mesmo grupo possuem alta similaridade entre si e apresentam baixa similaridade em relação aos objetos de grupos distintos (Han et al., 2011; Tan et al., 2006). Tal tarefa é considerada um método não supervisionado, onde os objetos não estão associados a uma classe, ou atributos de interesse (Jain e Dubes, 1988).

Atualmente, existe uma variedade de algoritmos disponíveis para a tarefa Agrupamento. Tais algoritmos são utilizados em várias áreas do conhecimento, estimulando ainda mais o desenvolvimento deste tipo de algoritmo. Neste caso, os algoritmos mais utilizados são divididos em diferentes categorias (Tan et al., 2006):

- *Algoritmos particionais*: por exemplo os algoritmos *K-Means* (MacQueen et al., 1967) e *X-Means* (Pelleg et al., 2000). Tais algoritmos geram partições simples visando descobrir grupos naturais presente no conjunto de dados. Este tipo de algoritmo é muito utilizado em aplicações de engenharia, onde partições simples são importantes (Jain e Dubes, 1988).

- *Algoritmos de densidade*: por exemplo o algoritmo *DBSCAN* (Ester et al., 1996),. Este tipo de algoritmo visa separar regiões dos dados que possuem alta densidade, de regiões dos dados com baixa densidade. Contudo, sua execução pode ser custosa para cenários com conjuntos de dados de alta dimensionalidade (Tan et al., 2006).

- *Algoritmos hierárquicos*: por exemplo os algoritmos *Complete Linkage*, *Single Linkage*, *WARD*, *Average (UPGMA)*, *PAM* e *ROCK*. Estes algoritmos podem ser aglomerativos ou divisivos (Jardine e Sibson, 1971) e organizam os dados em uma sequência de grupos próximos (Tan et al., 2006). Tal tipo de algoritmo é frequentemente utilizado em áreas biológicas, sociais e comportamentais (Jain e Dubes, 1988).

Abaixo são representados dois pseudo-códigos, um para o algoritmo particional *K-Means* (Algoritmo 2.1) e outro para um algoritmo hierárquico aglomerativo (Algoritmo 2.2).

Algoritmo 2.1: Pseudo-código do algoritmo *K-Means*. Adaptado de Tan et al. (2006).

```
1: Selecione k pontos como sendo os centróides iniciais.
2: repeat
3:     Compute k grupos atribuindo objetos ao centróide mais próximo.
4:     Recompute o centróide de cada grupo.
5: until que os centróides não mudem.
```

Algoritmo 2.2: Pseudo-código de um algoritmo hierárquico aglomerativo. Adaptado de Tan et al. (2006).

```
1: Compute a matriz de distâncias.
2: repeat
3:     Combine os dois grupos mais próximos.
4:     Atualize as distâncias entre o novo grupo e os grupos originais.
5: until que somente exista um grupo.
```

Muitos algoritmos para a tarefa de Agrupamento encontram desafios quando o conjunto de dados possui alta dimensionalidade, principalmente por serem baseado em proximidade ou densidade. Quando tal dimensionalidade aumenta, problemas que são relacionados ao tamanho do conjunto de dados, como esparcidade, ruídos e *outliers*, podem atrapalhar tal tarefa (Tan et al., 2006). Em geral, estes algoritmos podem se adaptar a diferentes tipos de cenários, e a definição do mesmo é considerada uma difícil tarefa (Han et al., 2011). Além disso, a maioria dos algoritmos de agrupamento precisa receber como entrada um parâmetro k, o qual representa o número de grupos que devem ser formados. Definir tal número é complicado, uma vez que muitos problemas do mundo

real não possuem as classes dos objetos. Neste sentido, diferentes medidas visam avaliar o melhor número de grupos a serem formados (Vendramin et al., 2010), entre elas:

- *SWC* (Largura da Silhueta - *Silhouette Width Criterion*): envolve o cálculo da dissimilaridade média do j-ésimo objeto ao seu grupo $\mathbf{p}$ ($a_{\mathbf{p},j}$), onde $\mathbf{p} \in \{1, ..., k\}$, e a menor dissimilaridade média do j-ésimo objeto em relação aos demais grupos ($b_{\mathbf{p},j}$), sendo $b_{\mathbf{p},j}$ o menor valor computado sobre $\mathbf{q} \in \{1, ..., k\}, \mathbf{q} \neq \mathbf{p}$. $a_{\mathbf{p},j}$ e $b_{\mathbf{p},j}$ são calculados como a distância média do j-ésimo objeto a todos os demais objetos do grupo em questão por uma medida de similaridade adotada. Assim, a *SWC* é calculada sobre todos os n objetos de acordo com a Equação 2.10 e a silhueta de cada objeto $\mathbf{x}_j$ é calculada de acordo com a Equação 2.11. Esta medida varia entre -1 e 1, onde valores positivos e distantes de zero indicam que os objetos são compatíveis com seus próprios grupos e não combinam com os grupos vizinhos.

$$SWC_k = \frac{1}{n} \sum_{j=1}^{n} s_{\mathbf{x}_j} \tag{2.10}$$

$$s_{\mathbf{x}_j} = \frac{b_{\mathbf{p},j} - a_{\mathbf{p},j}}{max(a_{\mathbf{p},j}, b_{\mathbf{p},j})} \tag{2.11}$$

- *DBI* (*Davis-Bouldin Index*) Equação 2.12: representa a similaridade média entre cada um dos k grupos e o grupo mais semelhante correspondente. $\overline{d}_{\mathbf{p}}$ é a distância média de todos os objetos do grupo $\mathbf{p}$ ao seu centróide $\overline{\mathbf{x}}_{\mathbf{p}}$, enquanto $\overline{d}_{\mathbf{q}}$ é a distância média de todos os objetos do grupo $\mathbf{q}$ ao seu centróide $\overline{\mathbf{x}}_{\mathbf{q}}$ e $d_{\overline{\mathbf{x}}_{\mathbf{p}},\overline{\mathbf{x}}_{\mathbf{q}}}$ é a distância entre os centróides $\overline{\mathbf{x}}_{\mathbf{p}}$ e $\overline{\mathbf{x}}_{\mathbf{q}}$. Neste caso, o menor valor de *DBI* indica uma boa coesão interna de cada grupo e a separação entre tais grupos.

$$DBI_k = \frac{1}{k} \sum_{\mathbf{p}=1}^{k} \underset{\mathbf{p} \neq \mathbf{q}}{max} \left[\frac{\overline{d}_{\mathbf{p}} + \overline{d}_{\mathbf{q}}}{d_{\overline{\mathbf{x}}_{\mathbf{p}},\overline{\mathbf{x}}_{\mathbf{q}}}} \right] \tag{2.12}$$

- *DUNN* (*Dunn Index*) Equação 2.13: envolve a menor distância entre um par de objetos pertencentes aos grupos $\mathbf{p}$ e $\mathbf{q}$ ($d_{c_{\mathbf{p}},c_{\mathbf{q}}}$) e a máxima distância entre dois objetos do mesmo grupo ($diam(c_{\mathbf{z}})$), onde $\mathbf{z} \in \{1, ..., k\}$. Assim, valores elevados de *DUNN* indicam grupos compactos e bem separados.

$$DUNN_k = \underset{\mathbf{p}=1,...,k}{min} \left\{ \underset{\mathbf{q}=\mathbf{p}+1,...,k}{min} \left(\frac{d_{c_{\mathbf{p}},c_{\mathbf{q}}}}{\underset{\mathbf{z}=1,...,k}{max}\, diam(c_{\mathbf{z}})} \right) \right\} \tag{2.13}$$

- *CH* (*Calinski-H Criterion*) Equação 2.14: é definido como a relação entre a dispersão interna dos grupos ($W(k)$) (Equação 2.16) e a dispersão entre os grupos e o centro do conjunto de dados ($B(k)$) (Equação 2.15). $\|\overline{\mathbf{x}}_{\mathbf{p}} - \mathbf{x}\|^2$ é a distância euclidiana entre o centróide do grupo $\mathbf{p}$ e a média geral do conjunto de dados, enquanto $\|\mathbf{x}_j - \overline{\mathbf{x}}_{\mathbf{p}}\|^2$ é a distância euclidiana entre o

j-ésimo objeto do grupo **p** e o centróide do grupo **p**. Maiores valores de *CH* indicam a melhor estratégia de agrupamento.

$$CH_k = \frac{B_k/(k-1)}{W_k/(n-k)} \tag{2.14}$$

$$B_k = \sum_{\mathbf{p}=1}^{k} n_{\mathbf{p}} \left\| \bar{\mathbf{x}}_{\mathbf{p}} - \mathbf{x} \right\|^2 \tag{2.15}$$

$$W_k = \sum_{\mathbf{p}=1}^{k} \sum_{j=1}^{n_{\mathbf{p}}} \left\| \mathbf{x}_j - \bar{\mathbf{x}}_{\mathbf{p}} \right\|^2 \tag{2.16}$$

- *Elbow method*: também se baseia na dispersão interna dos grupos (Equação 2.16), que também é chamada de variância *intra-cluster* ou *SSE* (Soma do Erro Quadrático - *Sum of Squared Error*). Nesta medida, o melhor número de grupos é indicado quando se verifica um "joelho" entre os valores de *SSE* calculados para cada k. Conforme o k cresce o valor da dispersão interna tende a diminuir monotonicamente e o "joelho" pode ser observado quando existir uma grande desaceleração de tal valor. Na Figura 2.2 são apresentados valores de *SSE* (eixo y) com k variando de 2 até 10 (eixo x), sendo possível identificar um "joelho" quando $k = 4$. Contudo, tal identificação nem sempre é possível, uma vez que o "joelho" nem sempre pode ser identificado.

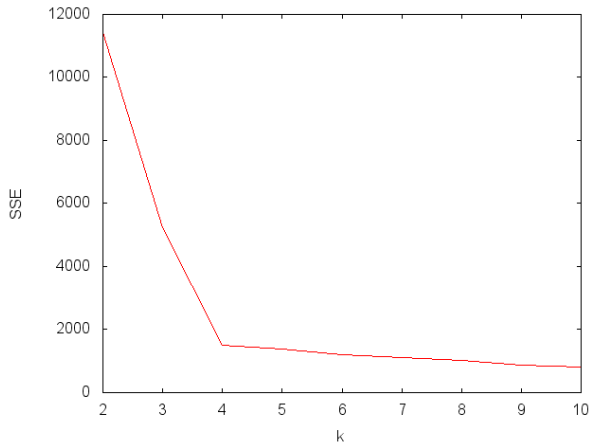


Figura 2.2: Exemplo da visualização do *Elbow method* para avaliação do melhor número de grupos em uma tarefa de agrupamento.

- *GAP* (*Gap Statistic*) (Tibshirani et al., 2001): é uma abordagem estatística que visa formalizar a diminuição monotônica da dispersão interna dos grupos. Em resumo, busca-se padronizar o gráfico de *log* da distribuição interna ($log(W_k)$), comparando tal resultado com a sua expectativa sob uma distribuição de referência nula apropriada dos dados. A dispersão interna W_k pode ser calculada como a soma agrupada das distâncias a cada par de objetos de todos os grupos **p** (Equação 2.17). Deste modo, devem ser criadas R referências nulas dos dados,

as quais são agrupadas em k grupos. Assim, o valor de *Gap* para cada k é computado de acordo com a Equação 2.18, onde $E_R^* log(W_k^*)$ é a média do valor de $log(W_{kr}^*)$ das R referências criadas. Nesta medida, o melhor número de grupos pode ser indicado pelo menor k onde $Gap(k) \geq Gap(k+1) - S_{k+1}$. Para este fim, é necessário o calculo do desvio padrão dos resultados de $log(W_{kr}^*)$ das R referências (Equação 2.19), onde $\bar{l} = (1/R) \sum_r log(W_{kr}^*)$. Então, o resultado de sd_k é utilizado para calcular o erro de simulação s_k (Equação 2.20), o qual será utilizado para verificar a variação do resultado do *Gap* para cada valor de k em relação a $k+1$. Contudo, esta não é a única forma de verificar a provável quantidade de grupos existes com o *Gap*. Em casos onde existem grupos menores dentro de grupos maiores e bem separados, o *Gap* pode exibir um comportamento não monótono, sendo necessário examinar toda a curva do intervalo, em vez de simplesmente encontrar a posição do seu máximo.

$$W_k = \sum_{\mathbf{p}=1}^{k} \frac{1}{2n_{\mathbf{p}}} \sum_{j=1}^{n_{\mathbf{p}}} d_{\mathbf{x}_j, \mathbf{x}_j'} \tag{2.17}$$

$$Gap_k = E_R^* \{log(W_k^*)\} - log(W_k) \tag{2.18}$$

$$sd_k = \left[(1/R) \sum_r \left\{ log(W_{kr}^*) - \bar{l} \right\}^2 \right]^{1/2} \tag{2.19}$$

$$s_k = \sqrt{1 + 1/R}\, sd_k \tag{2.20}$$

Por fim, pode-se afirmar que é possível apontar e determinar o melhor número de grupos existentes em um determinado conjunto de dados, desde que sejam investigadas e aplicadas medidas de avaliação de agrupamento (Vendramin et al., 2010).

Distância entre Instâncias

Alguns algoritmos para tarefa de Agrupamento, bem como algoritmos para outras tarefas de Aprendizado de Máquina, necessitam de um valor de similaridade entre as instâncias que compõem o conjunto de dados analisado (Jain e Dubes, 1988). Uma das maneiras de mensurar a similaridade entre estes objetos é calculando a distância entre as características de tais objetos. Esta distância pode ser calculada de diferentes formas, de acordo com os atributos existentes. Dentre as possibilidades, tal distância pode ser calculada por medidas como: *Euclidiana*, *Manhattan*, *Minkowski* e *Jaccard Index* (Han et al., 2011).

A distância *Euclidiana* (Equação 2.21) é uma das mais utilizadas, principalmente em algoritmos particionais e em áreas como a Engenharia (Jain e Dubes, 1988). Neste caso, a similaridade

entre dois objetos $\mathbf{x}_j$ e $\mathbf{x}_j^{'}$ é obtida da seguinte maneira. Para cada atributo **j** do conjunto de dados, onde **j** varia de 1 até m, é calculada, a diferença entre os dois objetos, a qual é elevada ao quadrado. Ao final é calculada a raiz quadrada do somatório das m diferenças entre tais objetos.

$$Dist\_E\left(\mathbf{x}_j, \mathbf{x}_j^{'}\right) = \sqrt{\sum_{\mathbf{j}=1}^{m}\left(\mathbf{x}_{j,\mathbf{j}} - \mathbf{x}_{j,\mathbf{j}}^{'}\right)^2} \tag{2.21}$$

A medida *Jaccard* (Equação 2.22) é utilizada para comparar a dissimilaridade entre dois conjuntos de dados. Esta medida tem como objetivo mensurar quais itens são compartilhados e quais são distintos. Desta forma, a dissimilaridade entre dois conjuntos I e $I^{'}$ é calculada com base no tamanho do conjunto de intersecção $|I \bigcap I^{'}|$ dividido pelo tamanho do conjunto de união $|I \bigcup I^{'}|$ entre tais conjuntos. Neste caso, pela teoria dos conjuntos, temos $|I \bigcap I^{'}| = \{\mathbf{i}|\mathbf{i} \in I \bigwedge \mathbf{i} \in I^{'}\}$ e $|I \bigcup I^{'}| = \{\mathbf{i}|\mathbf{i} \in I \bigvee \mathbf{i} \in I^{'}\}$. Por fim, caso não exista uma intersecção entre os conjuntos ($|I \bigcap I^{'}| = \varnothing$) a distância entre I e $I^{'}$ é definida como 1.

$$Dist\_J(I, I^{'}) = 1 - \frac{\left|I \bigcap I^{'}\right|}{\left|I \bigcup I^{'}\right|} = 1 - \frac{\left|I \bigcap I^{'}\right|}{|I| + |I^{'}| - |I \bigcap I^{'}|} \tag{2.22}$$

Ambas as medidas são calculadas a cada par de objetos em um espaço multidimensional e então são utilizadas para o seu devido fim. Por exemplo, tais medidas podem ser utilizadas para a tarefa de Agrupamento por meio de um determinado algoritmo. Contudo, de acordo com Gan et al. (2007) a melhor forma de medir a similaridade entre instâncias é frequentemente obtida por meio da combinação de experiência, habilidade, conhecimento e sorte.

2.3 Considerações Finais do Capítulo

Neste capítulo foram apresentados conceitos de Mineração de Dados e Aprendizado de Máquina utilizados no desenvolvimento desta pesquisa. Em particular, detalhes das técnicas de Pré-processamento de dados e das tarefas de Aprendizado de Máquina não supervisionado, como Transformação de dados, Discretização, Regras de Associação, Agrupamento e Distância entre Instâncias.

Nesta tese, dados de uso de aplicativos em dispositivos móveis são investigados visando identificar e monitorar diferentes tipos de perfis de uso de tais aplicativos. Neste sentido, as técnicas de Pré-processamento visam aprimorar a representação dos dados e melhorar os resultados da execução de algoritmos de Aprendizado de Máquina não supervisionado. Enquanto a tarefa de Regras de Associação é vantajosa para a descoberta de conhecimento e correlações de uso de aplicativos, a tarefa de Agrupamento possibilita a identificação de conceitos (perfis) com base na similaridade dos conjuntos de itens frequentes investigados.

No Capítulo 3 são apresentados conceitos e características de FCD assim como técnicas deste tipo de cenário, como Janela de Eventos e Detecção de Novidade. Tais técnicas são utiliza-

das no desenvolvimento desta pesquisa buscando monitorar os perfis e comportamentos de uso de aplicativos em dispositivos móveis ao longo do tempo.

3. FLUXO CONTÍNUO DE DADOS

Nas últimas décadas, grande parte das abordagens de Aprendizado de Máquina foram realizadas em cenários estacionários (*batch*). Neste tipo de cenário, um conjunto de dados é previamente disponível em um banco de dados convencional e sobre uma distribuição que não muda com o passar do tempo. Tal conjunto permanece disponível para ser processado por um determinado algoritmo, uma ou mais vezes durante a fase de treinamento, de acordo com a necessidade da abordagem utilizada. Este algoritmo produz um modelo, que uma vez aprendido não muda ao longo do tempo (Gama, 2010).

Atualmente, pesquisas e aplicações em Aprendizado de Máquina possuem novos desafios, principalmente pelo grande progresso da ciência e da tecnologia da informação. Este progresso traz novas fontes de produção de dados, aumentando a complexidade dos conjuntos de dados disponíveis (Gama e Gaber, 2007). Normalmente são dados gerados de forma contínua, possuindo um grande Volume (ex: *Big Data*), uma grande Variedade (ex: diferente tipos de dados) e sendo produzidos em alta Velocidade (ex: em tempo real), o que dificulta o armazenamento e a análise destes novos tipos de conjuntos, chamados FCD. Tais dificuldades são geradas devido às limitações físicas dos recursos computacionais atuais em comparação a grande quantidade de dados gerados (Gama, 2010; Aggarwal, 2007). Dado este novo tipo de dados, a aplicação de Mineração de Dados e tarefas de Aprendizado de Máquina depende de novas abordagens que visam superar as dificuldades da geração de grandes quantidade de dados em tempo real. Além disso, técnicas de Detecção de Novidades são utilizadas com o objetivo de identificar novos conceitos claramente diferentes dos conceitos aprendidos anteriormente, o que é necessário dado a mudança na distribuição dos dados presente em novos cenários Markou e Singh (2003). Assim, neste Capítulo são apresentados conceitos de FCD e suas principais características (ver Seção 3.1), Mineração de Dados e Aprendizado de Máquina em FCD (ver Seção 3.2) e Detecção de Novidade (ver Seção 3.3). Por fim, a Seção 3.4 apresenta as considerações finais do Capítulo.

3.1 Características do FCD

Gama (2010) define FCD como um processo estocástico em que eventos ocorrem continuamente e independentemente ao longo do tempo. Assim, o que distingue os conjuntos de dados mais atuais é a produção de dados continuamente. Neste caso, novos conceitos podem surgir e conceitos existentes podem mudar ou desaparecer ao longo do tempo. Desta forma, a obtenção de conhecimento sobre este tipo de dado é considerada uma tarefa difícil, uma vez que é necessário um processo eficaz que seja capaz de evoluir constantemente sobre tais FCDs (Fan e Bifet, 2013; PhridviRaj e GuruRao, 2014).

Os dispositivos móveis estão constantemente em evolução. Por meio do uso de aplicativos instalados nestes dispositivos, clientes podem fazer diferentes atividades. Tais atividades podem va-

riar conforme o surgimento de novos aplicativos ou novas necessidades dos consumidores. Portanto, são produzidos eventos que devem ser analisados em tempo real e juntos formam grandes volumes de dados mudando ao longo do tempo. Desta forma, um FCD D pode ser considerado como um fluxo de eventos potencialmente infinito, ou seja, $D = \{\varepsilon_1, \varepsilon_2, ...\varepsilon_n\}$ *onde* $D = \{\varepsilon_j\}_{j=1}^{n}$ *e* $(n \to \infty)$, o qual contém um conjunto de atributos conhecido e finito (Aggarwal, 2003). Em resumo, um FCD é uma sequência de eventos que podem ser lidos somente uma vez ou um número reduzido de vezes, utilizando capacidades de computação e de armazenamento limitadas. Neste sentido, Babcock et al. (2002a) cita algumas das principais diferenças entre Aprendizado de Máquina em FCDs e Aprendizado de Máquina em *batch*:

- O FCD possui uma infinita sequência de eventos $D = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_n\}$ que chegam continuamente, em tempo real e em alta velocidade;
- Os objetos do FCD são uma sequência não ordenada de m dimensões;
- Geralmente o FCD é muito grande pra ser armazenado na memória;
- Normalmente, objetos do FCD devem ser analisados somente uma vez pelo algoritmo de Aprendizado de Máquina e então devem ser descartados.

Ao longo da última década, a área de FCDs tem atraído a atenção da comunidade de Mineração de Dados e Aprendizado de Máquina. Pesquisas estão sendo realizadas tanto para desenvolver novas técnicas ou adaptar técnicas existentes, visando melhorar o processamento e a análise destes FCDs (Gaber, 2012). Por fim, é importante salientar que dispositivos móveis estão se tornando a porta de entrada para se obter dados em tempo real. Tais dados são, por exemplo, dados de clientes e aplicativos, que possuem diferentes aspectos. Além disso, a grande quantidade de dados que empresas fabricantes destes dispositivos podem potencialmente processar, tem aumentado gradativamente. Tal quantidade ultrapassou significativamente o passado simples de registro de dados de chamadas entre consumidores, onde o processamento era realizado apenas para fins de cobrança. Em geral, este tipo de FCDs possui grandes propriedades por serem infinitos e evolutivos, os quais são grandes e complexos para serem processados e analisados por técnicas de Mineração de Dados e Tarefas de Aprendizado de Máquina (Fan e Bifet, 2013).

3.2 Mineração de Dados em FCD

Como os FCDs possuem grande volume, grande variedade, e são produzidos em alta velocidade, existe a necessidade de métodos inovadores que possam analisar e processar este tipo de dado. Áreas como Inteligência Artificial, Aprendizado de Máquina e Estatística, apresentam técnicas para lidar com esse tipo de situação. Tais técnicas visam melhorar a qualidade dos dados, de modo a torná-los viáveis para futura utilização ou processamento, buscando melhorando o entendimento e beneficiar a tomada de decisão (Jain e Srivastava, 2013; PhridviRaj e GuruRao, 2014).

O objetivo da Mineração de FCDs é analisar e extrair o maior número de informações a partir de um conjunto de dados capturado ao longo de uma Janela de Eventos (Ver Seção 3.2.2), montando uma estrutura compreensível para utilização futura. Por exemplo, com o crescimento do número de eventos gerados a partir de vários aplicativos utilizados frequentemente em *smartphones*, o cenários dos dados não é mais estático, e sim dinâmico. Essa mudança traz uma série de desafios para a Mineração de Dados. Em alguns casos, algoritmos tradicionais, não são adequados para lidar com os FCDs, por necessitarem a realização de várias análises sobre os dados. Dessa forma, percebe-se o grande desafio para a Mineração de Dados com relação à FCDs (PhridviRaj e GuruRao, 2014).

A mineração de FCDs vêm crescendo e alcançando diferentes mercados. Neste caso, pode-se citar o sistema *MobiMine*, desenvolvido na década passada por Kargupta et al. (2002). Tal sistema tem como objetivo o monitoramento de um mercado de ações via um PDA (Assistente Digital Pessoal - *Personal Digital Assistant*). Pouco tempo depois Kargupta et al. (2004) desenvolveram outro sistema, chamado *VEDAS*, visando a mineração de FCDs oriundo de uma frota de veículos, onde foram analisados os comportamentos dos motoristas e a performance dos veículos de tal frota. Após as primeiras abordagens de Mineração de FCDs, percebeu-se uma invasão de informações que ultrapassou as capacidades tecnológicas de processar, analisar, armazenar e entender conjuntos de dados. Um dos cenários com grande crescimento é o mercado de dispositivos móveis, acelerado pelo desenvolvimento de novas funcionalidades bem como a evolução de aplicativos como, por exemplo, *Whatsapp*, *Facebook*, *Twitter*, *Instagram*. Consequentemente, tal progresso impacta as capacidades tecnológicas e dificulta o processamento de dados gerados por tantos aplicativos, em muitos dispositivos e em diferentes formatos (Fan e Bifet, 2013).

É possível observar um foco crescente sobre técnicas e estratégias para a Mineração de Dados de dispositivos móveis (Blondel et al., 2015). Embora exista um crescimento na investigação e na pesquisa desta área nos últimos anos, a Mineração em FCDs, principalmente em dados de dispositivos móveis, tem agora um nível de maturidade significativo e com muitas possibilidades de crescimento (Krishnaswamy et al., 2012). Sendo assim, é possível dizer que a Mineração em FCDs de dispositivos móveis é uma promissora área de pesquisa.

3.2.1 Aprendizado de Máquina em FCD

Atualmente, existem e estão disponíveis muitos algoritmos de Aprendizado de Máquina. Tais algoritmos estão sendo adequados para serem utilizados em cenários de FCDs. As estratégias de adaptação de algoritmos para FCDs tem sido um foco de investigação significativo (Krishnaswamy et al., 2012). Nesse sentido, tarefas de Agrupamento, de Regras de Associação e de Classificação são escolhas de interesse comum entre pesquisadores de Mineração de Dados que trabalham com FCDs (PhridviRaj e GuruRao, 2014).

A tarefa de Agrupamento em cenários de FCDs se torna ainda mais complicada. Neste cenário, muitos obstáculos surgem devido a chegada de dados de maneira contínua e também devido a necessidade de realizar a análise de tais dados, algumas vezes, em tempo real. Além disso, o FCD pode evoluir com o passar do tempo (Gama, 2012).

No decorrer das últimas décadas, alguns algoritmos foram criados ou modificados para a tarefa de Agrupamento em FCD, tais como: *CluStream* (Aggarwal, 2003), *StreamKM++* (Ackermann et al., 2012), e *Clus-Tree* (Kranen et al., 2011). Muitos destes algoritmos consideram que eventos de um FCD sejam uma representação completa de objetos independentes e que podem ser diretamente agrupados. Neste contexto, a tarefa de Agrupamento em FCDs pode ser resumida em duas fases Silva et al. (2013), chamadas de: Fase *online* e Fase *offline*, como mostra a Figura 3.1.

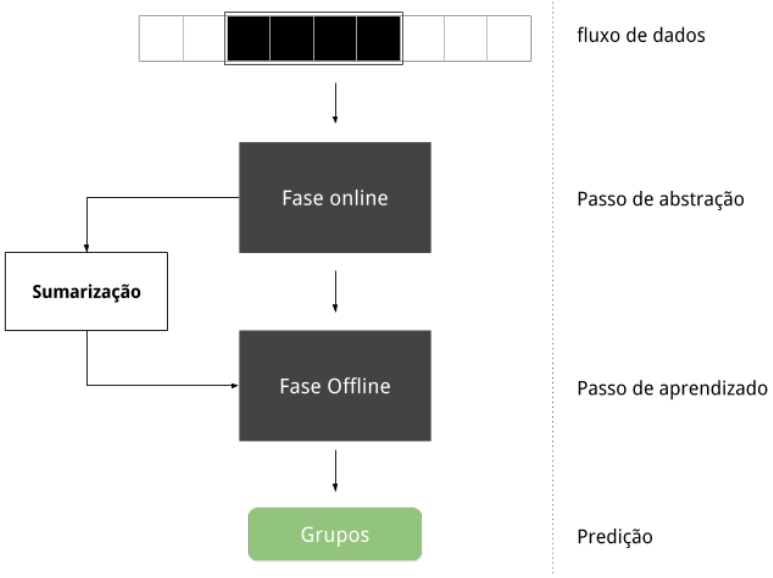


Figura 3.1: Processo da tarefa de Agrupamento em FCDs. Adaptado de Silva et al. (2013).

A fase *online*, ou passo de abstração, sumariza o FCD com a utilização de estruturas de dados que visam lidar com as restrições de espaço e memória de aplicações de FCDs. Tais estruturas tem como objetivo preservar o real significado dos dados originais sem a necessidade de armazenar todo o conjunto de dados. Uma abordagem popular para a sumarização de FCD consiste na definição de uma Janela de Eventos (Ver Seção 3.2.2). Estas janelas tem como objetivo trabalhar com um conjunto de dados menor e mais recente.

A fase *Offline*, ou passo de aprendizado, é aplicada juntamente com alguns outros parâmetros (ex: número de grupos e variação da janela). Tal fase é responsável por organizar um rápido entendimento dos conceitos gerados por meio da tarefa de Agrupamento (Silva et al., 2013). É importante notar que este processo, com as fases *online* e *offline*, se aplica também para tarefas como Classificação, bem como para o uso de diferentes algoritmos (Faria et al., 2013).

Entretanto, alguns destes algoritmos não podem ser aplicados em todos cenários reais dinâmicos, dado a forma em que os eventos do FCD são investigados. Por exemplo, no cenário abordado por esta pesquisa, os eventos de um FCD são registros (atividades de uso de aplicativos)

realizados em um único objeto (dispositivo móvel). Assim, tais algoritmos não podem ser praticados diretamente ao FCD abordado, sendo necessário realizar adaptações para realizar a tarefa de Agrupamento em tal cenário.

De modo geral, a generalização de redes de dispositivos móveis e o grande aumento no uso de aplicativos em tais dispositivos fornecem um cenário onde a tarefa de Agrupamento pode ser importante e necessária (Blondel et al., 2015). A exploração de informações, como o uso de aplicativos e outros dados de dispositivos móveis, bem como a investigação da evolução de tais dados ao decorrer do tempo, podem ajudar a estabelecer diferentes perfis e a identificar comportamentos de consumidores. Neste contexto, a tarefa de Agrupamento destes FCDs pode ajudar a analisar e compreender tais situações, permitindo que empresas, como fabricantes de tais dispositivos mantenham seus clientes oferecendo novos produtos e serviços.

3.2.2 Janela de Eventos

O uso de Janelas de Eventos, do inglês *Time Windows*, é a forma mais simples de manter uma quantidade praticável de dados em memória física. Tais janelas, também chamadas de mecanismos de esquecimento, podem ajudar na transição entre dados do passado recente e dados de um passado longínquo. Vários modelos de janelas tem sido apresentado na literatura. Babcock et al. (2002b) define dois tipos básicos de Janelas de Eventos:

- Janela baseada em sequência, Deslizante;
- Janela baseada em tempo, de Marcação.

A Figura 3.2 exemplifica o uso de duas Janelas Deslizantes, a *Sliding Window* (a) e a *Landmark Window* (b) e uma Janela de Marcação, a *Timestamp Window* (c).

As janelas baseada em sequência possuem um tamanho que é definido em termos de quantidade de objetos observados. Tais janelas são utilizadas de forma simples, para esquecer dados do passado, quando possuem um tamanho fixo. Entretanto, estas janelas podem ter um tamanho variável. Quando este tipo de janela possui tamanho fixo, ela segue uma estrutura baseada em uma fila (FIFO - *first in first out*), onde os dados mais recentes são mantidos e os dados mais antigos são desprezados. Neste caso, existem dois tipos de janelas, *Sliding Window* com um tamanho definido que desliza ao longo do FCD e *Landmark Window* onde a janela cresce a partir de um ponto específico t no tempo.

Por outro lado, janelas baseadas em tempo (de Marcação) são definidas em termos de duração. Este tipo de janela possui tamanho variável definido por um marcador de tempo e por isso é chamada de *Timestamp Window*. Tal janela w consiste em todos elementos que chegam dentro de um intervalo de tempo especificado (ex: $[t1, t2]$. Diferentemente de janelas deslizantes, as janelas de marcação iniciam uma nova janela e não deslizam ao longo do FCD. Contudo, os

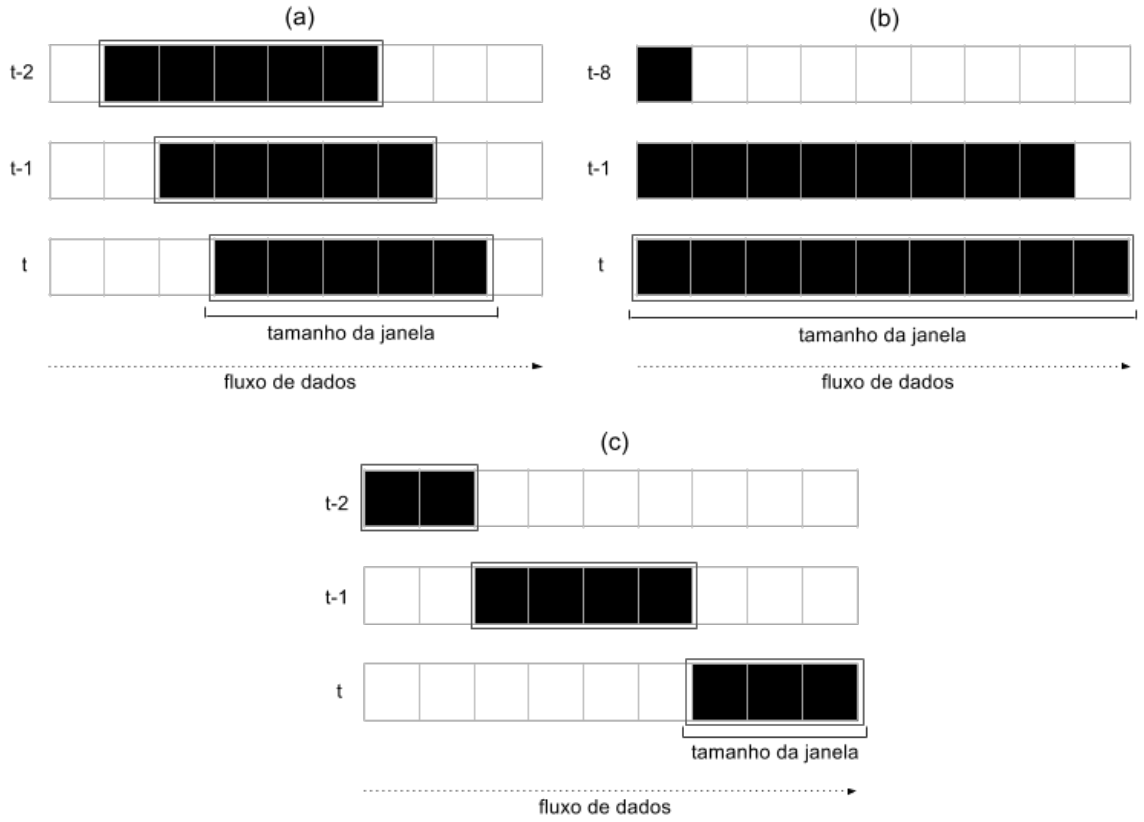


Figura 3.2: Modelos de Janelas de Eventos. Janelas Deslizantes: *Sliding Window* (a) e *Landmark Window* (b); e Janela de Marcação: *Timestamp Window* (c). Adaptado de Gama et al. (2014).

dados processados por ambos os tipos de janelas podem ser descartados ou sumarizados para serem armazenados.

Em cenários não estacionários como FCD, os dados podem evoluir com o passar das Janelas de Eventos analisadas. Esta evolução pode ocasionar mudanças nos conceitos aprendidos, sendo esta outra importante característica deste tipo de cenário. Assim, novos conceitos podem surgir e conceitos conhecidos podem desaparecer ou evoluir ao longo do tempo. Dessa forma, é necessário investigar mudanças na distribuição dos dados (Faria et al., 2013), o que é realizado por meio de técnicas de Detecção de Novidade.

3.3 Detecção de Novidade

Detecção de Novidade é a habilidade de identificar uma nova ou desconhecida instância ou um agrupamento delas, que represente um conceito claramente diferente dos conceitos aprendidos anteriormente (Markou e Singh, 2003). De acordo com Gama (2010), a Detecção de Novidade possibilita o reconhecimento de novos conceitos em dados não rotulados. Esta identificação pode indicar o surgimento de um novo conceito, uma mudança ou desaparecimento de conceitos já existentes. Desta forma, vários estudos mostram que a Detecção de Novidade é uma tarefa extremamente desafiadora e muito importante para sistemas de aprendizado. Neste sentido, pesquisadores de de

Mineração de Dados e Aprendizado de Máquina têm apresentado e discutido diferentes modelos de Detecção de Novidade para diferentes tipos de dados e problemas do mundo real (Markou e Singh, 2003).

Alguns modelos presentes na literatura tratam o problema de Detecção de Novidade em cenários *batch* onde não existem Janelas de Eventos. Contudo, a Detecção de Novidade é uma tarefa desafiadora em cenários do mundo real atual, onde FCDs estão em constante evolução e várias Janelas de Eventos precisam ser investigadas. Neste contexto, novos conceitos devem ser apreendidos a partir de novas observações que não correspondem à representação do cenário atual (Gama, 2012, 2010). Nesta perspectiva, Gama (2010) apresenta alguns desafios para a Detecção de Novidade em FCDs, são elas: Detecção de Anomalia, Detecção de *Outlier*, Mudança de Conceito e Evolução de Conceito.

3.3.1 Detecção de Novidade, de Anomalia e de Outlier

De acordo com Gama (2010) Detecção de Novidade, de Anomalia e de *Outlier* são definições similares que estão atraindo cada vez mais a atenção da comunidade de Mineração de Dados e Aprendizado de Máquina. Em geral, estas três tarefas tem como objetivo encontrar conceitos que são desconhecidos ou diferentes dos conceitos existentes. Contudo, tais termos são facilmente diferenciados e podem ser aplicados para diferentes tipos de problemas.

Conforme Chandola et al. (2009) a Detecção de Novidade é uma tarefa que visa detectar conceitos não observados indicando um conceito novo ou emergente que precisa ser aprendido. Por outro lado, Chandola et al. (2009) define Detecção de Anomalia como uma tarefa que tem como objetivo encontrar conceitos que não estão em conformidade com os padrões esperados sinalizando, muitas vezes, um padrão indesejado.

Aggarwal (2013) define Detecção de *Outlier* como uma tarefa que visa encontrar um objeto, ou um conjunto de objetos que se diferenciam extremadamente dos demais objetos, os quais podem ser considerados anormais. Dessa forma, é importante ressaltar que um dado anômalo é um dado específico de *Outlier*, acusando um padrão indesejado ou um comportamento que está sendo buscado.

Em geral, Detecção de Novidade indica o reconhecimento de um novo padrão reconhecido que precisa ser apreendido. Em alternativa, Detecção de Anomalia e Detecção de *Outlier* buscam encontrar problemas muito similares. As duas técnicas apresentam termos semelhantes e buscam a constatação de um padrão indesejado, o que diferencia ambas as técnicas da Detecção de Novidade.

Diferentes motivos impulsionam pesquisadores na aplicação de Detecção de Novidade. Markou e Singh (2003) afirmam que a Detecção de Novidade é uma importante tarefa, principalmente pelo fato de que em muitos problemas, os dados de treinamento não são capazes de representarem todos ao conceitos possíveis. Para Gama (2010), uma vez que os dados de teste contém informações sobre conceitos desconhecidos pelos dados de treinamento, a Detecção de Novidade é uma tarefa

fundamental para um bom sistema de Aprendizado de Máquina. Além disso, para Gama (2010) a Detecção de Novidade possibilita o reconhecimento de novos conceitos em dados onde as classes são desconhecidas.

3.3.2 Mudança e Evolução de Conceitos

No cenário de FCD, dois interessantes fenômenos podem acontecer, Mudança de Conceito e Evolução de Conceito. De acordo com Mitchell (1997), um conceito é uma função definida sobre um conjunto de dados. Tal função mapeia entradas e saídas e é apreendida por um algoritmo. Neste sentido, a Detecção de Mudança ou Evolução de Conceito é uma importante tarefa para o Aprendizado de Máquina em ambientes dinâmicos, principalmente porque as mudanças podem se manifestar em diferentes formas, como mostra a Figura 3.3 (Brzeziński, 2010).

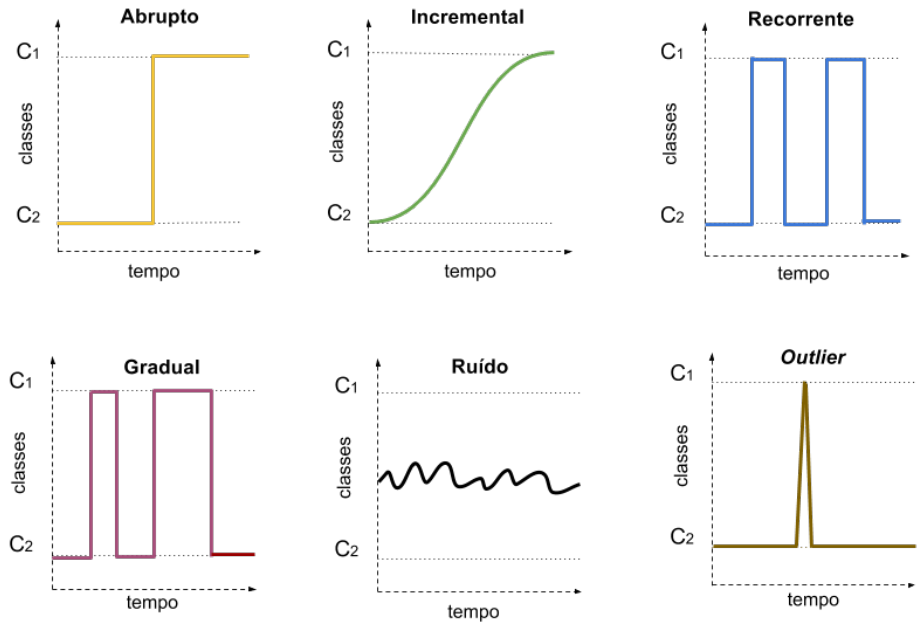


Figura 3.3: Diferentes tipos de mudanças de conceitos em FCDs. Adaptado de Brzeziński (2010).

Seis tipos básicos de Mudanças de Conceitos, as quais podem ocorrer ao longo do tempo, são representadas na Figura 3.3. Em geral, as mudanças podem ser: Abrupta, onde a classe muda instantaneamente; Incremental e Gradual, que apresentam mudanças que ocorrem lentamente, onde a mudança Incremental ocorre quando as variáveis mudam seus valores e a mudança Gradual ocorre quando a mudança envolve a distribuição da classe das variáveis, sendo ambas mudanças consideradas sinônimos por alguns pesquisadores; Recorrente, representa mudanças temporais que são revertidas depois de algum tempo; Ruído e *Outlier*, que não são considerados reais mudanças de conceito, onde Ruído é uma variação insignificante que pode afetar a Detecção de Mudanças

e *Outlier* são eventos raros que podem indicar comportamentos incomuns que ocorrem de forma aleatória (Brzeziński, 2010).

Em cenários tradicionais *batch*, o número de conceitos normalmente é conhecido pelo algoritmo, o qual reconhece que os objetos pertencem a uma das classes previamente conhecidas (Park e Shim, 2010). Contudo, em cenários do mundo real, onde o ambiente é dinâmico, modelos preditivos devem estar preparados para se adaptar as contínuas mudanças dos dados. Tais mudanças são esperadas devido ao não conhecimento de todas as classes na fase de treinamento, sendo possível o surgimento de novas classes ao longo do tempo.

Farid et al. (2013) definem a ocorrência de Mudança de Conceito da seguinte forma: Todo evento ε_j é produzido segundo uma fonte que reflete uma certa distribuição em um momento do tempo DI_t. Se para cada dois eventos ε_1 e ε_2 com tempos t_1 e t_2, $DI_1 \neq DI_2$, então uma Mudança de Conceito ocorre. Neste sentido, algoritmos devem ser capazes de detectar rapidamente novos conceitos e atuar com Janelas de Eventos, respeitando requisitos de uma fase de aprendizado. Por exemplo, executa-se a tarefa de Agrupamento em um conjunto de eventos de uma janela w_j e posteriormente realiza-se a mesma tarefa em um conjunto de eventos de uma janela posterior w_{j+1}, caso seja detectada Mudanças ou Evoluções de Conceitos, tais ocorrências podem ser classificadas de diferente formas (Spiliopoulou et al., 2006; Oliveira e Gama, 2010c; Ntoutsi et al., 2011).

- Um conceito pode sobreviver;
- Um conceito pode ser dividido em vários novos conceitos;
- Um novo conceito pode ser gerado pela união de diferentes conceitos;
- Um conceito pode desaparecer;
- Um novo conceito pode surgir.

Apesar de existirem modelos que tratam a Detecção de Novidade em diferentes áreas de aplicação, é possível perceber que algumas questões ainda precisam ser tratadas. Em cenários de dados estacionários a Detecção de Novidades se mostra bem definida. Entretanto, em cenários de FCD, diversas abordagens tem sido propostas por diferentes autores (Spiliopoulou et al., 2006; Ntoutsi et al., 2009; Oliveira e Gama, 2010c; Ntoutsi et al., 2011; Oliveira e Gama, 2012; Pereira e Mendes-Moreira, 2016)(Ver Secção 4). Entre as abordagens existentes, duas delas se destacam por serem exploradas em diversos áreas de aplicação, e ambas buscam as cinco possíveis transições apresentadas acima. São elas:

- *enumeração*: nesta abordagem as transições entre grupos são monitoradas em diferentes janelas de eventos $w_1, w_2, w_n, ...$, onde a cada janela w_n os objetos são novamente agrupados para que as variações em grupos existes e novos grupos possam ser monitorados. Assim, as transições podem ser detectadas mesmo quando ocorrem mudanças na distribuição dos dados. Neste caso, os objetos de cada grupo são monitorados visando encontrar sobreposição

de grupos, grupos similares, ou novos grupos. Por exemplo, se um grupo formado em uma janela w_2 possuir ao menos metade das instâncias do grupo formado na janela w_1, considera-se uma homogeneidade (um *match*), indicando que os dois grupos são similares. Além disso, as demais transições entre grupos são exploradas com base na variação dos objetos entre os grupos formados a cada janela w_n (Spiliopoulou et al., 2006; Ntoutsi et al., 2009, 2011).

- *compreensão*: nesta abordagem os grupos formados em uma janela w_n são caracterizados por diferentes estatísticas, por exemplo, raio, circunferência e densidade. Tal abordagem busca encontrar uma similaridade entre grupos formados em janelas distintas por meio da descoberta de sobreposição de regiões entre tais grupos com o uso de suas estatísticas. Neste sentido, tal processo busca descobrir se um grupo formado em uma janela w_1 continua existindo em uma janela posterior w_2, o que pode ocorre quando grupos apresentam alguma transposição dado suas circunferências. Além disso, as diferentes transições podem ser determinadas com a investigação de estatísticas dos grupos formados a cada janela w_n (Oliveira e Gama, 2010c, 2012; Pereira e Mendes-Moreira, 2016).

3.4 Considerações Finais do Capítulo

Neste capítulo foram apresentados os conceitos relacionados à Fluxo Contínuo de Dados (FCD). Primeiramente, as principais características destes FCDs e a forma como as tarefas de Mineração de Dados e Aprendizado de Máquina podem ser aplicadas neste tipo de cenário, por meio de Fases e Janelas de Eventos. Além disso, foram apresentados detalhes sobre Detecção de Novidade, em destaque Mudança e Evolução de conceito, que devem ser investigados em FCD, principalmente onde classes são desconhecidas.

Nesta tese é investigado o uso de aplicativos em dispositivos móveis em um cenário real de FCD, que possui características muito importantes. Além disso, diferentes fases, sumarização dos dados e técnicas de janelas de eventos devem ser empregadas para que seja possível realizar um acompanhamento dos perfis de uso que são buscados por meio da tarefa de Agrupamento em tais FCDs de dispositivos móveis. Mais ainda, é necessário um acompanhamento dos conceitos (perfis) e também dos comportamento dos consumidores que evoluem ao longo do tempo. Desse modo, técnicas de Detecção de Novidade para monitorar mudanças em perfis e comportamentos de consumidores devem ser utilizadas permitindo que empresas fabricantes de dispositivos móveis possam utilizar tal conhecimento na tomada de decisões.

Por fim, no Capítulo 4 são apresentados os trabalhos relacionados com esta pesquisa, os quais foram identificados por meio de uma revisão sistemática da literatura. Em tal Capítulo são discutidos trabalhos que buscam identificar perfil de uso e trabalhos que visam monitorar tais conceitos ao longo do tempo.

4. TRABALHOS RELACIONADOS

Este capítulo tem como objetivo apresentar a formalização de uma Revisão Sistemática da Literatura e demonstrar os resultados obtidos. Este tipo de revisão tem como objetivo identificar, avaliar e interpretar trabalhos relevantes e disponíveis sobre uma questão de pesquisa específica. Além disso, busca-se externar lacunas identificadas e revelar novas pesquisas a serem investigadas nesta área de pesquisa (Kitchenham, 2004). Em geral, são detalhados os trabalhos relacionados em termos de contribuição para o objetivo proposto por esta pesquisa, suas vantagens e desvantagens, sustentando o desenvolvimento desta tese. A Seção 4.1 descreve o planejamento de tal revisão, com a apresentação de questões de pesquisa e metodologias de buscas de artigos. Na Seção 4.2 são apresentados os critérios de inclusão e exclusão de artigos, bem como perguntas a serem respondidas por meio da leitura dos artigos selecionados. A Seção 4.3 descreve a análise dos resultados obtidos em relação as etapas da revisão sistemática. Na Seção 4.4 é apresentada uma avaliação geral sobre os trabalhos selecionados ao longo de tal revisão. Por fim, a Seção 4.5 apresenta as considerações finais do Capítulo.

4.1 Planejamento da Revisão Sistemática

Esta revisão sistemática segue o protocolo proposto por Kitchenham (2004), o qual é utilizado pela ferramenta *StArt* (State of the Art through Systematic Review) (Fabbri et al., 2016). A versão atual da ferramenta *StArt* suporta os três passos do processo de revisão propostos por Kitchenham (2004). São eles, *Planejamento*, onde os pesquisadores elaboram o protocolo; *Execução*, onde os pesquisadores devem buscar, adicionar e avaliar os estudos encontrado realizando uma revisão sistemática e Sumarização, onde gráficos e tabelas são gerados visando mostrar uma visão geral da revisão realizada e ajudando a descrever o estado da arte do tema de pesquisa. Por sua vez, o passo de *Execução* é realizado em duas etapas, *Etapa I - Seleção*, onde são selecionados os estudos mais relevantes e *Etapa II Extração*, onde são extraídas as principais informações dos estudos relevantes ao tema de pesquisa. Neste sentido, a Figura 4.1 apresenta uma captura de tela da ferramenta *StArt* durante o processo de revisão realizado nesta pesquisa.

Em resumo, realizou-se uma busca sistemática por estudos publicados em anais de congressos e revistas científicas, limitando-se a estudos com a língua inglesa, mas sem restrições de status de publicação ou ano de publicação. A data da pesquisa mais recente foi primeiro de maio de 2018 e os passos e a execução do protocolo são apresentados nas próximas Seções.

4.1.1 Objetivo e Questões Norteadoras

O Objetivo principal desta revisão é identificar o estado da arte nas pesquisas acerca da identificação e monitoramento de perfis e comportamentos por perfil de uso. Assim, buscou-se

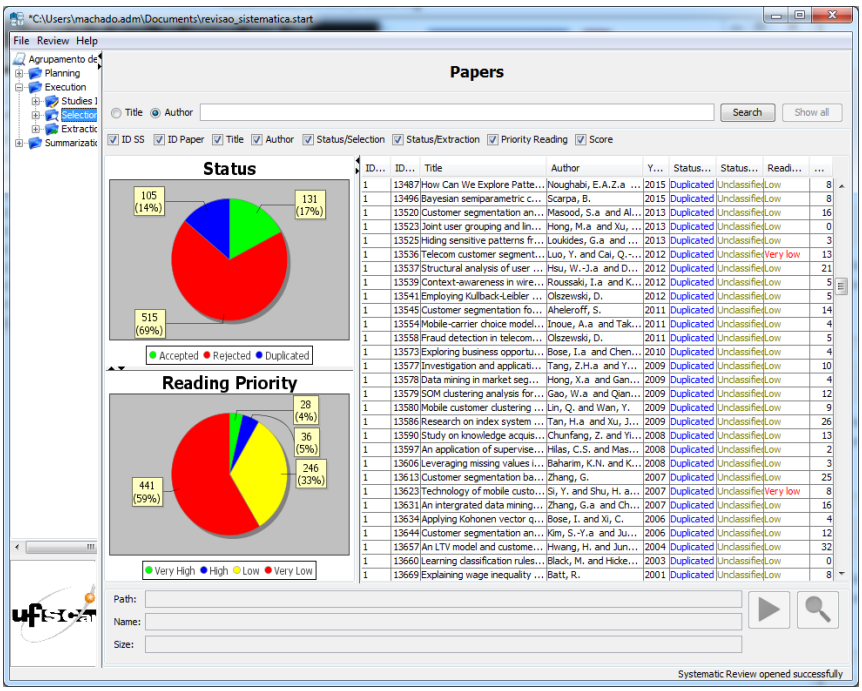


Figura 4.1: Ferramenta *StArt* (Fabbri et al., 2016) utilizada para a realização da revisão sistemática.

identificar soluções completas voltadas para a realização destas atividades com base em algumas questões norteadoras, as quais são:

1. Quais são as soluções existentes voltadas para o auxilio à identificação e ao monitoramento de perfis de uso (grupos)?
2. Quais são os tipos de abordagens utilizadas nestas soluções?
3. Qual a forma de validação destas soluções?
4. Quais conjuntos de dados são utilizados por tais soluções?
5. Como os métodos e técnicas evoluíram no decorrer dos anos, principalmente na área de FCD de dispositivos móveis?
6. Quais os pontos fortes e fracos destas soluções?

4.1.2 Artigos de Controle

Artigos de controles são utilizados para dar maior garantia aos resultados gerados por uma *string* de busca e para iniciar a técnica de *Snowballing* (Wohlin, 2014). Estes artigos devem aparecer nas consultas em bases de artigos científicos e caso não sejam encontrados a *string* deve ser ajustada. Os artigos (Chu et al., 2007) e Hamka et al. (2014) foram escolhidos como artigos de controle por serem mais referenciados e por terem suas abordagens comparadas e discutidas ao longo dos últimos anos.

4.1.3 Busca de Artigos

A busca de artigos foi realizada com o uso da técnica de *Snowballing* (Wohlin, 2014) e com pesquisas em bases de artigos científicos.

Busca por *Snowballing*

O *Snowballing* tem como objetivo identificar trabalhos relacionados anteriores e posteriores, que não tenham sido obtidos a partir da *string* de busca e que sejam relacionados ao objetivo desta revisão. O processo adotado para esta busca segue o que é proposto Wohlin (2014), e pode ser visualizado na Figura 4.2.

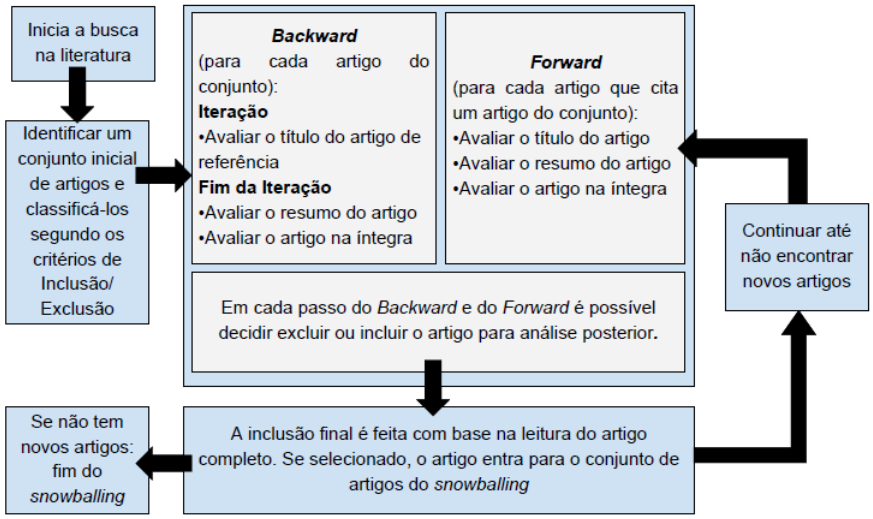


Figura 4.2: Processo de *Snowballing*. Adaptado de Wohlin (2014).

A fase de *Backward* é realizada a partir da leitura direta dos metadados presentes nas referências dos artigos, enquanto a fase de *Forward* é realizada com o auxilio do Google Scholar[1], onde são buscados os trabalhos que citaram um determinado artigo.

Busca em bases de artigos científicos

As bases escolhidas para a realização desta revisão foram: *Scopus*, *Science Direct*, *ACM Digital Library*, *IEEEXplore* e *Web of Science*. Além disso, uma string *PICo* foi formalizada, a qual é uma forma de realizar buscas focadas e é descrita da seguinte forma: *P* é *População* (por exemplo, subscriber clustering), *I* é *Intervenção* (por exemplo, usage pattern), and *Co* é *Contexto/Consequência* (por exemplo, framework). Assim, a *string* de busca final utilizada é apresentada a seguir:

((“subscriber clustering” OR “subscriber profiling” OR “subscriber grouping” OR “subscriber segmentation” OR “customer clustering” OR “customer profiling” OR “customer grouping” OR

[1] https://scholar.google.com.br/

"customer segmentation" OR "user clustering" OR "user profiling" OR "user grouping" OR "user segmentation") AND ("app usage" OR "mobile data" OR "call data record" OR CDR OR telecommunication OR "usage pattern" OR "behavior" OR "service usage" OR "pattern recognition" OR pattern OR usage) AND (method OR framework OR tool OR algorithm OR approach OR model OR strategy))

As palavras-chave contidas nesta *string* de busca tem como base a literatura sobre o tema e as palavras-chave comumente utilizadas em artigos:

4.2 Seleção de Artigos

A seleção de artigos é realizada em duas etapas. Na primeira etapa, chamada de Seleção, os artigos são avaliados com base nos dados do título, resumo e palavras-chave. Além disso, critérios de inclusão e exclusão são aplicados:

1. Critérios de inclusão:
   (a) Artigos que apresentem soluções de identificação ou de monitoramento de perfis de uso (grupos). As soluções encontradas devem incluir técnicas, métodos, modelos, estratégias, algoritmos ou qualquer outra iniciativa relacionada à identificação ou monitoramento de perfil de uso;
   (b) Artigos devem ser completos, incluindo: ano de publicação, conjuntos de dados, critérios de avaliação e resultados experimentais.
2. Critérios de Exclusão:
   (a) Artigos que não sejam completos (ex: *short paper*);
   (b) Artigos relacionados a identificação e monitoramento de perfis de uso, mas que não apresentem uma nova solução para este objetivo (ex: artigos que comparam soluções);
   (c) Trabalhos que indiquem desafios e direções futuras para a área de pesquisa em questão.

Na segunda etapa, chamada de Extração, é realizado o refinamento da revisão a partir da leitura do texto dos artigos selecionados durante a fase anterior. Assim, para cada artigo são respondidos os seguintes critérios que visam a avaliação da qualidade do estudo:

1. Quais métodos são utilizados no processo de identificação perfis de uso (grupos)?
2. Quais técnicas são aplicadas no processo de monitoramento destes perfis?
3. Qual a área de aplicação?
4. Quais tipos de conjuntos de dados são abordados?

5. Como avalia o desempenho da solução?

Essa fase do processo de revisão sistemática é essencial para garantir sua consistência e capturar percepções sobre sua viabilidade. O protocolo desta revisão sistemática visa dar rigor à metodologia. Nesse sentido, tal protocolo foi desenvolvido para orientar a revisão sistemática, pois permite a reutilização dos resultados obtidos. Além disso, outros pesquisadores interessados no mesmo assunto podem executá-lo, por exemplo, para julgar quão adequado é o protocolo, ou mesmo expandi-lo.

4.3 Análise dos Resultados da Revisão

Nesta seção são apresentados os trabalhos relacionados finais que foram extraídos por meio da revisão da literatura com a aplicação do protocolo apresentado anteriormente. Além disso, algumas considerações e resultados obtidos durante toda revisão são apresentados e discutidos.

Dada a investigação de trabalhos em cada um das bases de artigos científicos, é possível observar que o número de estudos encontrados apresentou disparidade entre as bases utilizadas (Ver Figura 4.3). Em relação ao tema, observa-se um crescimento no interesse em soluções voltadas para o problema abordado por esta pesquisa. Desde 2003, o número de estudos publicados vem crescendo. Especialmente nos anos de 2007 e 2015, este crescimento fica bastante evidente, como é possível observar na Figura 4.4.

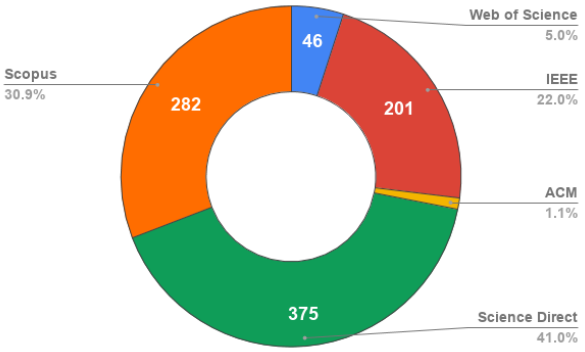


Figura 4.3: Quantidade de artigos retornados e suas respectivas porcentagens para cada base de artigos científicos.

Em relação às etapas da revisão sistemática, a Figura 4.5 mostra o número de estudos retornados, selecionados (Etapa I) e extraídos (Etapa II), bem como o número de estudos ao final da realização desta revisão. Na Etapa I, a pesquisa nas bases de artigos científicos recuperou 914 estudos, onde 95 estudos foram selecionados. No mesmo estágio, durante o processo de *Snowballing*, que começou com os dois artigos de controle, foram selecionados 44 estudos. Na Etapa II, 35 estudos foram extraídos a partir de estudos selecionados pela busca nas bases utilizadas. Por outro lado, 14 estudos foram extraídos de estudos selecionados pelo processo de *Snowballing* nesta mesma etapa.

Figura 4.4: Representação da evolução do tópico de pesquisa ao longo dos anos de acordo com o número de artigos encontrados por ano de publicação.

Ao final, depois de alguns estudos terem sido removidos, incluindo estudos duplicados, 20 estudos foram selecionados.

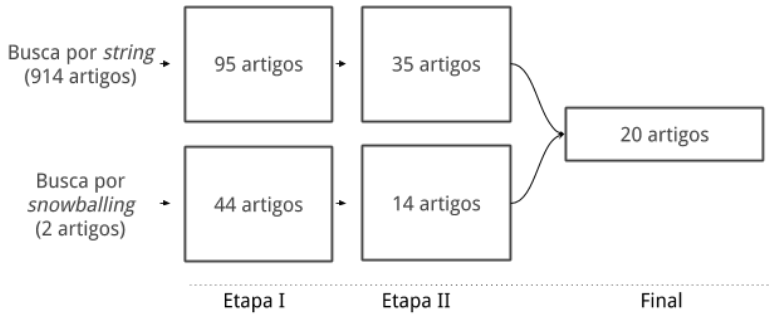


Figura 4.5: Resumo das etapas da Revisão Sistemática.

A maioria dos estudos selecionados ao final da extração não são aplicados em conjuntos de dados com dados de uso de aplicativos em dispositivos móveis. Contudo, estes trabalhos são atuais e apresentam importantes direções que podem ser abordadas ao cenário proposto nesta pesquisa. O grande número de artigos encontrados pela busca por *string* deve-se pelas palavras-chave utilizadas. Contudo uma busca mais restrita não retorna uma quantidade relevante de trabalhos. Além disso, palavras relacionadas a FCD ou a técnicas aplicadas a este tipo de cenário também não foram utilizadas por não serem comumente utilizadas. A busca por *Snowballing* levou a artigos mais concentrados na tarefa de monitoramento em cenários de FCD. Tais trabalhos motivados por Spiliopoulou et al. (2006) e Oliveira e Gama (2010c) focam no monitoramento de conceitos (perfis) em tal cenário.

4.3.1 Soluções Propostas para Identificação e Monitoramento de Perfis de Uso.

Foram selecionados 20 artigos que apresentam soluções para identificação e monitoramento de perfis de uso. A Tabela 4.1 apresenta a listagem final dos artigos selecionados com suas respectivas informações de: referência, ano de publicação e tipo de aplicação, dados os objetivos propostos. O tipo I indica estudo voltado somente a tarefa de identificação de perfis de uso (Seção 4.3.1), enquanto o tipo M significa que o estudo busca a identificação e o monitoramento de tais perfis (Seção 4.3.1). Os estudos selecionados foram encontrados em 17 bases de dados diferentes, indicando que o tema abordado por esta revisão possui significativa receptividade tanto em periódicos científicos como em anais de congressos (Ver Figura 4.6). Além disso, as conferências e periódicos em que os 20 manuscritos foram publicados estão listados na Tabela 4.2. Neste contexto, a revista Expert Systems with Applications é a única que apresentou mais de uma (3) publicação obtida por esta revisão.

Referência	Título	Ano	Tipo
(Ballea et al., 2013)	The architecture of a churn prediction system based on stream mining	2013	I
(Chu et al., 2007)	Toward a hybrid data mining model for customer retention	2007	I
(Hamka et al., 2014)	Mobile customer segmentation based on smartphone measurement	2014	I
(Hsu et al., 2012)	Segmenting customers by transaction data with concept hierarchy	2012	I
(Li e Deng, 2012)	Customer Churn Prediction of China Telecom Based on Cluster Analysis and Decision Tree Algorithm	2012	I
(Lauschke e Ntoutsi, 2012)	Monitoring user evolution in twitter	2012	M
(Ntoutsi et al., 2011)	Summarizing cluster evolution in dynamic environments	2011	M
(Oliveira e Gama, 2010a)	Bipartite graphs for monitoring clusters transitions	2010	M
(Oliveira e Gama, 2010b)	Understanding clusters evolution	2010	M
(Oliveira e Gama, 2010c)	Mec - Monitoring clusters' Transitions	2010	M
(Pereira e Mendes-Moreira, 2016)	Monitoring clusters in the telecom industry	2016	M
(Rehman e Raza Ali, 2015)	Customer churn prediction, segmentation and fraud detection in telecommunication industry	2015	I
(Rizoiu et al., 2015)	Cluspath: a temporal-driven clustering to infer typical evolution paths	2015	M
(Sohn e Kim, 2008)	Searching customer patterns of mobile service using clustering and quantitative association rule	2008	I
(Siddiqui et al., 2015)	Predicting the post-treatment recovery of patients suffering from traumatic brain injury (tbi)	2015	M
(Spiliopoulou et al., 2006)	Monic: modeling and monitoring cluster transitions	2006	M
(Siddiqui et al., 2012)	Where are we going? predicting the evolution of individuals	2012	M
Shabana et al. (2016)	A Multi-view Non-parametric Clustering Approach to Mobile Subscriber Segmentation	2016	I
(Zhang, 2007)	Customer segmentation based on survival character	2007	I
(Zhu et al., 2011)	Role defining using behavior-based clustering in telecommunication network	2011	I

Tabela 4.1: Artigos selecionados ao final da revisão sistemática, bem como suas referências, seus títulos, anos de publicação e os tipos.

Ao final desta revisão, foram encontrados 46 autores distintos. Destes, apenas seis possuem mais de uma publicação sobre o assunto, o que mostra muita diversidade em relação aos autores

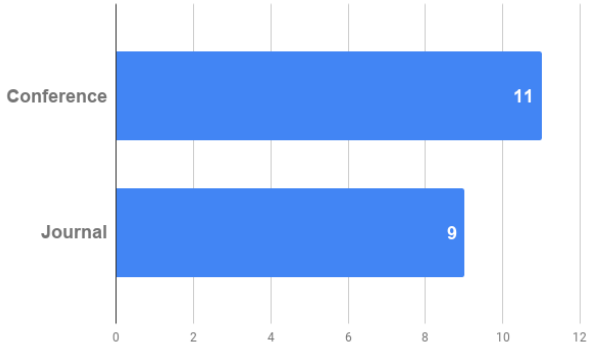


Figura 4.6: Comparação entre os formatos de publicação dos artigos selecionados pela Revisão Sistemática.

Nome	Tipo
IEEE Conference on Business Informatics	Conferência
International Symposium on Intelligent Data Analysis	Conferência
Ubiquitous Data Mining	Conferência
Starting Ai Researchers' Symposium	Conferência
Data Mining and Knowledge Discovery	Revista
Emerging Research in Artificial Intelligence and Computational Intelligence	Revista
ASE Big-Data/SocialInformatics/PASSAT/BioMedCom	Conferência
Wireless Communications, Networking and Mobile Computing	Conferência
Telematics and Informatics	Revista
International Conference on Knowledge Discovery and Data Mining	Conferência
Advances in Information Systems and Technologies	Revista
International Conferência on Advances in Social Networks Analysis and Mining	Conferência
Brain Informatics	Revista
International Symposium on Intelligent Data Analysis	Conferência
Expert Systems with Applications	Revista
International Conference on Computational Science and Its Applications	Conferência
International Conference of the Catalan Association for Artificial Intelligence	Conferência
Knowledge-Based Systems	Journal

Tabela 4.2: Nome das conferências e das revistas científicas onde os artigos finais desta revisão sistemática foram publicados.

que publicaram estudos neste tema de pesquisa. A Tabela 4.3 mostra os nomes de seis autores com pelo menos duas publicações sobre o assunto, juntamente com a afiliação e ordenado pelo número de publicações. Por fim, por meio desta revisão foi possível realizar um mapeamento sistemático da área de estudo. Neste sentido, nas próximas seções os artigos selecionados são explorados com maior profundidade tendo como base as questões definidas no protocolo desta revisão.

Trabalhos que visam a Identificação de Perfis de Uso

Os dez trabalhos selecionados que visam a identificação de perfis de uso (Ballea et al., 2013; Chu et al., 2007; Hamka et al., 2014; Hsu et al., 2012; Li e Deng, 2012; Rehman e Raza Ali,

Autor	Afiliação	Publicações
Spiliopoulou M.	Otto von Guericke University Magdeburg	4
Oliveira M.	University of Porto	4
Gama J.	University of Porto	4
Ntoutsi E.	Leibniz University of Hanover	3
Theodoridis Y.	University of Piraeus	2
Siddiqui Z. F.	Otto von Guericke University Magdeburg	2

Tabela 4.3: Número de publicações, afiliação e o nome dos autores que possuem mais de uma publicação no tema de pesquisa ao final da revisão sistemática.

2015; Sohn e Kim, 2008; Shabana et al., 2016; Zhang, 2007; Zhu et al., 2011) apresentam diferentes abordagens, como o uso de diferentes algoritmos para a tarefa de Agrupamento, buscando identificar diferentes perfis de uso nos cenários abordados. Estes trabalhos investigam em sua maioria, diferentes tipos de dados relacionados a dispositivos móveis e apresentam discussões e comparações que podem ser investigadas no cenário abordado por esta pesquisa.

Ballea et al. (2013) propõem uma arquitetura de duas etapas para a predição de *churn* de usuários. Esta arquitetura visa, em uma de uas etapas, identificar perfis de uso por meio da tarefa de Agrupamento. Além disso, tal tarefa é executada usando diferentes algoritmos e a definição do melhor algoritmo não é apresentada adequadamente. Posteriormente, em uma etapa de treinamento, os perfis de usuários que abandonaram a marca estudada são utilizados como aprendizado visando melhorar a previsão de futuros abandonos. A proposta de Ballea et al. (2013) foi projetada para ser executado em cenários de FCD de dispositivos móveis, onde eventos de uso de tais dispositivos, oriundos de empresas de telecom, são monitorados. Nesse sentido, dada a privacidade dos dados reais da empresa patrocinadora, Ballea et al. (2013) utilizaram um conjunto de dados simulado. Além disso, um gerador de dados que simula eventos obtidos por essa empresa também é apresentado. Em resumo, os eventos simulados combinam dados distintos, como CDR, pessoal (por exemplo, sexo e idade) e faturamento (por exemplo, impostos pagos).

Chu et al. (2007) propõem uma abordagem de Mineração de Dados que é híbrida e focada na retenção de usuários. Tal proposta contém dois modelos, um para aprender e outro para prever. O objetivo do modelo de aprendizado é detectar perfis (grupos) de acordo com os atributos mais significativos do conjunto de dados. Para esse fim, são criadas políticas de retenção de usuários, que são aplicadas no modelo de previsão, no caso de novos usuários apresentarem probabilidade de abandono relevante. Chu et al. (2007) usam o algoritmo de agrupamento hierárquico GHSOM[2]. Esse algoritmo faz sua própria determinação do melhor número de grupos a serem formados. Além disso, Chu et al. (2007) fazem uso de um conjunto de dados de 65.516 usuários reais para testar sua abordagem. Tal conjunto inclui registro de chamadas (CDR), faturamento e dados pessoais dos usuários. Além disso, informações de um ano de cada usuário são resumidas a partir dos dados de uma empresa de telecom real.

[2]http://www.ifs.tuwien.ac.at/ andi/ghsom/description.html

Hamka et al. (2014) apresentam uma nova abordagem que visa segmentar usuários de dispositivos móveis com base em métricas relevantes de uso em diferentes perspectivas de negócio (ex: operadora de rede, desenvolvedores de aplicativos). Tal abordagem de segmentação tem como principal objetivo compreender os estilos de vida de diversos usuários dado o uso de dispositivos móveis por tais usuários. Para este fim, Hamka et al. (2014) utilizam a ferramenta *Latent Gold 4.5* para realizar a tarefa de Agrupamento. Além disso, são utilizadas várias medidas de avaliação buscando o melhor número de perfis. No entanto, Hamka et al. (2014) discutem apenas os resultados obtidos por meio da aplicação do algoritmo de agrupamento *Expectation Maximization* (EM), sem incluir outras abordagens disponíveis em tal ferramenta. Para mostrar a aplicabilidade de sua abordagem, Hamka et al. (2014) utilizam um único conjunto de dados do tipo CDR combinado com dados do Serviço de Mensagens Curtas (SMS), tráfego de dados da internet, número de aplicativos instalados e número de aplicativos usados pelos usuários. Todo tipo de dado é oriundo de 129 usuários reais. Além disso, um total de 130.000 eventos, contendo a média dos respectivos dados, foram capturados durante um período de 28 dias.

Hsu et al. (2012) propõem uma nova metodologia que investiga hierarquias de itens utilizados visando identificar similaridade entre usuários com base no consumo de itens. A novidade deste trabalho está em avaliar os itens consumidos de uma forma hierárquica que ajuda na definição da similaridade entre os usuários. Esta similaridade é aplicada em uma tarefa de Agrupamento que visa segmentar tais usuários em diferentes perfis. Hsu et al. (2012) avaliam a tarefa de Agrupamento por meio de três medidas de avaliação, incluindo *SWC*, em busca do melhor número de grupos a serem formados. Além disso, Hsu et al. (2012) fazem uma comparação de sua abordagem em relação à aplicação de algoritmos hierárquicos tradicionais. Hsu et al. (2012) obtiveram melhores resultados para todos números de grupos utilizados na aplicação de sua abordagem proposta em comparação as abordagens tradicionais. Para tais testes, Hsu et al. (2012) utilizaram um conjunto de dados contendo informações sobre o uso de livros em uma biblioteca real. Os eventos deste conjunto de dados foram capturados durante um período de 3 meses, de 01/01/2009 a 31/03/2009. Os dados de tais transações contêm informações de usuário (por exemplo, identificador), de livros (por exemplo, título e categoria) e de retirada de livros (por exemplo, data e hora). Um total de 385 transações são usadas durante o estudo, e Hsu et al. (2012) afirmam que esse tipo de transação é muito semelhante a transações que ocorrem em outros tipos de negócios do mundo real.

Li e Deng (2012), investigam as tarefas de Agrupamento e de Árvores de Decisão visando prever o abandono de usuários em um cenário de FCD de dispositivos móveis. Visando alcançar tal objetivo Li e Deng (2012) identificam os perfis de uso com a aplicação da tarefa de Agrupamento. As informações destes perfis são utilizadas na tarefa de Árvore de Decisão que tem como objetivo prever o abandono dos usuários. Li e Deng (2012) usam a ferramenta *SAS*[3] para determinar e avaliar o melhor número de grupos com a aplicação do algoritmo *K-Means*. No entanto, Li e Deng (2012) não apresentaram como essa ferramenta avalia o melhor número de grupos. Para realizar seus experimentos, Li e Deng (2012) capturaram eventos de uso de dispositivos móveis de uma empresa

[3]https://www.sas.com/en_us/home.html

chinesa de telecom durante 7 meses. Tais dados foram capturados de Junho a Dezembro de 2011. Vários registros, como CDR, dados pessoais, dados de faturamento e dados de hardware, compõem tal conjunto. Li e Deng (2012) investigaram um total de 2.155 usuários reais, embora não tenham mencionado o número de eventos capturados no período monitorado.

Rehman e Raza Ali (2015) também estudam a predição de *churn* de usuários. Para este fim, Rehman e Raza Ali (2015) buscam segmentar usuários e identificar fraudes no cenário abordado de modo a contribuir com tal predição. Rehman e Raza Ali (2015), aplicam a metodologia proposta segmentando os usuários em diferentes perfis de uso visando encontrar produtos e serviços que possam ser oferecidos em caso de predição de *churn* ser positiva para tais usuários. Neste estudo, dois algoritmos para a tarefa de Agrupamento são usados, *K-Means* e *Two-Step*. Nesse sentido, Rehman e Raza Ali (2015) predizem a rotatividade de usuários separadamente com cada um dos algoritmos e ao final escolhem aquele com a melhor precisão. Para seus experimentos, Rehman e Raza Ali (2015) fazem uso de um FCD produzido pela captura diária de transações de usuário, o qual contém dados de CDR, faturamento e informações pessoais dos usuários. Esse FCD resume o uso diário desses usuários em mais de 6 meses de análise de dados. No total, 31.964 usuários foram monitorados. Além disso, Rehman e Raza Ali (2015) abordaram um cenário de FCD com Janelas de Eventos baseadas em *timestamp*. No entanto, essas janelas têm tamanhos diferentes ao longo do período analisado.

Sohn e Kim (2008) apresentam um trabalho que visa identificar padrões de uso de usuários por meio da tarefa de Mineração de Regras de Associação. Esta tarefa é realizada após a identificação de perfis de usuários por meio da tarefa de Agrupamento. Com os perfis identificados, padrões são obtidos e utilizados visando caracterizar cada perfil encontrado. Sohn e Kim (2008) comparam o desempenho de diferentes algoritmos para a tarefa de Agrupamento combinando estes com técnicas de pré-processamento de dados. Além disso, Sohn e Kim (2008) definem o melhor número de grupos principalmente comparando a distribuição de usuários nos grupos obtidos. Para executar seus testes, Sohn e Kim (2008) abordam um conjunto de dados de CDR disponibilizado por uma empresa de telecom da Coréia do Sul, o qual contêm eventos capturados durante três meses. Os dados foram capturados de Março a Maio 2001 de 17.000 usuários reais. No entanto, Sohn e Kim (2008) não mencionam o número total de eventos capturados.

Shabana et al. (2016) propõem uma abordagem de visualização múltipla para a segmentação de usuários. Para este fim, Shabana et al. (2016) apresentam um novo algoritmo para a tarefa de Agrupamento, em múltiplas escalas, denominado *PD-means*. Este algoritmo visa segmentar cada tipo de informação presente em um conjunto de dados separadamente em escalas, sendo bastante distinto dos algoritmos tradicionais para tal tarefa. No entanto, Shabana et al. (2016) não mencionam qual processo realizado para definir o melhor número de grupos a serem formados, assim como não fazem qualquer comparação do novo algoritmo com outras abordagens tradicionais. Ao final, os grupos gerados para cada característica são apresentas por uma nova abordagem de visualização. Tal abordagem busca mostrar de forma eficiente as diferentes características que podem existir na segmentação de usuários. Para seus testes, Shabana et al. (2016) fazem uso de um conjunto de

dados de um dos maiores operadores de telecom da Ásia. Este conjunto de dados consiste em CDR, tráfego pessoal, de faturamento e de dados. Além disso, esse conjunto de dados consiste em aproximadamente 700.000 usuários reais, não sendo mencionado o número de eventos que compõem tal conjunto de dados.

Zhang (2007) apresenta um *framework* que visa segmentar usuários em perfis de uso baseado em técnicas de Mineração de Dados combinadas à técnicas de análise de sobrevivência. A combinação de tais técnicas visa monitorar as característica de usuários por um determinado período buscando identificar posições no tempo em que tais características ocorrem com mais frequência. Em resumo, esta análise procura identificar as principais características dos usuários e sua evolução. Além disso, aprender essa evolução é o principal objetivo para identificar indicadores úteis para a tarefa de Agrupamento. Para este fim, Zhang (2007) usa a ferramenta *SPSS*[4] para executar o algoritmo de agrupamento *K-Means*. No entanto, Zhang (2007) não mostra como o melhor número de grupos gerados é avaliado e a abordagem proposta não é comparada a outros estudos. Por outro lado, Zhang (2007) discute os resultados obtidos para demonstrar as diferenças entre os grupos obtidos. Zhang (2007) usa um conjunto de dados que consiste em CDR, informações pessoais e de faturamento de 1.000 usuários reais para realizar seus testes. Tal conjunto foi fornecido por uma grande empresa de telecom da China.

Zhu et al. (2011) propõem um novo método que visa detectar padrões de usuários e segmentá-los em diferentes perfis de uso. Para este fim, o método proposto por Zhu et al. (2011) faz uso de diferentes métricas que são combinadas visando melhorar a obtenção da similaridade entre usuários e a definição de tais perfis, a qual é realizada por meio de uma tarefa de Agrupamento. Zhu et al. (2011) utilizam o algoritmo de agrupamento *K-Means*. Além disso, Zhu et al. (2011) utilizam a medida de avaliação *BIC* para determinar o melhor número de grupos a serem formados. Da mesma forma que Zhang (2007), Zhu et al. (2011) conduzem uma discussão sobre os grupos obtidos mostrando a diversidade de tais grupos. No entanto, Zhu et al. (2011) não comparam sua abordagem com demais estudos. No total, 3 conjuntos de dados são explorados por Zhu et al. (2011). Todos os conjuntos de dados contêm apenas CDR. O primeiro conjunto de dados foi capturado por 10 dias a partir de uma empresa real, enquanto outros dois conjuntos de dados são de outra empresa real, onde ambos são capturados por aproximadamente 3 meses. No entanto, Zhu et al. (2011) não mencionam explicitamente o número de usuários e eventos presentes em cada um desses conjuntos de dados.

Áreas de Aplicação

Os trabalhos desta seção são aplicados, em sua maioria, em áreas que envolvem o uso de dispositivos móveis. Contudo, os trabalhos são aplicados principalmente na área de telecom.

Ballea et al. (2013), Chu et al. (2007), Hamka et al. (2014) Li e Deng (2012), Rehman e Raza Ali (2015), Sohn e Kim (2008), Shabana et al. (2016), Zhang (2007) e Zhu et al. (2011),

[4] https://spss.en.softonic.com/

apresentam abordagens que são aplicadas na área de telecom. Diversas empresas desta área são investigadas por tais trabalhos. Este tipo de área é muito investigada quando se busca a identificação de perfis de uso. Principalmente quando o objetivo final é investigar e prever a perda de usuários.

O único trabalho que aborda uma área diferente é o de Hsu et al. (2012). A abordagem proposta para segmentação de usuários baseada em hierarquias de transações é aplicada na área de Bibliotecas. Contudo, os Hsu et al. evidenciam que a aplicabilidade da sua abordagem abrange demais áreas em que é possível a obtenção de transações de conjuntos de dados.

Conjuntos de Dados

Similar ao que ocorre na área de aplicação dos estudos investigados nesta seção, os conjuntos de dados, que são usados em experimentos e validação de abordagens, são geralmente compostos de dados derivados de dispositivos móveis. Em geral, tais conjuntos são compostos por CDR, e as vezes combinados com o informações de aplicativos, dados pessoais, de cobrança, de hardware ou de tráfego de dados. Uma comparação entre os conjuntos de dados é apresentada na Tabela 4.4. Nesta tabela é indicado se: os estudos são aplicados em um cenário de FCD, a quantidade de eventos, a quantidade de objetos, o período de tempo e o tipo dos dados analisados pelos estudos.

Referência	FCD	Conjunto de dados			
		Eventos	Objetos	Tempo	Dados
(Ballea et al., 2013)	Sim	-	-	-	*CDR*, pessoais e faturamento
(Chu et al., 2007)	Não	-	65.516	1 ano	*CDR*, pessoais e faturamento
(Hamka et al., 2014)	Não	130.000	129	28 dias	*CDR*, uso de aplicativos e tráfego de dados
(Hsu et al., 2012)	Não	385	-	3 meses	Transações biblioteca
(Li e Deng, 2012)	Não	-	2.155	7 meses	*CDR*, pessoais, faturamento e *hardware*
(Rehman e Raza Ali, 2015)	Sim	-	31.964	6 meses	*CDR*, pessoais e faturamento
(Sohn e Kim, 2008)	Não	-	17.000	3 meses	*CDR*
(Shabana et al., 2016)	Não	-	700.000	-	*CDR*, pessoais e faturamento
(Zhang, 2007)	Sim	-	1.000	-	*CDR*, pessoais e faturamento
(Zhu et al., 2011)	Não	-	-	[10 dias e 3 meses]	*CDR*

Tabela 4.4: Comparação entre os conjuntos de dados utilizados pelos estudos que buscam a identificação de perfis de uso. A coluna FCD indica se o conjunto utilizado é em FCD ou não. As demais colunas apresentam a quantidade de eventos, quantidade de objetos, tempo de captura dos dados e o tipo de dado analisado.

Considerando o número de objetos (usuários), alguns estudos não mencionam este número em seus artigos, são os casos de Ballea et al. (2013), Hsu et al. (2012) e Zhu et al. (2011). No entanto, diferentes números de objetos são investigado por outros estudos. Apesar do conjunto de dados utilizado por Hamka et al. (2014) possuir um grande número de eventos, os dados utilizados são de apenas 129 objetos, sendo este o menor número analisado entre os estudos. Por usa vez, Zhang (2007), Li e Deng (2012) e Sohn e Kim (2008) investigam respectivamente 1.000, 2.155 e 17.000 objetos. Além disso, Rehman e Raza Ali (2015) analisam 31.000 objetos enquanto Chu

et al. (2007) investigam 65.516 objetos em seu conjunto de dados. Chu et al. (2007) também apresentam o maior período de tempo para captura de dados (1 ano). Por outro lado, o maior número de objetos investigados é apresentado em no estudo de Shabana et al. (2016) (700.000). No entanto, Shabana et al. (2016) não mencionam o número de eventos que compõem tal conjunto.

Por fim, Hsu et al. (2012) faz uso de um conjunto de dados sobre o uso de livros em bibliotecas enquanto todos os demais estudos apresentam o uso de dados do tipo CDR. Nesse sentido, Zhu et al. (2011) e Sohn e Kim (2008) abordam apenas um conjunto com CDR. Por outro lado, os demais estudos combinam diferentes tipos de dados com CDR. Dados pessoais e de faturamento são mais os comuns e são combinados pelos estudo de Ballea et al. (2013), Chu et al. (2007), Rehman e Raza Ali (2015), Shabana et al. (2016) e Zhang (2007). Contudo, CDR também são combinados com informações de aplicativos por Hamka et al. (2014), tráfego de dados também por Hamka et al. (2014) e informações de hardware por Li e Deng (2012).

Validação

Estudos que visam a identificação de perfis têm como principal problema a ausência de classes reais dos objetos em seus conjuntos de dados. A falta deste tipo de classe faz com que os trabalhos recorram a medidas de avaliação de agrupamentos buscando demonstrar que os resultados obtidos são satisfatórios. Além disso, com a indisponibilidade de conjuntos de dados tipicamente devido à proteção de dados pessoais, nenhum dos artigos analisados compara seus resultados com outras abordagens focadas no mesmo cenário, como a área de telecom. Mais ainda, a maioria dos estudos mostrados nesta seção não realiza uma avaliação adequada ou uma comparação de suas abordagens. Isso pode ser observado na comparação entre a forma de validação das abordagens que é apresentado pela Tabela 4.5. Em tal Tabela são apresentados os algoritmos utilizados para a tarefa de Agrupamento agrupamento, as medidas utilizadas para avaliar o melhor número de grupos, o tipo de validação e os critérios de comparação aplicados.

Diferentes algoritmos para a tarefa de Agrupamento são utilizados pelos estudos investigados nesta seção. Alguns usam algoritmos hierárquicos tradicionais enquanto outros aplicam algoritmos particionais. Entre eles, se destaca o algoritmo particional *K-means*, que é o mais utilizado e está presente nos trabalhos de Li e Deng (2012), Rehman e Raza Ali (2015), Sohn e Kim (2008), Zhang (2007) e Zhu et al. (2011). De todos trabalhos investigados, os estudos de Rehman e Raza Ali (2015) e Sohn e Kim (2008) são os únicos que apresentam resultados nos quais mais de um algoritmo foi empregado. No estudo de Rehman e Raza Ali (2015), dois algoritmos são usados, *K-Means* e *Two-Step*, enquanto que no trabalho de Sohn e Kim (2008) cinco algoritmos são abordados, são eles *K-means*, *Average*, *Ward*, *Centroid* e *Midrange*).

Outros algoritmos para a tarefa de Agrupamento também são abordados pelos estudos, são eles *GHSOM*, em Chu et al. (2007) e *EM* no trabalho de Hamka et al. (2014). Por outro lado, Hsu et al. (2012) apresentam um novo algoritmo hierárquico, bem como Shabana et al. (2016) que apresentam outro novo algoritmo chamado *DP-Means*. Além disso, Hsu et al. (2012) são os únicos

Referências	Detecção de Perfis		Validação	Comparação
	Algoritmos	Medidas		
(Ballea et al., 2013)	-	-	Qualitativa	-
(Chu et al., 2007)	*GHSOM*	próprio algoritmo	Qualitativa	-
(Hamka et al., 2014)	*EM*	*BIC*	Quantitativa	-
(Hsu et al., 2012)	Novo método hierárquico	*SWC*, *C* e *Average Index*	Quantitativa	Algoritmos hierárquicos tradicionais
(Li e Deng, 2012)	*K-Means*	-	Qualitativa	-
(Rehman e Raza Ali, 2015)	*K-Means* e *Two-Step*	Acurácia	Quantitativa	-
(Sohn e Kim, 2008)	*K-means*, *Average*, *Ward*, *Centroid* and *Midrange*	Distribuição dos objetos	Quantitativa	-
(Shabana et al., 2016)	*DP-means*	-	Quantitativa	-
(Zhang, 2007)	*K-Means*	-	Quantitativa	-
(Zhu et al., 2011)	*K-Means*	*BIC*	Quantitativa	-

Tabela 4.5: Comparação entre formas de validação dos estudos que buscam a identificação de perfis de uso. Os algoritmos utilizados para a tarefa de Agrupamento, as medidas de avaliação de grupos, o método de validação e os critérios de comparação aplicados.

a realizar uma comparação de sua abordagem. Nesse sentido, Hsu et al. (2012) comparam sua abordagem hierárquica a alguns algoritmos hierárquicos tradicionais. Por fim, Ballea et al. (2013) utilizam a ferramenta *MOA* (Massive Online Analysis) para realizar seu estudo. Apesar de Ballea et al. (2013) mencionar a identificação do perfil dos usuários, eles não indicam os algoritmos e as medidas aplicadas.

Considerando a avaliação do melhor número de grupos a serem formados, alguns estudos, como o de Ballea et al. (2013), Li e Deng (2012), Shabana et al. (2016) e Zhang (2007) não relacionam adequadamente quais medidas são utilizadas. Das informações capturas, é possível observar que o *BIC* foi a medida mais utilizada para tal objetivo. No entanto, Hsu et al. (2012) apresentam um estudo onde o maior número de medidas foram aplicadas, são elas *SWC*, *C* e *Average Index*. Por outro lado, Rehman e Raza Ali (2015) e Sohn e Kim (2008) usam formas alternativas de definição de melhor número de grupos Rehman e Raza Ali (2015) investigam a precisão de seus resultados. Para esse fim, Rehman e Raza Ali (2015) utilizaram dois algoritmos para diferentes número de grupos e compararam os resultados obtidos com a a classe real do seu conjunto de dados. Por sua vez, Sohn e Kim (2008) definem o melhor número de grupos principalmente pela comparação da distribuição dos usuários nos grupos obtidos. Por fim, Chu et al. (2007) utilizam o algoritmo hierárquico *GHSOM* que faz a sua própria determinação do melhor número de grupos a serem formados.

Trabalhos com o objetivo final de predizer o abandono de usuários tentam realizar uma validação dos seus modelos de predição, porém não se preocupam em validar seus perfis obtidos. Neste sentido, trabalhos como de Ballea et al. (2013) e Rehman e Raza Ali (2015) apresentam uma validação de suas propostas somente em relação ao seu modelo de predição de *churn*. No entanto, a proposta de ambos trabalhos não é comparada a outras abordagens. Da mesma forma, os demais estudos não fazem comparação de suas abordagens com outras estratégias (Shabana et al., 2016;

Zhang, 2007; Zhu et al., 2011; Hamka et al., 2014; Chu et al., 2007; Hsu et al., 2012; Li e Deng, 2012; Sohn e Kim, 2008). Dessa forma, pode-se afirmar que a maioria dos trabalhos desta seção não realizam uma devida avaliação ou comparação de suas abordagens de identificação de perfis de uso.

Trabalhos que visam o Monitoramento de Perfis de Uso

Nesta seção, são apresentados os outros dez trabalhos que tem como objetivo principal identificar e monitorar grupos (perfis) (Lauschke e Ntoutsi, 2012; Ntoutsi et al., 2011; Oliveira e Gama, 2010a,b,c; Siddiqui et al., 2012; Rizoiu et al., 2015; Siddiqui et al., 2015; Spiliopoulou et al., 2006; Siddiqui et al., 2012). Tais estudos aplicam diferentes técnicas, como técnicas de Detecção de Novidade, que buscam aprimorar o processo de evolução dos conceitos aprendidos ao longo do período de análise dos grupos formados. Estes trabalhos são relevantes e atuais apresentando importantes direções que podem ser investigadas no cenário proposto por esta pesquisa. Uma breve introdução para cada um desses estudos é descrita a seguir e uma comparação entre tais estudos é detalhada nas demais seções.

Spiliopoulou et al. (2006) apresentam um *framework* chamado *MONIC*. *MONIC* tem como objetivo detectar e monitorar transições em grupos identificados ao longo do tempo. Nesse sentido, o *MONIC* executa uma caracterização dos grupos por *enumeração* (ver Seção 3.3). Em resumo, neste tipo de caracterização os grupos são representados por seus elementos (objetos), onde tais objetos são monitorados visando detectar para quais grupos estes elementos se movimentam em Janelas de Eventos posteriores e a fim de encontrar transições que podem ocorrer ao longo do tempo. Spiliopoulou et al. (2006) apresentam o emprego do *MONIC* na área de banco de documentos, visando principalmente a implementação da tarefa de monitoramento. Neste sentido, o trabalho de Spiliopoulou et al. (2006) investiga classes de artigos científicos e suas variações em um determinado período de tempo. Spiliopoulou et al. (2006) realizam experimentos usando sua abordagem de Enumeração comparando os resultados de grupos formados ao longo do tempo com as classes reais conhecidas a priori em seu conjunto de dados. Spiliopoulou et al. (2006) aplicam o *MONIC* em um conjunto de dados de artigos classificados da seção H.2.8 da Biblioteca Digital ACM. Este conjunto de dados contém publicações de diferentes áreas, que são usadas como classes, publicadas entre 1997 e 2004, em um total em 4.920 publicações. Além disso, Spiliopoulou et al. (2006) agora em Spiliopoulou et al. (2013) apresentam resultados de diferentes pesquisas em que o *framework MONIC* é abordado, mostrando que tal proposta é interessante para diferentes cenários de FCDs. Mais ainda, Spiliopoulou et al. (2013) apresentam tais aplicações do *MONIC* em outros tipos de conjuntos de dados. Também é importante citar que o trabalho de Spiliopoulou et al. (2006) foi o pioneiro no monitoramento de agrupamentos.

Após o desenvolvimento do *framework MONIC* outros estudos com o objetivo de monitorar grupos foram apresentados. Oliveira e Gama (2010a) e Oliveira e Gama (2010b) apresentaram estudos iniciais para o desenvolvimento de um novo mecanismo para o monitoramento de grupos e afirmam que tal tarefa é valiosa podendo ajudar na compreensão da evolução de perfis ao longo do

tempo. Logo, Oliveira e Gama (2010c) apresentam um *framework* chamado *MEC* visando caracterizar e monitorar transições de grupos ao longo do tempo. O *framework MEC* investiga diferentes métodos e transições de monitoramento de grupos, o que depende do tipo da técnica de caracterização de destes grupos, bem como algoritmos para detectar mudanças em conceitos aprendidos. O *framework MEC* emprega duas abordagens de caracterização de grupos, por *enumeração* similar ao *MONIC*, e por *ompreensão* onde diferentes medidas são calculadas para evitar o armazenamento de todos os elementos do grupo na memória (por exemplo, densidade e raio) (ver secão 3.3). Para cada um dos tipos de caracterização é apresentado um método de monitoramento de grupos. Para testar suas abordagens, Oliveira e Gama (2010c) aplicaram o *MEC* em áreas de Economia e Estatística, onde os grupos são investigados e monitorados. De fato, Oliveira e Gama (2010c) realizaram vários experimentos com o objetivo de mostrar as progressões de seu método de monitoramento (Oliveira e Gama, 2010a,b). Tais testes foram realizadas em vários conjuntos de dados, a fim de revelar seus pontos fortes e suas limitações. Por exemplo, Oliveira e Gama (2010a) aplicam suas abordagens a dois conjuntos de dados distintos. O primeiro é formado por dados de atividades econômicas de Portugal obtidos durante três anos a partir de 39 objetos. Enquanto, o segundo é composto por eventos estatísticos sobre o desenvolvimento econômico de Portugal (por exemplo, índice de desenvolvimento regional) obtidos ao longo de dois anos e contendo dados de 30 objetos. O segundo conjunto de dados também é utilizado em outros estudos (Oliveira e Gama, 2010b,c) e com outro conjunto de dados que é composto de dados estatísticos de estudantes obtidos durante três anos a partir de 30 objetos.

Em um cenário muito similar Ntoutsi et al. (2011) buscam, além de monitorar mudanças de conceitos, sumarizar de maneira eficiente e em uma representação gráfica tais mudanças. A sumarização proposta, chamada de *FINGERPRINT*, visa melhorar a representação das evoluções dos conceitos que podem ocorrer ao longo do tempo. Nesse sentido, Ntoutsi et al. (2011) aplicam o *FINGERPRINT* em várias áreas, são elas, banco de dados de documentos, assim como Spiliopoulou et al. (2006), redes de computadores e transações. No cenário de redes de computadores, são investigados o acesso e evolução de perfis em tais redes. Na área de transações, são investigados os diferentes perfis de transações apresentados ao longo do tempo por usuários. Em seus experimentos, Ntoutsi et al. (2011) buscam validar o processo de sumarização executado pela *FINGERPRINT* comparando seus métodos de monitoramento e sumarização em vários conjuntos de dados. Entre eles o mesmo conjunto usado por Spiliopoulou et al. (2006) e mais dois. O primeiro consiste em dados de intrusão de redes de computadores do KDD CUP de 1999, contendo 424.021 eventos de intrusão capturados durante duas semanas. O segundo é um conjunto de dados de transações de doação da KDD CUP de 1998, que inclui 95.412 eventos sobre doações. Ntoutsi et al. (2011) perceberam que o conjunto de dados do ACM é desequilibrado, o conjunto de dados de intrusão é altamente dinâmico e o conjunto de dados de doação é razoavelmente estável mas sua abordagem se aplica a todos os conjuntos testados.

Uma vez que as mudanças nos comportamentos dos indivíduos (por exemplo, usuários que compõem grupos monitorados) não é amplamente pesquisada, Siddiqui et al. (2012) apresentam um

novo *framework* visando monitorar o comportamento dos indivíduo, tendo como base o *framework MEC* e o estudo de Ntoutsi et al. (2011). Siddiqui et al. (2012) tem como motivação o fato de que, principalmente quando novos conceitos não são conhecidos a priori, não é possível saber o que irá ocorrer com os indivíduos de um determinado grupo em Janelas de Eventos posteriores. Neste caso, Siddiqui et al. (2012) propõem um novo *framework* capaz de aprender com grupos de indivíduos em diferentes momentos do tempo. Tal abordagem utiliza este aprendizado, realizado em gráficos de transição de estados, para aprender um modelo de *cadeias de Markov*, que é utilizado para prever o próximo grupo em que um individuo será agrupado. Semelhante aos trabalhos de Oliveira e Gama (2010a), Oliveira e Gama (2010b) e Oliveira e Gama (2010c), Siddiqui et al. (2012) propõem o emprego de sua investigação na área econômica. Além de prever o que ocorrerá com os indivíduos em janelas posteriores, Siddiqui et al. (2012) também pretendem monitorar os grupos formados. Em seus experimentos, Siddiqui et al. (2012) abordam um conjunto de dados econômicos europeus de 836 empresas distintas capturadas de 2003 a 2007. Em tais experimentos Siddiqui et al. (2012) buscam evidenciar a previsão para as próximas janelas com a tarefa de monitoramento dos grupos, demonstrando a aplicabilidade de sua proposta no cenário abordado. Além disso, Siddiqui et al. (2012) discutiram os resultados obtidos e os compararam com os resultados produzidos pelo uso do *MEC*.

Em outro trabalho, Lauschke e Ntoutsi (2012) propõem o monitoramento de perfis de usuários do *Twitter* com base em tópicos e termos de interesse do dia-a-dia dos usuários. Esta abordagem visa a detecção de mudanças de perfis, bem como de comportamento destes usuários ao longo do tempo. No enatando, somente três possíveis transições de perfis são abordadas, sobrevivência, desaparecimento e surgimento de novos conceitos. Para caracterizar cada agrupamento e entender o comportamentos de tais usuários é proposto uma abordagem por meio de *TF–IDF* (*Term Frequency–Inverse Document Frequency*) aplicada ao texto publicado visando detectar o tópico de tais textos e ajudar na caracterização dos perfis e no entendimento do comportamento dos usuários. Além disso, esta abordagem trabalha com diferentes Janelas de Eventos de que variam de acordo com a intensidade a intensidade dos eventos produzidos por cada usuário. Buscando destacar a qualidade de sua abordagem, Lauschke e Ntoutsi (2012) realizam alguns experimentos e analisam seus resultados discutindo a evolução dos mesmos durante a execução de suas pesquisas. Neste sentido, Lauschke e Ntoutsi (2012) usam um conjunto de dados do Twitter capturado por aproximadamente 6 meses. No total, 561 eventos de 6 usuários distintos compõem tal conjunto.

Siddiqui et al. (2015) investigam a evolução de perfis de pacientes, antes e depois de tratamentos, visando a predição do avanço da saúde de tais pacientes de acordo com os comportamentos apresentados ao decorrer do tempo. Siddiqui et al. (2015) fazem uso do *framework* proposto em Siddiqui et al. (2012) com alguns ajustes necessários para a área de Medicina. Por exemplo, Siddiqui et al. (2015) investigam a progressão de pacientes durante seus tratamentos. Para validar as previsões do comportamento do paciente, Siddiqui et al. (2015) monitoram os perfis dos pacientes após o tratamento, comparando os resultados obtidos com os dados reais dos pacientes. Essa validação foi possível, uma vez que o conjunto de dados usado em experimentos tem registros

reais de pacientes, bem como tratamentos aplicados e resultados de testes clínicos. Além disso, o conjunto de dados possui informações sobre indivíduos com traumatismo cranioencefálico composto por 29 pacientes.

Por sua vez, Rizoiu et al. (2015) apresentam um novo algoritmo chamado *ClusPath*, que visa inferir as relações entre grupos evidenciando os caminhos (*paths*) que tais grupos percorrem ao longo do tempo. Este estudo tem como principal novidade uma nova medida para avaliar a similaridade entre grupos em um cenário de FCD. Rizoiu et al. (2015) empregam seu algoritmo *ClusPath* em dados políticos e econômicos, bem como Siddiqui et al. (2012) e Oliveira e Gama (2010a,b,c). Para validar sua proposta, Rizoiu et al. (2015) realizam uma comparação da execução da sua abordagem com outros algoritmos, são eles *TDCK-Means* e *Temporal Driven K-Means*. Para este fim, Rizoiu et al. (2015) abordam o mesmo conjunto de dados empregado por Siddiqui et al. (2012). Além disso, Rizoiu et al. (2015) abordam outro conjunto de dados de política que foi capturado entre 1960 e 2009 com dados de 23 países. No geral, os experimentos dos autores demonstram que o *ClusPath* supera consistentemente os outros algoritmos comparados mostrando os caminhos dos grupos analisados ao longo do tempo.

Em um dos trabalhos mais atuais, a tarefa de monitoramento de transições de grupos foi proposta na área de telecom. Pereira e Mendes-Moreira (2016) propõem o monitoramento de transições de grupos de usuários utilizando dados de uma operadora de telecom. Para a realização de tal tarefa, Pereira e Mendes-Moreira (2016) utilizam com base o método de compreensão do *framework MEC* como forma de caracterizar os grupos formados. Além disso, Pereira e Mendes-Moreira (2016) propõem a adição de um novo sumário estatístico, o qual visa melhorar a detecção de transições de grupos no cenários proposto. Como resultado final, Pereira e Mendes-Moreira (2016) mostram que é possível identificar características que afetam a transição de grupos em um cenário real similar ao contexto abordado por esta tese. Para validar sua proposta, Pereira e Mendes-Moreira (2016) comparam os resultados obtidos por sua abordagem com os resultados gerados pela aplicação do *MEC*, discutindo as relações positivas obtidas por sua variante em tal cenário. Para os seus experimentos, Pereira e Mendes-Moreira (2016) abordam um conjunto de dados de CDR capturados por 15 dias em Dezembro de 2012, de aproximadamente 900.000 usuários. De fato, Pereira e Mendes-Moreira (2016) são pioneiros na realização de monitoramento de grupos envolvendo dados de dispositivos móveis, em que os grupos são perfis de uso baseados em dados de ligações entre dispositivos móveis.

Para resumir, a Tabela 4.6 apresenta os mecanismos utilizados pelos estudos descritos nesta seção. Nesse sentido, a Tabela 4.6 mostra o monitoramento do estudo proposto, o qual pode ser de perfis e/ou comportamentos de indivíduos, se estes trabalhos aplicam um processo de sumarização e se tais trabalhos utilizam técnicas de Detecção de Novidade.

Referência	Monitoramento		Sumarização	Detecção de Novidade
	Perfis	Comportamentos		
(Lauschke e Ntoutsi, 2012)	Não	Sim	Sim	Sim
(Ntoutsi et al., 2011)	Sim	Não	Sim	Sim
(Oliveira e Gama, 2010a)	Sim	Não	Sim	Sim
(Oliveira e Gama, 2010b)	Sim	Não	Sim	Sim
(Oliveira e Gama, 2010c)	Sim	Não	Sim	Sim
(Pereira e Mendes-Moreira, 2016)	Sim	Não	Sim	Sim
(Rizoiu et al., 2015)	Sim	Não	Sim	Sim
(Siddiqui et al., 2015)	Sim	Sim	Sim	Sim
(Spiliopoulou et al., 2006)	Sim	Não	Sim	Sim
(Siddiqui et al., 2012)	Sim	Sim	Sim	Sim

Tabela 4.6: Descrição dos estudos que visam o monitoramento de perfis. Os objetivos do monitoramento, a existência da sumarização dos dados e a utilização de técnicas de Detecção de Novidades.

Áreas de Aplicação

Os trabalhos desta seção são aplicados em diferentes áreas. Tais áreas envolvem eventos temporais (FCD - ver Seção 3) que são de interesse para a aplicação de monitoramento de grupos. Em alguns trabalhos áreas comuns são investigadas, contudo todas as áreas possuem relevâncias para a execução da tarefa de monitoramento. Nesse sentido, a Tabela 4.7 compara como os estudos apresentam suas investigações em cenários de FCD mostrando cada área abordada, a quantidade de Janelas de Eventos analisadas, o tipo de janela utilizada e o período de tempo de tais janelas.

Referências	Area de Aplicação	Janela de Eventos		
		Qtd	Tipo	Tamanho
(Lauschke e Ntoutsi, 2012)	RS	≈ 7	Marcação	Mensal
(Ntoutsi et al., 2011)	RC, T e BDD	[-; -; 7]	Deslizante	[4.000, 400, ≈ 727]
(Oliveira e Gama, 2010a)	EC e ES	[3; 2]	Marcação	Anual
(Oliveira e Gama, 2010b)	EC	3	Marcação	Anual
(Oliveira e Gama, 2010c)	EC e ES	[3; 3]	Marcação	Anual
(Pereira e Mendes-Moreira, 2016)	DM	60	Marcação	6 hours
(Rizoiu et al., 2015)	P e EC	[49; 5]	Marcação	Anual
(Siddiqui et al., 2015)	M	-	-	-
(Spiliopoulou et al., 2006)	BDD	7	Deslizante	≈ 727
(Siddiqui et al., 2012)	EC	5	Marcação	Anual

Tabela 4.7: Comparação entre as áreas abordadas pelos estudos que visam o monitoramento de grupos. A quantidade, tipo e tamanho das Janelas de Eventos utilizadas por cada estudo.

A área de Econômica (EC) é extensivamente investigada por estudos que buscam monitoramento de perfis de uso. Essa área é abordada em cinco estudos, são eles Oliveira e Gama (2010a), Oliveira e Gama (2010b) e Oliveira e Gama (2010c), Rizoiu et al. (2015) e Siddiqui et al. (2012). Os outros 50% dos estudos investigam diferentes áreas, como redes sociais (RS) por Lauschke e Ntoutsi (2012), Redes de computadores (RC) e Transações (T) Ntoutsi et al. (2011), Banco de dados de documentos (BDD) Spiliopoulou et al. (2006) e Ntoutsi et al. (2011), Estatística (ES) por Oliveira e Gama (2010a) e Oliveira e Gama (2010c), Política (P) Rizoiu et al. (2015) e Dispositivos Móveis (DM) por Pereira e Mendes-Moreira (2016). Além disso, Ntoutsi et al. (2011) cobriram muitas áreas de aplicação (RC, T e BDD). No entanto, apesar de Siddiqui et al. (2015) aplicar a sua abordagem para a área de Medicina (M), esse trabalho não descreve adequadamente o uso de Janelas de Eventos. Além disso, o número de Janelas de Eventos, seu tipo e seu tamanho não foram mencionados por Siddiqui et al. (2015).

Estudos que investigam mais de uma área, normalmente utilizam diferentes quantidades de Janelas de Eventos em seus experimentos. Nesse sentido, Ntoutsi et al. (2011), com sua aplicação na área de BDD, empregaram 7 janelas deslizantes com base no estudo anterior de Spiliopoulou et al. (2006). É importante citar que ambos os estudos são os únicos a utilizar janelas deslizantes. No entanto, Ntoutsi et al. (2011) não mencionam a quantidade de Janelas de Eventos para as áreas RC e T. Além disso, as janelas deslizantes empregadas se baseiam um número máximo de eventos a serem processados (4.000, 400, ≈ 727) para determinar o tamanho de cada janela em tais áreas. Por sua vez, Oliveira e Gama (2010a), Oliveira e Gama (2010b) e Oliveira e Gama (2010c) utilizam janelas de marcação com tamanho de um ano e com no máximo 3 janelas analisadas. Tal restrição é devido a natureza dos conjuntos de dados utilizados, os quais são utilizados nos três estudos. Rizoiu et al. (2015) também investigam janelas de marcação com tamanho anual, bem como Siddiqui et al. (2012). De fato Rizoiu et al. (2015) empregam, respectivamente, 49 e 5 Janelas de Eventos para as áreas P e ES, enquanto Siddiqui et al. (2012) abordam 5 janelas em seu estudo.

Entre os demais estudos que usam janelas de marcação, é possível observar diferentes tamanhos de janela, por exemplo, janelas de 6 horas utilizadas por Pereira e Mendes-Moreira (2016) e janelas de um mês utilizadas por Lauschke e Ntoutsi (2012). Pereira e Mendes-Moreira (2016) apresentam um estudo onde mais janelas são investigadas (60), principalmente devido ao tamanho de janela definido (horas). Por outro lado, Lauschke e Ntoutsi (2012) empregaram aproximadamente 7 janelas, uma vez que as frequências de *tweets* de usuários no FCD observado foi variado. Nesse sentido, alguns usuários do Twitter são monitorados por mais tempo do que outros. Em resumo, os estudos desta Seção apresentam abordagens distintas para o uso de Janelas de Eventos (por exemplo, tipo, quantidade e tamanho), o que depende do conjunto de dados e da área de aplicação abordada.

Conjuntos de Dados

Assim como há estudos em diferentes domínios, esses estudos empregam suas propostas em vários conjuntos de dados. Tais conjuntos possuem suas especificações e todos são FCD, mas

em poucos casos tais conjuntos são utilizados por diferentes trabalhos. Neste sentido, a Tabela 4.8 apresenta uma comparação entre os conjuntos de dados utilizados pelos estudos apresentados nesta seção. De fato, a Tabela 4.8 mostra o número de eventos, número de objetos e o período de tempo dos conjuntos de dados investigados por tais estudos.

Referências	Conjunto de Dados		
	Eventos	Objetos	Tempo
(Lauschke e Ntoutsi, 2012)	561	6	≈ 13 meses
(Ntoutsi et al., 2011)	[424.021; 95.412; 4.920]	-	[-; -; 7 anos]
(Oliveira e Gama, 2010a)	-	[439; 30]	[3 anos; 2 anos]
(Oliveira e Gama, 2010b)	-	439	3 anos
(Oliveira e Gama, 2010c)	-	[439; 30]	[3 anos; 2 anos]
(Pereira e Mendes-Moreira, 2016)	-	≈ 900.000	15 dias
(Rizoiu et al., 2015)	-	[23; 836]	[49 anos; 5 anos]
(Siddiqui et al., 2015)	-	29	-
(Spiliopoulou et al., 2006)	4.920	-	7 anos
(Siddiqui et al., 2012)	-	836	5 anos

Tabela 4.8: Comparação entre os conjuntos de dados dos estudo que visam o monitoramento de perfil de uso. Todos estudos são em FCD utilizando conjuntos que possuem diferentes quantidades de eventos, de objetos e de períodos de tempo.

Apenas três estudos descreveram o número de eventos capturados, são eles Lauschke e Ntoutsi (2012), Spiliopoulou et al. (2006) e Ntoutsi et al. (2011). Lauschke e Ntoutsi (2012) observam 6 objetos que geraram 561 eventos sendo o menor número de objetos observados. Além disso, alguns objetos são monitorados por mais meses do que outros, durante um período total de aproximadamente 13 meses. Spiliopoulou et al. (2006) investigam um FCD composto de 4.920 eventos capturados ao longo de 7 anos, bem como Ntoutsi et al. (2011) que utilizam o mesmo conjunto em seus experimentos. No entanto, o número de objetos de tal FCD não foi descrito por ambos os estudos. Por outro lado, Ntoutsi et al. (2011) investigam outros dois FCD com, respectivamente, 424.021 e 95.412 eventos. Contudo, o número de objetos e o tempo cobertos por ambos FCD não foram mencionados no estudo.

Considerando o número de objetos, outros estudos investigam FCDs com diferentes quantidades. Siddiqui et al. (2015) observam 29 pacientes em seus experimentos. Porém, o número de eventos e período de tempo sobre essa observação não foi descrito. Apesar do grande número de objetos presentes no conjunto de dados usado por Pereira e Mendes-Moreira (2016) (≈ 900.000), esses dados são de apenas 15 dias, sendo o menor período de tempo entre os estudos. Ao todo, Oliveira e Gama (2010a), Oliveira e Gama (2010c) e Oliveira e Gama (2010b) investigam dois conjuntos de dados, um contendo 439 e outro que possui 30 objetos. Esses FCDs foram capturados, respectivamente, por três e dois anos. Por sua vez, Rizoiu et al. (2015) também investigam dois conjuntos de dados. O primeiro contendo 23 e o segundo incluindo 836 objetos. Tais conjuntos de dados foram monitorados respectivamente por 49 e 5 anos, sendo 49 o maior período de tempo entre os estudos.

Finalmente, Siddiqui et al. (2012) abordam um FCD monitorado por cinco anos que contêm 836 objetos.

Validação

Apesar da etapa de validação ser um requisito importante para comprovação da utilidade e aplicabilidade da uma solução, ela não foi realizada por todos os estudos. Os trabalhos que apresentam validação realizam esta etapa de diferentes formas. Alguns trabalhos fazem comparações com outras abordagens, enquanto outros comparam seus resultados em diferentes conjuntos de dados. Neste sentido, uma comparação entre o processo de validação dos estudos é apresentada pela Tabela 4.9, a qual apresenta os algoritmos de agrupamento utilizados, o tipo de validação e a comparação aplicada pelos estudos.

Referências	Algoritmos	Validação	Comparação
(Lauschke e Ntoutsi, 2012)	Novo algoritmo	Qualitativo	-
(Ntoutsi et al., 2011)	*K-Means*	Qualitativo	-
(Oliveira e Gama, 2010a)	*MClusT*	Qualitativo	-
(Oliveira e Gama, 2010b)	*Ward* e *K-Means*	Qualitativo	-
(Oliveira e Gama, 2010c)	*Ward* e *K-Means*	Qualitativo	-
(Pereira e Mendes-Moreira, 2016)	*X-Means*	Quantitativo	*MEC*
(Rizoiu et al., 2015)	*ClusPath*	Quantitativo	*TDCK-Means*
(Siddiqui et al., 2015)	*EvoLabelPred*	Quantitativo	Resultados reais
(Spiliopoulou et al., 2006)	*EM*, *Single Linkage*, *CLUTO* e *bisecting K-Means*	Qualitativo	-
(Siddiqui et al., 2012)	*EM*	Quantitativo	*MEC*

Tabela 4.9: Os algoritmos de Agrupamento, os métodos de validação, e comparações aplicadas pelos estudos que visam o monitoramento de perfis de uso.

Diversos algoritmos para a tarefa de Agrupamento foram usados pelos estudos desta seção. Alguns adotaram algoritmos hierárquicos tradicionais enquanto outros algoritmos particionais. Porém, poucos desenvolveram algoritmos próprios. Desta forma, o algoritmo particional *K-means* é o mais utilizado entre os estudos, sendo abordado por Ntoutsi et al. (2011), por Oliveira e Gama (2010b), por Oliveira e Gama (2010c) e com uma variação por Spiliopoulou et al. (2006). Na sequência aparecem os algoritmos hierárquicos *Ward* aplicado por Oliveira e Gama (2010b) e Oliveira e Gama (2010c), *Single Linkage* por Spiliopoulou et al. (2006), e *EM* aplicado por Spiliopoulou et al. (2006) e por Siddiqui et al. (2012). Alguns dos estudos abordaram mais de um algoritmo visando comparar seus resultados. Nesse sentido, Oliveira e Gama (2010b), Oliveira e Gama (2010c) e Spiliopoulou et al. (2006) aplicaram mais de um algoritmo. Em ambos os estudos, Oliveira e Gama (2010b) e Oliveira e Gama (2010c) usaram dois algoritmos, *Ward* e *K-Means*. Por outro lado Spiliopoulou et al. (2006) abordaram quatro algoritmos, são eles *EM*, *CLUTO*, *Single linkage* e *bisecting K-Means*. Outros algoritmos também são aplicados, são eles *MClusT* por Oliveira e Gama (2010a), *X-Means* Pereira e Mendes-Moreira (2016), *ClusPath* por Rizoiu et al. (2015) e

EvoLabelPred por Siddiqui et al. (2015). Além disso, Lauschke e Ntoutsi (2012) apresentam um novo algoritmo que não foi nomeado.

Considerando a validação dos estudos, a maioria deles tentou realizar uma validação do mecanismo de monitoramento, não detalhando a validação dos grupos obtidos ao longo do processo. Nesse sentido, Ntoutsi et al. (2011) pretendeu essencialmente validar o processo de sumarização executado pelo *framework FINGERPRINT*. Para este fim, Ntoutsi et al. (2011) comparou os resultados da utilização de diferentes parâmetros da aplicação do algoritmo de agrupamento juntamente com o *MONIC*, o qual foi desenvolvido por Spiliopoulou et al. (2006). No entanto, Ntoutsi et al. (2011) não apresentou comparação com outros estudos. Por outro lado, o estudo de Spiliopoulou et al. (2006) foi um dos primeiros na área de monitoramento de grupos. Dentre os estudos, o *MEC* foi o método de monitoramento mais comparada. Tal comparação foi realizada pelos estudo de Siddiqui et al. (2012) e Pereira e Mendes-Moreira (2016).

Outros estudos também correlacionaram seus resultados com abordagens distintas. Rizoiu et al. (2015) realizam uma comparação entre a execução de sua abordagem e outro método denominado *TDCK-Means*. Por sua vez, Siddiqui et al. (2015) validam seu mecanismo comparando os resultados obtidos com os casos reais ocorridos com seus objetos. Essa validação foi possível, uma vez que o conjunto de dados utilizado em seus experimentos contêm os registros reais dos pacientes, bem como os tratamentos e exames aplicados a eles. No entanto, outros estudos não fizeram qualquer comparação das suas abordagens, são eles Lauschke e Ntoutsi (2012), Ntoutsi et al. (2011), Oliveira e Gama (2010a), Oliveira e Gama (2010b) e Oliveira e Gama (2010c).

4.4 Avaliação Geral

Vinte estudos (listados na Tabela 4.1, em ordem alfabética por título) foram selecionados por esta revisão sistemática até primeiro de maio de 2018. Esses estudos apresentaram modelos, métodos, metodologias, ferramentas, *frameworks*, técnicas e abordagens para o problema da identificação e monitoramento de grupos em contexto de utilização. Conforme apresentado nas seções anteriores, esses estudos foram distribuídos em duas categorias distintas, são elas: identificação e monitoramento de grupos. Nesse sentido, categorizamos como identificação 50% dos estudos e como monitoramento outros 50% dos estudos. Esta categorização é necessária, uma vez que alguns estudos exploram a tarefa de Agrupamento sem considerar o monitoramento dos grupos obtidos ao longo do tempo. A partir da pesquisa realizada (Etapa I) e da revisão sistemática subsequente (Etapa II), conforme especificado em seu protocolo, é plausível responder às questões de pesquisa propostas:

- **Quais métodos são utilizados no processo de identificação perfis de uso (grupos)?** Nesta revisão sistemática, foram econtrados estudos aplicados em diferentes cenários. Em cenários de FCD ou *Big Data*, os dados são esparsos, exigindo o emprego de técnicas de mineração de dados destinadas a melhorar a representatividade dos dados capturados. Por

outro lado, os cenários em *batch* também precisam de técnicas de mineração de dados, mas na maioria dos casos, em pequena proporção em relação ao cenários de *Big Data* ou FCD. Nesse sentido, alguns estudos propõem a identificação de grupos em cenários *batch* (Chu et al., 2007; Hamka et al., 2014; Li e Deng, 2012; Sohn e Kim, 2008; Shabana et al., 2016; Zhu et al., 2011), enquanto outros abordam tal tarefa em um cenário de DFC (Ballea et al., 2013; Rehman e Raza Ali, 2015; Zhang, 2007; Lauschke e Ntoutsi, 2012; Ntoutsi et al., 2011; Oliveira e Gama, 2010b,a,c; Pereira e Mendes-Moreira, 2016; Rizoiu et al., 2015; Siddiqui et al., 2012, 2015; Spiliopoulou et al., 2006).

Foram encontrados vários algoritmos para a tarefa de Agrupamento, são eles, *GHSOM*, *EM*, *K-Means*, *Ward*, *Average*, *Two-step*, *Centroid*, *Midrange*, *MClusT*, *X-Means*, *Single Linkage*, *Bisecting K-means* e *CLUTO*. Tais algoritmos são amplamente utilizados e discutidos na maioria dos estudos apresentados nas duas categorias propostas por esta revisão. Além disso, alguns estudos propuseram novos algoritmos, como *DP-Means* (Shabana et al., 2016), *ClusPath* (Rizoiu et al., 2015) e *EvoLabelPred* (Siddiqui et al., 2015), além de outros dois sem nome (Hsu et al., 2012; Lauschke e Ntoutsi, 2012). Neste contexto, alguns estudos empregaram um único algoritmo enquanto outros aplicaram dois ou mais algoritmos. Em geral, a definição do melhor número de grupos é uma tarefa complexa. Quando grandes conjuntos de dados são investigados, essa tarefa se torna ainda mais complicada, o que é comum na maioria dos problemas do mundo real. É possível observar que estudos de ambas as categorias sugeridas investigam a aplicação de várias medidas de avaliação visando descrever o melhor número de grupos a serem formados. De fato, encontramos várias medidas, como *SWC*, *BIC*, *Average index*, *C*, *Accuracy* e a *distribuição de dados*. Tais técnicas também foram aplicadas individualmente ou em combinação.

- **Quais técnicas são aplicadas no processo de monitoramento destes perfis?**

  Em relação aos estudos que também buscam o monitoramento de grupos, é possível reconhecer novas propostas que estão sendo exploradas principalmente nos últimos anos. Contudo, alguns estudos selecionados por esta revisão não apresentaram novas soluções. De fato, alguns manuscritos mostram progressões da mesma proposta publicada por uma equipe de autores. Em geral, todos os estudos que visam o monitoramento de grupos abordaram o cenários de FCD. No entanto, tais estudos utilizaram diferentes tipos de Janelas de Eventos, *Marcação* (Lauschke e Ntoutsi, 2012; Oliveira e Gama, 2010b,a,c; Pereira e Mendes-Moreira, 2016; Siddiqui et al., 2012) e *Deslizante* (Rehman e Raza Ali, 2015; Zhang, 2007; Ntoutsi et al., 2011; Spiliopoulou et al., 2006), bem como vários tamanhos para essas janelas. Nesse caso, foram usadas janelas baseadas em *horas*, *meses* e *anos* para janelas de *Marcação* e *um número de objetos* como tamanho das janelas *Deslizantes*.

  Por fim, todos os estudos que buscam o monitoramento de grupos realizam a sumarização de dados para lidar com restrições de memória e processamento de dados. Em particular, o estudo de Ntoutsi et al. (2012), que propõem um novo método de sumarização de FCD chamado *FINGERPRINT*. Além disso, dois procedimentos principais foram propostos para o monitora-

mento de perfis, que são de *Enumeração* (Spiliopoulou et al., 2006) e *Compreensão* (Oliveira e Gama, 2010c).

- **Qual a área de aplicação?**

  A maioria dos estudos classificados como de identificação de grupos apresenta propostas em áreas que envolvem o uso de dispositivos móveis, principalmente na área de telecom, exceto um estudo dessa categoria que busca identificar grupos na área de bibliotecas. Por outro lado, os estudos listados como monitoramento de grupos são aplicados em diversas áreas, como *bancos de dados de documentos*, *Redes de computadores*, *Transações*, *Economia*, *Estatística*, *Dispositivos Móveis*, *Rede social*, *Política* e também *Medicina*. Em geral, os estudos rotulados como identificação de grupos foram substancialmente aplicados em áreas que envolvem o uso de dispositivos móveis. Em contraste, o emprego de tarefas de monitoramento neste tipo de área ainda é pouco investigado. Pereira e Mendes-Moreira (2016) desenvolveram o único trabalho encontrado por esta revisão aplicando a tarefa de monitoramento em uma área de dispositivo móvel. Este trabalho investigou uma variação do estudo de Oliveira e Gama (2010c) mas sem o uso de dados de uso de aplicativos, usando somente CDR.

- **Quais tipos de conjuntos de dados são abordados?**

  A maioria dos estudos visando identificação de grupos utiliza conjuntos de dados de dispositivos móveis. Nesse caso, a maioria dos conjuntos de dados é formada por dados de CDR, pessoais e de faturamento. Nesse contexto, alguns estudos também aumentam seus conjuntos de dados com dados de hardware de dispositivos móveis, uso de aplicativos e tráfego da internet. Por outro lado, os estudos de monitoramento abordaram vários tipos de conjuntos de dados. Em alguns casos, os mesmos conjuntos de dados são usados por diferentes estudos, por exemplo, Spiliopoulou et al. (2006) e Ntoutsi et al. (2011). No entanto, o uso do mesmo conjunto de dados foi realizado por estudos do mesmo grupo de autores (Oliveira e Gama, 2010b,a,c). É importante notar que os estudos analisados por esta revisão utilizaram conjuntos de dados reais e artificiais, em sua maioria privados e não disponíveis. Também é importante observar que, mesmo com o alto uso de conjuntos de dados de dispositivos móveis, há uma falta de pesquisa sobre o uso de aplicativos em tais dispositivos. Esse tipo de dado é gerado em grandes proporções, mas ainda pouco explorado pelos dois cenários, identificação e monitoramento de grupos.

- **Como avalia o desempenho da solução?**

  Os estudos analisados por esta revisão realizaram validações de suas propostas de diferentes maneiras. Nesse sentido, alguns estudos apenas discutem seus resultados experimentais com métricas de desempenho em um conjunto de dados. Outros estudos realizam experimentos com mais de um conjunto de dados mostrando o desempenho de sua solução em diferentes situações. Por outro lado, propostas semelhantes forneceram comparações entre si. Essa comparação ocorre em poucos estudos, especialmente naqueles que buscam o monitoramento

de grupos, por exemplo, Pereira e Mendes-Moreira (2016) e Siddiqui et al. (2012) onde ambos comparam seus resultados com Oliveira e Gama (2010c).

4.5 Considerações Finais do Capítulo

Este capítulo apresentou os trabalhos relacionados a esta pesquisa que foram selecionados por meio de uma revisão da literatura, os quais foram discutidos e comparados. Tais trabalhos visam a identificação e o monitoramento de perfis de uso, principalmente em cenários de FCD. Contudo, não foram encontrados dados suficientes para recomendar o uso rotineiro de algum algoritmo de Agrupamento ou algum método de monitoramento de grupos. As implicações desta revisão sistemática podem ser devido a viés de publicações e limitações de estudo individuais. Além disso, é impossível recomendar com segurança quaisquer soluções em diferentes domínios de aplicação, uma vez que há falta de validação, bem como restrições impostas por conjuntos de dados privados. Com base nos resultados obtidos por esta revisão, é possível identificar lacunas nas soluções existentes, tais como:

1. Explorar de forma efetiva grandes FCDs, em termos de eventos de uso de aplicativos, aplicativos utilizados e usuários, sem incorrer em uma explosão computacional e sem necessariamente armazenar o FCD ao longo de diferentes Janelas de Eventos;
2. Buscar definir um número limitado de perfis de uso de aplicativos por meio de tarefas de Aprendizado de Máquina e avaliação do número de grupos a serem formados;
3. Monitorar mudanças de comportamentos de usuários dado a evolução de FCD de uso de dispositivos móveis bem como de aplicativos.
4. Validar amplamente a proposta, levando em consideração os grandes FCD e a falta de classe nos conjuntos de dados.

As contribuições da revisão sistemática realizada durante esta pesquisa, visam a busca do aprimoramento sobre o conhecimento e o incremento das discussões sobre as tarefas de identificação e monitoramento de grupos, principalmente no que diz respeito à criação e evolução de soluções para suporte a tais tarefas. Os detalhes mais importantes que encontramos são a evidência e o aprimoramento das técnicas de Agrupamento e também os métodos de monitoramento de grupos, bem como a evolução dos domínios em que as soluções existentes estão sendo abordadas. Neste caso, notamos o baixo número de validações em estudos que buscam a identificação de grupos e a baixa quantidade de estudos que exploram o monitoramento de grupos. Nesse sentido, como a validação é um passo importante e vários estudos de monitoramento são evolutivos uns dos outros, a possibilidade de novos pesquisadores focados nessas tarefas é viável. Outros resultados significativos são relacionados a identificação das validações de soluções que buscam o monitoramento de grupos, bem como a identificação de conjuntos de dados centrais utilizados para tais estudos, contribuindo para aumentar a qualidade de soluções futuras de tal tarefa.

5. FRAMEWORK f-DOPE

Neste Capítulo é apresentado o *Framework for iDentification and mOnitoring of Profiles and bEhaviors - f-DOPE*, proposto para mineração de FCD que visa a identificação e o monitoramento de perfis e comportamentos baseado na utilização de aplicativos em dispositivos móveis. Como diferencial, o *f-DOPE* possui diferentes fases responsáveis pelo pré-processamento dos dados, detecção de padrões de uso de aplicativos, identificação e monitoramento de perfis e comportamentos, possibilitando a segmentação de tais comportamentos reagindo de acordo com a distribuição do FCD. Para o desenvolvimento deste *framework* foram utilizadas as linguagens de programação *Python* e *R* aplicando algoritmos de Mineração de Dados, Aprendizado de Máquina e de Detecção de Novidade, qualificadas para a proposta desejada. Os algoritmos que foram desenvolvidos durante esta pesquisa estão em um repositório no *Github*[1]. Nas próximas seções são apresentados maiores detalhes sobre o *f-DOPE*. A Seção 5.1 apresenta a visão esquemática do *f-DOPE*. Na Seção 5.2 é demonstrada a primeira etapa do *f-DOPE*. Por sua vez, a Seção 5.3 expõe a segunda etapa do *f-DOPE*. Por fim, na Seção 5.4 são apresentadas as considerações finais deste Capítulo.

5.1 Visão Esquemática do *Framework* Proposto

Como verificado pela revisão da literatura no Capítulo 4, eventos de uso de aplicativos não são comumente utilizados para identificar perfis de uso. No entanto, o uso de aplicativos está em constante crescimento. Além disso, a tarefa de monitoramento de perfis e comportamentos ainda é pouco explorada em áreas relacionadas a dispositivos móveis, as quais ainda usam eventos de CDR. Nesse sentido, este estudo propõe um *framework* que possa extrair diferentes padrões de uso de aplicativos, fornecer um número limitado de perfis de uso e permitir o monitoramento de perfis e comportamentos ao longo do tempo com base em um cenário real.

A exploração de uso de aplicativos em dispositivos móveis pode permitir a identificação de perfis e estabelecer diferentes comportamentos. Neste contexto, tarefas de aprendizado de máquina como regras de associação e agrupamento podem ajudar na investigação destas situações, permitindo, por exemplo, que empresas fabricantes de dispositivos móveis possam buscar a fidelização de seus consumidores oferecendo novos serviços e/ou produtos. O diagrama de atividades da Figura 5.1. apresenta o funcionamento do *framework*, chamado *f-DOPE*.

O *f-DOPE* é resumido em duas etapas principais, chamadas de *Mineração do FCD* e *Acompanhamento do FCD*. A primeira etapa possui três fases, de *Absorção*, *Associação* e *Caracterização*, visando a identificação de perfis de uso. Ao final da primeira etapa, para cada janela, são formados grupos representando os perfis de uso identificados. Esses perfis são utilizados como entrada na segunda etapa que possui duas fases, de *Monitoramento* e *Segmentação*, buscando analisar os perfis

[1] https://github.com/nielsenrechia/Framework

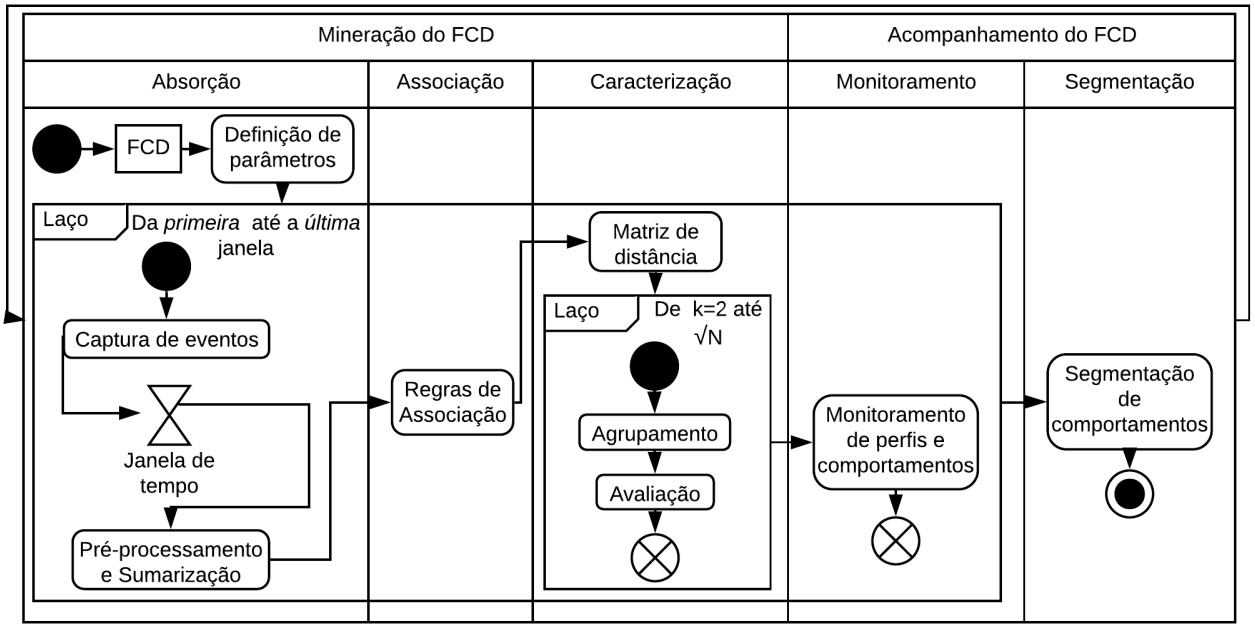


Figura 5.1: Diagrama de atividades do *framework f-DOPE*.

e comportamentos ao longo das janelas e permitindo a segmentação de comportamentos desejados por partes interessadas.

5.2 Etapa 1: Mineração do FCD

A exploração de uso de aplicativos em dispositivos móveis pode permitir a identificação de perfis e estabelecer diferentes comportamentos de consumidores. Neste contexto, algoritmos de Mineração de Dados e Aprendizado de Máquina não supervisionado podem ajudar na investigação destas situações. Desta forma, cada fase da primeira etapa do *f-DOPE* é proposta com a aplicação de algoritmos de Mineração de Dados e de Aprendizado não supervisionado em busca da identificação de perfis de uso de aplicativos em dispositivos móveis. Por meio de estudos de caso, foram realizados testes que demonstraram as necessidades e as melhores abordagens para cada uma das fases desta etapa. Assim, os Apêndices B, C e D apresentam alguns estudos de caso que apontam para a indispensabilidade de tais fases.

Nesta pesquisa, diferentemente dos FCDs tradicionais que consideram os eventos de um FCD como uma representação completa de objetos independentes, o FCD utilizado possui eventos que são registros (atividades), realizados por um único objeto (dispositivo). Assim um único dispositivo pode gerar muitos eventos em uma única janela. A evolução dos eventos ao longo do FCD e as tarefas de aprendizado de máquina que são utilizadas, motivam a necessidade de que os eventos sejam pré-processados e sumarizados. É importante notar que é necessário definir uma janela de eventos (Babcock et al., 2002b) (Ver Seção 3.2.2) a qual pode ou não ser fixa e uma janela de eventos inicial, que aponte para o início do monitoramento de eventos de uso de aplicativos. Também é

necessário a definição do número de consumidores n a serem monitorados a partir da janela inicial. Essas definições devem ser determinadas de acordo com os recursos computacionais disponíveis.

5.2.1 Fase de Absorção

Nesta fase, todos os eventos do FCD são continuamente pré-processados e sumarizados para lidar com as diversas atividades executadas em um único dispositivo e também pelas restrições de espaço e memória física. Em um cenário real é necessário um pré-processamento dos dados capturados, pois são produzidos centenas de milhares de eventos em um curto período de tempo. Esse grande volume de registros torna a sumarização de tais dados importante, pois visa a preservação do significado de todos estes eventos sem de fato armazená-los. A execução desta fase ocorre de acordo com o Algoritmo 5.1.

Input

A função *ABSORPTION* recebe como entrada um conjunto D que é composto por eventos ε, onde ε é uma tupla $(\varepsilon_i, \varepsilon_p, \varepsilon_d, \varepsilon_{et})$, conforme descrito no Capítulo 1, Seção 1.3. Além disso, D é um conjunto de eventos capturado em uma janela de eventos $\mathbf{w}$. Neste caso, sendo $D_{\mathbf{w}}$ a janela inicial, então, $D_{\mathbf{w}} \cap D_{\mathbf{w+1}} = \varnothing$. Caso seja necessário, é possível atribuir um peso aos eventos pertences a D e para cada janela $\mathbf{w}$, principalmente se forem escolhidas janelas deslizantes. Além disso, são necessários valores para os limiares, τ_{most}, τ_{rem}, τ_{pop} e $minTime$, que são aplicados na execução do pré-processamento dos dados. A definição de tais limiares é apresentada no decorrer desta seção.

Output

Como saída da função *ABSORPTION* é gerada uma matriz de dados ω, a qual é o resultado do pré-processamento e da sumarização de D.

Pre-processamento

Inicialmente (linha 2), são instanciadas a variável **D**, a qual é um subconjunto de D que armazenará eventos dos aplicativos *mais utilizados*, e a variável *mostUsed*, que receberá os aplicativos *mais utilizados*. Logo após (linhas 3-9), busca-se descobrir os aplicativos *mais utilizados*. Para cada aplicativo único p existente em D (linha 3), são selecionados e armazenados em S_p os eventos ε onde $\varepsilon_p = p$ (linha 4). Para ser considerado um aplicativo *mais utilizado* p tem que ser utilizado por uma porcentagem mínima aceitável de dispositivos. Nesse sentido, adota-se o limiar τ_{most} e verifica-se a porcentagem de dispositivos que utilizam p que deve ser $\geq \tau_{most}$ (linha 5), sendo $|S[i]|$ a quantidade de dispositivos únicos em S_p e $|D[i]|$ a quantidade de dispositivos únicos em D. Quando um aplicativo p for considerado *mais utilizado* este é armazenado na variável *mostUsed*

Algoritmo 5.1: Fase de Absorção do *framework f-DOPE*.

1: **function** ABSORPTION(D, τ_{most}, τ_{rem}, τ_{pop}, $minTime$) ▷ D é um conjunto de eventos
2: **D** = $mostUsed = \varnothing$ ▷ Variáveis auxiliares
3: **for** $p \in [p_1, p_m]$ **do** ▷ Aplicativos mais usados - Def. 1
4: $S_p = \forall\ \varepsilon \in D \mid \varepsilon_p = p$ ▷ Eventos do aplicativo p
5: **if** $|S_p[i]| \ / \ |D[i]| \geq \tau_{most}$ **then** ▷ Verificar se p é um aplicativo *mais utilizado*
6: $mostUsed = mostUsed \cup p$ ▷ p é considerado um aplicativo *mais utilizado*
7: **end if**
8: **end for**
9: $D = \forall\ \varepsilon \in D \mid \varepsilon_p \in mostUsed$ ▷ Manter eventos dos *mais utilizados*
10: **for** $i \in [i_1, i_n]$ **do** ▷ Aplicativos remanescentes - Def. 2
11: $U_i = \forall\ \varepsilon \in D \mid \varepsilon_i = i$ ▷ Conjunto de eventos de um dispositivo i
12: $T_{min} = UTUT(i) \times \tau_{rem}$ ▷ Definir T_{min} com base no $UTUT(i)$ (Equação 5.2) e τ_{rem}
13: **for** $p \in [p_1, p_m]$ **do**
14: $SU_{i,p} = \forall\ \varepsilon \in D \mid \varepsilon_i = i \wedge \varepsilon_p = p$ ▷ Eventos de um aplicativo p em um dispositivo i
15: **D** = **D** $\cup\ \ \forall \varepsilon \in SU_{i,p} \mid ATUT(i,p) \geqslant T_{min} \wedge ATUT(i,p) \geqslant minTime$
▷ Manter eventos dos *remanescentes*
16: **end for**
17: **end for**
18: $mostUsed = \ popular = \varnothing$ ▷ Variáveis auxiliares
19: **for** $p \in [p_1, p_m]$ **do** ▷ Aplicativos populares - Def. 3
20: $S_p = \forall \varepsilon \in$ **D** $\mid \varepsilon_p = p$ ▷ Conjunto de eventos do aplicativo p
21: **if** $|S_p[i]| \ / \ |$**D**$[i]| \geq \tau_{pop}$ **then** ▷ Verificar se p é um aplicativo *popular*
22: $mostUsed = mostUsed \cup p$ ▷ p é considerado um aplicativo *mais utilizado*
23: $popular = popular \cup p$ ▷ p é considerado um aplicativo *popular*
24: **else if** $|S_p[i]| \ / \ |$**D**$[i]| \geq \tau_{most}$ **then** ▷ Verificar se p é um aplicativo *mais utilizado*
25: $mostUsed = mostUsed \cup p$ ▷ p é considerado um aplicativo *mais utilizado*
26: **end if**
27: **end for**
28: **D** $= \forall \varepsilon \in$ **D** $\mid \varepsilon_p \in mostUsed$ ▷ Manter eventos dos aplicativos mais utilizados em **D**
29: $\omega = matrix[|$**D**$[i]|][|$**D**$[p]|]$ ▷ Matriz de sumarização dos eventos
30: **for** $i \in [i_1, i_n]$ **do** ▷ Discretização/sumarização
31: **for** $p \in [p_1, p_m]$ **do**
32: $SU_{i,p} = \forall\ \varepsilon \in$ **D** $\mid \varepsilon_i = i \wedge \varepsilon_p = p$ ▷ Eventos de um aplicativo p em um dispositivo i
33: $\omega[i][p] = ATUT(i,p)$ ▷ Calcular $ATUT(i,p)$ para posição i e p
34: **end for**
35: **end for**
36: **for** $p \in mostUsed$ **do**
37: **if** $p \in popular$ **then**
38: $IP(\omega[p])$ ▷ Discretização dos aplicativos *populares*
39: **else**
40: *Transformar p em um atributo categórico*
41: **end if**
42: **end for**
43: **return** ω
44: **end function**

(linha 6). Desta forma, todos os eventos ε onde ε_p é um aplicativo *mais utilizado* são mantidos em D enquanto que os demais são desconsiderados (linha 9).

Definição 1 *(*Aplicativos mais utilizados*) Assuma $p_1, p_2, ..., p_m$ sendo os aplicativos únicos utilizados em D e S_p como um conjunto de eventos de um determinado aplicativo p ($S_p \subseteq D$), tal que $p_j \neq p_q : S_{p_j} \cap S_{p_q} = \varnothing$, $\bigcup_{p=1}^{m} S_p = D$. Ademais, assuma $i_1, i_2, ..., i_n$ sendo os dispositivos monitorados e τ_{most} sendo um valor percentual mínimo aceitável. Um aplicativo p é um aplicativo mais utilizado se $|S_p[i]| \,/\, |D[i]| \geq \tau_{most}$ (linhas 5-7) e $|S_p[i]| \,/\, |\mathbf{D}[i]| \geq \tau_{most}$ (linhas 24-26)*

O limiar τ_{most} é um valor que pode ser ajustado, por exemplo, pode ser escolhido um valor dentro do intervalo $[0,01;\ 0,1]$. Em resumo, a seleção dos aplicativos *mais utilizados* visa a redução da complexidade e tempo de processamento mantendo os aplicativos mais frequentes.

A definição de aplicativos *mais utilizados* foi elaborada após a percepção da existência de um número substancial de aplicativos usados por apenas um único dispositivo, enquanto outros aplicativos são utilizados em poucos dispositivos (por exemplo, menos de 1%) (ver Apêndice B). Assim, na etapa de pré-processamento é conveniente selecionar apenas aplicativos usados por um número significativo de dispositivos, que possuem atividades mínimas necessárias permitindo realizar uma análise mais realista do uso de tais aplicativos. Tal definição reduz o escopo dos aplicativos presentes, mantendo os aplicativos relevantes e removendo os aplicativos que são utilizados por poucos dispositivos.

Em continuidade (linhas 10-17), busca-se descobrir os aplicativos *remanescentes*. Para cada dispositivo monitorado i (linha 10) são selecionados e armazenados em U_i os eventos ε onde $\varepsilon_i = i$ (linha 11). Para ser considerado um aplicativo *remanescente* p tem que ser utilizado por uma quantidade mínima aceitável de tempo. Assim, são adotados os limiares τ_{rem} e $minTime$. Então (linha 12), calcula-se o $UTUT(i)$ (Equação 5.2), que tem seu resultado multiplicado por τ_{rem} indicando o tempo de uso mínimo para o dispositivo i (T_{min}). Depois (linha 13), para cada p utilizado pelo dispositivo i são selecionados e armazenados em $SU_{i,p}$ os eventos ε onde $\varepsilon_i = i \wedge \varepsilon_p = p$ (linha 14). Após (linha 15), verifica-se o tempo total de uso $ATUT(i,p)$, Equação 5.1 de cada aplicativo p utilizado pelo dispositivo i, o qual deve ser $\geqslant\ T_{min} \wedge \geqslant minTime$, onde $minTime$ é um tempo mínimo aceitável independente do tempo total de uso do dispositivo.

Definição 2 *(*Aplicativos remanescentes*) Assuma $p_1, p_2, ..., p_m$ sendo os aplicativos mais utilizados de acordo com a Def. 1 e τ_{rem} com um valor percentual mínimo aceitável. Além disso, assuma U_i sendo um conjunto de eventos gerados em um dispositivo i, ($U_i \subseteq D$), tal que $i_j \neq i_q : U_{i_j} \cap U_{i_q} = \varnothing$, $\bigcup_{i=1}^{n} U_i = D$, e $SU_{i,j}$ como um conjunto de eventos gerados por um dispositivo i por meio de um único aplicativo p, ($SU_{i,j} \subseteq D$). $ATUT(i,p)$ é a soma do tempo de uso de um aplicativo p gerado no dispositivo i:*

$$ATUT(i,p) = \sum_{\varepsilon \in SU_{i,p}} \varepsilon_d \tag{5.1}$$

O $UTUT(i)$ é a soma do tempo de uso de todos os eventos gerados por um dispositivo i:

$$UTUT(i) = \sum_{\varepsilon \in U_i} \varepsilon_d \tag{5.2}$$

Portanto, um aplicativo p é considerado um aplicativo remanescente em um dispositivo i se $ATUT(i,p) \geqslant T_{min} \wedge ATUT(i,p) \geqslant minTime$.

Os limiares τ_{rem} e *minTime*, assim como τ_{most}, podem ser ajustados. Contudo, enquanto τ_{rem} pode ser, por exemplo, um valor do itervalo $[0,01; 0,1]$, *minTime* deve ser uma porção de tempo. Além disso, pode-se considerar o $UTUT(i)$ como o somatório dos $ATUT(i,p)$ para um mesmo dispositivo i.

Ferreira et al. (2014) afirmam que o tempo de uso mínimo aceitável para um aplicativo é de cerca de 15 segundos. Ferreira et al. (2014) chamam esse tipo de utilização de *micro-uso*. Assim, os aplicativos utilizados por um tempo menor que um limite aceitável podem ser desconsiderados dado que representam um uso irrelevante em tal dispositivo. Além disso, em dispositivos móveis são utilizados, em média, 10 aplicativos por dia. Em um mês, esse número fica próximo de 30 aplicativos utilizados (Annie, 2017). O uso de um aplicativo por menos de um limite aceitável pode simplesmente representar um erro cometido ao tentar abrir outro aplicativo. Além disso, alguns aplicativos são usados diariamente, enquanto outros são esporádicos. Nesse sentido, buscou-se determinar os aplicativos que realmente são utilizados em cada dispositivos.

Na sequência (linha 18), são instanciadas a variável *mostUsed*, que receberá os aplicativos que ainda são considerados *mais utilizados*, e a variável *popular*, a qual recebe os aplicativos considerados *populares*. Assim (linhas 19-28), busca-se descobrir os aplicativos *populares* e os aplicativos que ainda são *mais utilizados*, pois alguns aplicativos considerados *mais utilizados* anteriormente podem não ser mais usados pelo mesmo percentual de dispositivos dado a definição dos aplicativos *remanescentes*. Desta forma, para cada aplicativo único p de $\mathbf{D}$ (linha 19) são selecionados e armazenados em S_p os eventos ε onde $\varepsilon_p = p$ (linha 20). Similar a definição de aplicativos *mais utilizados*, para ser considerado um aplicativo *popular* p tem que ser utilizado por uma porcentagem mínima aceitável de dispositivos. Assim, adota-se o limiar τ_{pop} e verifica-se a porcentagem de dispositivos que utilizam p que deve ser $> \tau_{pop}$ (linha 21), sendo $|\mathbf{D}[i]|$ a quantidade de dispositivos únicos em $\mathbf{D}$. Quando um aplicativo p for considerado *popular*, este é armazenado na variável *mostUsed* (linha 22) e também na variável *popular* (linha 23). Caso o aplicativo p não seja considerado *popular* ele ainda pode ser considerado *mais utilizado* (linha 24), sendo armazenado na variável *mostUsed* (linha 25). Desta forma, todos os eventos ε onde ε_p é um aplicativo *mais utilizado* são mantidos em $\mathbf{D}$ enquanto que os demais são descartados (linha 28).

Definição 3 *(*Aplicativos Populares*) Assuma* S_p *sendo um conjunto de eventos de um aplicativo p, o qual é um aplicativo mais utilizado e remanescente de acordo com Def. 1 e Def. 2. Além disso, assuma* τ_{pop} *como um valor percentual mínimo aceitável. Um aplicativo p é um aplicativo popular se* $|S_p[i]| \,/\, |\mathbf{D}[i]| \geq \tau_{pop}$.

Assim como os limiares τ_{rem} e τ_{most}, τ_{pop} pode ser ajustado, por exemplo, pode ser escolhido um valor entre o intervalo $[0,5; 1,0]$. Além disso, um aplicativo considerado *popular* também deve ser considerado *mais utilizado* pois o limiar τ_{pop} deve ser ao menos igual ao limiar τ_{most}.

Sumarização

Dentre os aplicativos *mais utilizados* que permanecem de acordo com a Def. 2, alguns são utilizados em menos dispositivos, enquanto outros são utilizados em muitos dispositivos. Por exemplo, alguns aplicativos são utilizados por apenas 30% dos dispositivos (ex: Google Docs) e outros chegam a ser utilizados por até 98% dos dispositivos (ex: Whatsapp) em uma mesmo período de tempo. Nesse sentido, alguns aplicativos são considerados *populares*. Constatou-se que tais aplicativos são utilizados em muitos dispositivos e precisam ser tratados de maneira especial dado que são muito frequentes e têm muitas formas de utilização. Por exemplo, um aplicativo *popular* pode ser utilizado por alguns minutos em um dispositivo, enquanto este mesmo aplicativo pode ser utilizado por horas em outro dispositivo em uma única janela de eventos.

Em prosseguimento (linha 29), é instanciada a matriz ω que resumirá os eventos de $\mathbf{D}$, onde $|\mathbf{D}[p]|$ é a quantidade de aplicativos únicos em $\mathbf{D}$. Portanto, tal matriz possui como objetos os dispositivos monitorados, os quais são representadas pelo IMEI anônimo do dispositivo (ε_i). Por sua vez, os aplicativos são considerados os atributos, os quais são simbolizados pelo nome do pacote do aplicativo (ε_p). Nesse sentido, os eventos gerados em uma determinada janela são sumarizados em ω com o cálculo do $ATUT(i,p)$ para todos todos dispositivos i e todos aplicativos p, onde $\omega[i_n][p_m]$ (linhas 30-35). É importante notar, que em alguns casos, dispositivos i não apresentam uso de aplicativos p ocasionando $ATUT(i,p) = 0$. Além disso, ω pode ser modificada de acordo com a aplicação da discretização dos aplicativos *populares* (linhas 36-42). Desta forma, os valores de $ATUT(i,p)$ calculados para os aplicativos *populares* são transformados em valores categóricos que representam os intervalos da discretização (linhas 37-38), enquanto que valores de $ATUT(i,p)$ dos aplicativos não *populares* são modificados em atributos categóricos (linhas 39-40) que caracterizam um único intervalo de tempo de uso. Assim, os eventos são resumidos e armazenados permitindo o descarte dos eventos reais. A matriz ω é um dataset que resume os eventos capturados do FCD refletindo as características de tal FCD em um formato diferente, o que ajuda nos resultados dos algoritmos de aprendizado de máquina.

Dado os diferentes tempo de uso dos aplicativos *populares*, decidiu-se explorar a hipótese de que tais variações poderiam ser separadas em diferentes intervalos de tempo. Assim, dispositivos com tempo de uso similar estariam juntos em uma mesmo intervalo de tempo, o qual seria diferente de outros intervalos que abrangem dispositivos com tempo de uso desigual. Nesse caso, optou-se pela aplicação de um processo de discretização não supervisionada (Tan et al., 2006) (ver Capítulo 2, Seção 2.1.2) em tais aplicativos. Alguns algoritmos, especialmente aqueles para a tarefa de mineração de regras de associação, exigem dados categóricos. Além disso, técnicas de discretização podem ser realizadas para a redução e transformação de atributos contínuos. Três técnicas de discretização foram abordadas: por *frequência igual*, por *Agrupamento* e por *IP* (Han et al., 2011) (Ver

Apêndice B). Os aplicativos *populares* apresentaram uma melhor discretização pelo uso da técnica *IP*. Além disso, a técnica de discretização *IP* satisfaz dois requisitos para a tarefa de discretização não supervisionada: i) agrupa os valores em grupos do mesmo tamanho e ii) agrupa os valores em um número limitado de intervalos de uma maneira mais natural. Esse tipo de discretização é uma maneira eficaz de dividir os atributos contínuos em um pequeno ou reduzido número de intervalos (de 3 a 6 intervalos). Além disso, ao manter a divisão em intervalos naturais, a discretização por *IP* torna-se mais intuitiva.

Em resumo, ao final do pré-processamento, aplicativos menos utilizados, tanto por quantidade de dispositivos quanto por tempo de uso, são desconsiderados. Tais aplicativos podem ser removidos por serem: i) utilizados por um ou por poucos dispositivos monitorados (Def. 1) e ii) utilizados por uma quantidade de tempo desinteressante (Def. 2). Por outro lado, existem aplicativos que são utilizados por uma grande quantidade de dispositivos mas em diferentes frequências, os quais são divididos em diferentes intervalos de tempo de uso (Def. 3).

Por fim, as definições e abordagens escolhidas para esta fase foram definidas após um conjunto de estudos de caso que são apresentados no Apêndice B. Além disso, resultados que demonstram a importância desta fase podem ser encontrados no Capítulo 6, Seção 6.3.1.

A complexidade computacional do Algoritmo 5.1 é extremamente relacionada ao número de aplicativos e ao número de dispositivos monitorados. Desta forma, para selecionar os aplicativos *mais utilizados* a complexidade é $O(p)$, onde p é o número total de aplicativos encontrados em $D_{\mathbf{w}}$. A identificação dos aplicativos *remanescentes* possui complexidade $O(i \times p)$, onde i é o número de dispositivos monitorados e p é a quantidade de aplicativos *mais utilizados*. Para selecionar os aplicativos *populares* a complexidade é $O(p)$, onde p é o número de aplicativos *mais utilizados* e *remanescentes*. A sumarização dos eventos em ω possui complexidade $O(i \times p)$. Por fim a discretização possui complexidade $O(p)$.

5.2.2 Fase de Associação

Com base no estudo de caso B, não é possível identificar padrões de uso de aplicativos utilizando o tempo total de uso ou a categoria dos mesmos. Tais padrões não são suficientes para a identificação de perfis de uso buscados. Assim, foi desenvolvida uma fase de Associação por meio da tarefa de Regras de Associação (ver Seção 2.2.1). Tal tarefa é aplicada com uso da matriz ω, resultante da fase de Absorção, buscando encontrar correlações positivas de uso de aplicativos. Desta forma, foi possível utilizar os padrões identificados como base para calcular a similaridade de uso de aplicativos entre os dispositivos monitorados visando a definição de perfis na fase de Caracterização. O Algoritmo 5.2 apresenta a execução desta fase.

Input

A função *ASSOCIATION* recebe como entrada a matriz ω que sumariza os eventos de uso de aplicativos, assim como valores de limites mínimos para algumas medidas: *minSup*, *minAllConf*, *minConf*, *minLift*, *maxLen*. Tais limiares são utilizados para gerar e selecionar os padrões de uso mais significantes em ω.

Output

Como saída, a função *ASSOCIATION* retorna uma coleção de conjunto de itens (*itemsets*) que são os padrões identificados.

Algoritmo 5.2: Fase de Associação do *framework f-DOPE*.

```
 1: function ITEMSET-GEN(ω, minSup, minAllConf, maxLen)
 2:     Items = itemsets = ∅                                   ▷ Variáveis auxiliares
 3:     Items = apriori-gen(ω, minSup, maxLen)        ▷ Função do Apriori (Agrawal et al., 1994)
 4:     for I ∈ Items do          ▷ Manter conjunto com all-confidence (Equação 2.9) mínimo
 5:         itemsets = itemsets ∪ I ∈ Items | all-confidence(I) > minAllConf
 6:     end for
 7:     return itemsets
 8: end function
 9: function RULES-GEN(itemsets, minSup, minConf, minLift)
10:     AR = rules = ∅                                         ▷ Variáveis auxiliares
11:     AR = apriori-genrules(itemsets, minSup, minConf)       ▷ Função do Apriori
12:     for X ⇒ Y ∈ AR do          ▷ Manter regras com lift (Equação 2.7) mínimo
13:         rules = rules ∪ X ⇒ Y ∈ AR | lift(X ⇒ Y) > minLift
14:     end for
15:     return rules
16: end function
17: function ASSOCIATION(ω, minSup, minAllConf, minConf, minLift, maxLen)  ▷ Função principal
18:     itemsets = rules = ∅                                   ▷ Variáveis auxiliares
19:     itemsets = itemset-gen(ω, minSup, minAllConf)   ▷ Chamada para gerar conjunto de itens
20:     rules = rules-gen(itemsets, minSup, minConf, minLift)  ▷ Chamada para gerar regras
21:     itemsets = apriori-generatingItemsets(rules)  ▷ Função do Apriori (Agrawal et al., 1994)
22:     return itemsets
23: end function
```

Associação

A força de uma regra de associação pode ser medida em termos de *suporte* (*minSup*), *all-confidence* (*minAllConf*), *confiança* (*minConf*) e *lift* (*minLift*) (Agrawal et al., 1994). Tais medidas são aplicadas ao longo desta fase para estimar a significância de cada regra com o objetivo de selecionar padrões interessantes. Além disso, para a execução de tal fase adotou-se o algoritmo *Apriori* (Agrawal et al., 1994) pois este é um algoritmo eficiente para mineração de regras de associação.

Inicialmente, a matriz ω é transformada em um conjunto de transações. Este conjunto indica por dispositivo quais atributos (aplicativos) são utilizados. Logo após, é necessário realizar a geração dos conjuntos de itens candidatos, que ocorre na função *ITEMSET-GEN* (linhas 1-8). Nesta função, é aplicada a rotina **apriori-gen** (linha 3) que é do algoritmo *Apriori* e visa a geração de tais conjuntos que são armazenados na variável *Items*. Em tal rotina é importante verificar o tamanho dos conjunto a serem gerados (*maxLen*). Tal verificação é realizada pois correlações de uso entre um número muito alto de itens distintos tornam a execução computacional lenta e onerosa. Além disso, é utilizado o limiar de *minSup* para gerar todos os possíveis conjuntos de itens candidatos frequentes. Por fim, para cada *itemset I* em *Items* (linha 4) são selecionados todos conjuntos de itens *I* que possuam $\textit{all-conf}(I) > \textit{minAllConf}$, buscando desconsiderar padrões de suporte cruzado (ver Apêndice C) (linha 5). Logo após, os limiares *minConf* e *minLift* juntamente com *minSup* são usados como entrada na função *RULES-GEN* (linhas 9-16). Em tal função, é empregada a rotina **apriori-genrules** também do algoritmo *Apriori* (linha 11). Nesta rotina, são geradas as regras de associação por meio dos conjuntos de itens (*itemsets*) gerados anteriormente, assim como dos limiares citados acima. As regras geradas são armazenadas na variável *AR*. Então, para cada regra $X \Rightarrow Y$ em *AR* (linha 12) são selecionadas todas as regras $X \Rightarrow Y$ que apresentem $\textit{lift}(X \Rightarrow Y) > \textit{minLift}$ (linha 13), as quas são armazenadas na variável *rules*, visando encontrar padrões com correlações positivas (ver Capítulo 2, Seção 2.2.1).

Conforme encontrado nos estudos de caso (ver Apêndice C), podem existir regras *redundantes* (Tan et al., 2006). Tais regras são, em muitos casos, regras similares, dado o aumento no tamanho máximo (*maxLen*) de tais regras. Estas regras apresentam os mesmos conjuntos de itens, os quais se apresentam em uma ordem distinta possuindo o mesmo valor de *suporte* e diferentes valores de *confiança* e de *lift*. Dado este cenário e os estudo de caso do Apêndice C, ao final desta fase os conjuntos de itens frequentes, os quais dão suporte as regras finais geradas e não apresentam tal redundância, são selecionados. Esta seleção é realizada pela rotina **apriori-generatingitemsets** do algoritmo *Apriori* (linha 21).

Em resumo, a geração de regras se faz importante dado que somente neste passo é possível definir os padrões que se correlacionam positivamente com a medida de *lift*. Sem a geração das regras, os conjuntos de itens iniciais não seriam filtrados podendo resultar em padrões negativos ou neutros. Além disso, para a definição do tamanho de regras e valores dos limiares, realizou-se um conjunto de estudos de caso que são apresentados no Apêndice C. Tais resultados foram validados com experimentos apresentados no Capítulo 6, Seção 6.3.2.

A complexidade computacional do Algoritmo 5.2 é relacionada ao número de itens possíveis, ao número de transações, ao tamanho das transações, a geração de conjunto de itens e a contagem de suporte. Em resumo, a geração de conjunto de itens pode ser muito expansivo e requer $O(i \times M \times maxLen)$ comparações, onde i é o número de transações (uma por dispositivo), $M = 2^p - 1$, onde p são os itens possíveis (aplicativos) e M é a quantidade de *itemsets*. Para a contagem de suporte o custo é $O(i \times \sum_p \binom{maxLen}{k} \times \alpha_p)$ onde α_p é o custo para atualizar a contagem de suporte de um conjunto de item candidato (Tan et al., 2006).

5.2.3 Fase de Caracterização

Na fase de Caracterização, os *itemsets* gerados pela fase de Associação são abordados como uma base de conhecimento para calcular a similaridade entre os padrões de uso dos dispositivos visando ajudar na identificação de perfis de uso. O Algoritmo 5.3 apresenta como tal fase é realizada.

Input

A função *IDENTIFICATION* recebe como entrada a matriz ω que sumariza os eventos de uso de aplicativos, bem como os *itemsets* que formaram as regras finais na fase de Associação.

Output

Ao final desta fase, a função *IDENTIFICATION* retorna a solução de agrupamento obtido ζ, o qual contém os rótulos dos dispositivos, bem como os dispositivos considerados *outliers* O e a matriz de distância Δ computada.

Identificação de perfis

Na função *DISTANCE* (linhas 1-17) é instanciada uma variável O (linha 2) que receberá os dispositivos i considerados *outliers*. Os *itemsets*, obtidos pela fase de Associação, são suportados por diferentes dispositivos e para que um dispositivo dê suporte para um determinado conjunto de itens ele deve apresentar uso de todos os aplicativos presentes em tal conjunto. Assim, um dispositivo é considerado um *outlier* quando este não utiliza todos os aplicativos de pelo menos um dos conjuntos de item existentes em *itemsets*. Assim, para cada dispositivo i em ω (linha 3) verifica-se os aplicativos p, os quais são atributos de ω, que apresentam tempo ou intervalo de uso por tal dispositivo i, ou seja, onde $\omega[i][p] \neq 0$, sendo tais aplicativos armazenados em UA (linha 4). Depois, verifica-se a ausência de interseção entre a potência do conjunto de aplicativos utilizados pelo dispositivo i ($P(UA)$) e os conjuntos de itens em *itemsets* (linha 5). Se existir uma interseção nada é modificado pois o dispositivo é suporte de ao menos um conjunto de itens. Caso contrário, o dispositivo é removido de ω (linha 6) e mantido em O (linha 7). Dispositivos considerados *outliers* geralmente representam uma parte mínima da população analisada (menos de 1% da população, ver Apêndice D). Essa falta de suporte mostra que estes dispositivos apresentam uso de poucos aplicativos *mais utilizados* ou usam aplicativos *mais utilizados* que não estão presentes nos *itemsets*.

Na sequência, é estruturada uma matriz de distância Δ (linha 10). Tal matriz possui $[|\omega[i]|][|\omega[i]|]$ dimensões, onde $|\omega[i]|$ é a quantidade de dispositivos em ω. Ela será computada (linhas 11-15) com base no conjunto de itens frequentes obtidos pela fase de Associação, com a aplicação do Índice de *Jaccard* apresentado pelo Capítulo 2, Seção 2.2.1, Equação 2.22, porém, com uma modificação (ver Equação 5.3). A mudança é a substituição da subtração ($1-$) existente no índice original, pela negação de uma função logarítmica ($-log$), o que proporciona um ganho

Algoritmo 5.3: Fase de Caracterização do *framework f-DOPE*.

1: **function** DISTANCE(ω, *itemsets*)
2: $O = \varnothing$ ▷ Variável para aramazenar *outliers*
3: **for** $i \in \omega[i]$ **do**
4: $UA = p \in \omega[p] | \omega[i][p] != 0 \wedge \omega[i] = i$ ▷ Identifica aplicativos utilizados pelo dispositivo
5: **if** $P(UA) \bigcap I \; \forall \; I \in itemsets = \varnothing$ **then** ▷ Verifica os aplicativos nos conjuntos de itens
6: $\omega = \omega - i$ ▷ Desconsidera dispositivo
7: $O = O \cup i$ ▷ Considera dispositivo *outlier*
8: **end if**
9: **end for**
10: $\Delta = matrix[[|\omega[i]|]][|\omega[i]|]]$ ▷ Matriz para armazenar similaridades
11: **for** i_1 *to* $i_{n-1} \in \omega[i]$ **do**
12: **for** $i_2^{'}$ *to* $i_n^{'} \in \omega[i]$ **do**
13: $\Delta_{i,i'} = dist(i, i^{'})$ ▷ Calcula similaridade entre dispositivos (Equação 5.3)
14: **end for**
15: **end for**
16: **return** Δ, O
17: **end function**
18: **function** CLUSTERING(Δ)
19: $evalRes = \varnothing$ ▷ Variável que armazena resultados dos k agrupamentos
20: **for** $k = 1$ *to* $k = \sqrt{n}$ **do**
21: $evalRes = evalRes \cup$ **GAP**(***WARD***(Δ)) ▷ Tarefa de Agrupamento
22: **end for**
23: $\zeta =$ **bestGAP**($EvalRes$) ▷ Definição do melhor número de k
24: **return** ζ
25: **end function**
26: **function** IDENTIFICATION(ω, *itemsets*)
27: Δ, $O = distance(\omega, itemsets)$ ▷ Função para calcular a similaridade
28: $\zeta = clustering(\Delta)$ ▷ Função para definição dos perfis
29: **return** ζ, Δ, O
30: **end function**

no intervalo de valores possíveis para as distâncias calculadas. Na Figura 5.2 é apresentado um exemplo do que ocorre com a aplicação das duas medidas dado um valor que representa a razão do tamanho da diferença entre os conjuntos (eixo x). Para a medida tradicional é utilizada a linha azul pontilhada, enquanto que para sua modificação é utilizada a linha vermelha. É possível perceber que conforme os valores do eixo x aumentam, as duas medidas apresentam resultados próximo. Por outro lado, quando os valores do eixo x são baixos, por exemplo, entre 0 e 0,5, existe uma perceptível diferença dos valores das duas medidas, onde a medida modificada apresenta valores maiores. Assim, as distâncias são calculadas com base nos conjuntos de itens para quais os dispositivos apresentam suporte de acordo com a Equação 5.3 (linha 13), onde o tamanho da interseção de tais conjuntos é dividido pelo tamanho da união destes conjuntos. Cada conjunto deve possuir todos aplicativos *remanescentes* que foram utilizados pelos dispositivos. Assim, a distância é calculada a cada par de dispositivos (i e i') com base em seus conjuntos de itens (I e I'), para os quais tais dispositivos são suporte. Neste caso, o conjunto I pode ou não conter os mesmo itens do conjunto de itens I' pois diferentes dispositivos podem utilizar diferentes aplicativos. O resultado do cálculo de distância, para dois dispositivos que sejam suporte para os os mesmos conjuntos de itens, será 0. Por outro lado, se dois dispositivos são suporte de conjuntos de itens diferentes, a distância tenderá ao infinito.

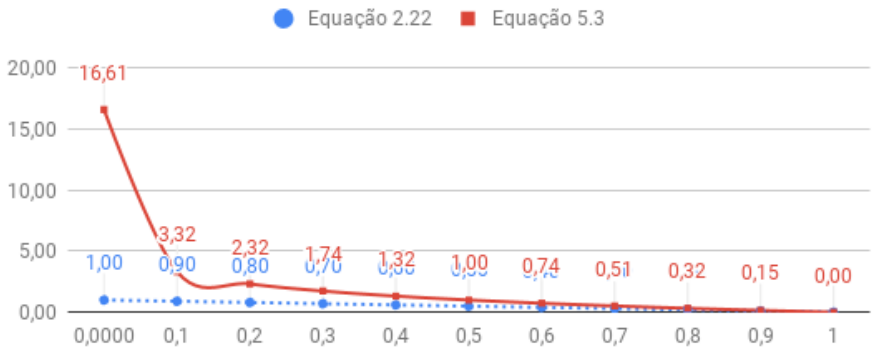


Figura 5.2: Comparação do intervalo de valores com aplicação do índice *Jaccard* original (1−) e sua modificação ($-log$).

$$dist(i, i') = -log\left(\frac{|I \cap I'|}{|I| + |I'| - |I \cap I'|}\right) \tag{5.3}$$

Em seguida (linhas 18-25), é realizada a tarefa de agrupamento (Ver Seção 2.2.1) visando a identificação de perfis de uso, onde a função *CLUSTERING* recebe como entrada a matriz de similaridade Δ (linha 18). Após é instanciada uma variável que armazenará os resultados da tarefa de Agrupamento (linha 19). A tarefa de agrupamento começa com $k = 2$, onde k é o número de perfis a serem formados, variando até $k = \sqrt{n}$ (MacQueen et al., 1967) (linha 20). De fato, deve-se aplicar o algoritmo escolhido $\sqrt{n} - 1$ vezes em Δ. Além disso, é necessário aplicar ao menos uma medida de avaliação para determinar o melhor número de perfis a serem formados. Nesse sentido, para cada k é necessário calcular a medida de avaliação (linha 21) para depois definir o melhor número de perfis em cada janela dado os resultados de tal medida para cada k (linha 23).

Apesar da existência de diferentes algoritmos de agrupamento a utilização de uma matriz de distância pré-computada impossibilita a execução de alguns destes algoritmos. Isso ocorre pois alguns algoritmos possuem sua própria função de similaridade. Para a execução da tarefa de agrupamento nesta fase foram adotados o algoritmo de agrupamento *WARD* e a medida de avaliação de agrupamentos *GAP*, os quais foram utilizados como exemplo no Algoritmo 5.3. A escolha deste algoritmo e desta medida foi dada por meio de estudos de caso apresentados no Apêndice D. Em resumo, os algoritmos hierárquicos, em conjunto com as medidas de avaliação testadas, apresentaram melhores resultados. Além disso, este tipo de algoritmo permite a possibilidade de observar dendrogramas que auxiliam na definição do melhor número de perfis. Entre os algoritmos hierárquicos testados estão *Complete Linkage*, que apresenta os piores resultados, seguido pelos algoritmos *Single Linkage* e *UPGMA* os quais apresentaram resultados similares mas piores em comparação com *WARD*. Além disso, algumas das medidas de avaliação testadas não contribuem na definição de mais de 3 perfis para o cenário abordado. Portanto, os melhores resultados foram obtidos com a combinação do algoritmo *WARD* e a medida *GAP*, indicando tais abordagens como as melhores para o cenário em questão.

Definição 4 *(*Perfis de Uso*) Um agrupamento ζ é uma partição de ω em perfis de uso $A_1, A_2, ..., A_k$ tal que $\forall\ \mathbf{p} \neq \mathbf{q}$: $A_\mathbf{p} \cap A_\mathbf{q} = \varnothing$, $O + \bigcup_{\mathbf{p}=1}^{k} A_\mathbf{p} = \omega$ e algum critério de otimização é satisfatório (ex: os objetos de um perfil A_k são mais similares entre si do que aos objetos em outro perfil).*

A Def. 4 assume um particionamento completo de Δ que é computado para todos os n dispositivos que suportam conjuntos de itens, ou seja, não são *outliers*. É importante lembrar que essa matriz é simétrica (Tan et al., 2006) podendo ser condensada em um vetor visando economizar espaço e processamento. Por exemplo, Δ pode ser um vetor onde as distâncias estão organizadas na ordem $(i_1, i_2), (i_1, i_3), ..., (i_1, i_n), (i_2, i_3), ..., (i_2, i_n), ..., ..., (i_{n-1}, i_n)$, uma vez que a distância na posição $[i][i']$ será a mesma na posição $[i'][i]$. O cálculo de Δ tem como principal objetivo representar a similaridade entre os dispositivos com base nos padrões de uso de aplicativos.

A complexidade computacional do Algoritmo 5.3 é extremamente relacionada ao número de entradas do conjunto de dados e do número k de grupos a serem formados (Wu et al., 2008). Em resumo, calcular a matriz de similaridade necessita de tempo $O(n^2)$. O tempo total necessário para a execução de algoritmos aglomerativos como *WARD* que armazenam e acompanham os grupos é $O(n^2 \times logn)$, onde *logn* é a complexidade adicional de manter os dados em uma lista classificada.

5.2.4 Saída da Etapa 1

Ao final desta etapa, uma solução de perfis de uso é obtida, ou seja, é gerado um conjunto de grupos $(A_1, A_2, ...A_k)$ que reúnem dispositivos $(i_1, i_2, ..., i_n)$. Estes grupos são subconjuntos do conjunto de entrada ω e em união aos *outliers* (O) formam tal conjunto por completo. Mesmo

adotando um algoritmo hierárquico para a execução da fase de Caracterização, onde os perfis são aninhados em forma de uma árvore hierárquica, a solução de perfis é composta por grupos que são subconjuntos não sobrepostos (Tan et al., 2006). Na etapa de Acompanhamento, propõe-se uma maneira de analisar as mudanças nos comportamento de uso de aplicativos com base nas mudanças e evoluções dos perfis obtidos. Os perfis devem ser investigados pois podem mudar, evoluir ou desaparecer ao longo das janelas analisadas. É importante notar que grupos, em janelas distintas, podem representar um único perfil. Em uma única janela cada rótulo representa um perfil diferente, enquanto rótulos diferentes podem representar o mesmo perfil em diferentes janelas. Isso pode ocorrer pelo fato de que o algoritmo de agrupamento rotula automaticamente cada um dos agrupamento. Contudo, tal situação não interfere na realização da etapa de Acompanhamento pois nela os perfis são investigados por meio de seus objetos. Além disso, quando um dispositivo deixa de enviar eventos de uso de aplicativos, este não é agrupado nem considerado *outlier* e sim um dispositivo ausente (M), o qual é explorado na próxima etapa.

5.3 Etapa 2: Acompanhamento do FCD

O acompanhamento dos perfis de uso é necessário pois os dados evoluem ao longo das janelas de evento. Nesse contexto, técnicas de detecção de novidade (Markou e Singh, 2003) (Ver Seção 3.3), como mudanças de conceito, visam ajudar no monitoramento de tais perfis. Além disso, também pode-se avaliar os diferentes tipos de comportamentos que os consumidores apresentam ao longo do tempo. Com base nos resultados obtidos por esta etapa, as empresas podem entender o comportamento de seus clientes ao longo do tempo e utilizar tal conhecimento para melhorar a tomada de decisões em ralação a produtos/serviços para tais clientes. Assim, cada fase da etapa de Acompanhamento do *f-DOPE* é proposta com a aplicação de técnicas de detecção de novidade buscando o monitoramento dos perfis de uso de aplicativos em dispositivos móveis em janelas de eventos adjacentes. Via estudos de caso, foram elaborados verificações que apontaram as melhores abordagens para cada uma das fases desta etapa. Os Apêndices E e F apresentam dois estudos de caso que indicam para a essencialidade de tais fases.

5.3.1 Fase de Monitoramento

De acordo com Markou e Singh (2003), a detecção de novidade pode identificar uma instância nova ou desconhecida ou um agrupamento delas, representando um conceito diferente daqueles aprendidos anteriormente. Para analisar os perfis de uso obtidos, pode-se empregar técnicas capazes de identificar possíveis mudanças entre eles. Algumas abordagens foram propostas por diferentes autores (Spiliopoulou et al., 2006; Oliveira e Gama, 2010c) (ver Seção 3.3.2). Uma técnica possível para o cenário de FCD é chamada de *enumeração* presente em MONIC (Spiliopoulou et al., 2006). Nesta abordagem, cada perfil identificado é monitorado ao longo das janelas de eventos,

onde é verificado o número de objetos que compõem tais perfis. Assim, para o *framework f-DOPE* o monitoramento da evolução dos perfis ao longo das janelas é realizada por meio do uso da técnica de *enumeração* (Spiliopoulou et al., 2006) (ver Seção 3.3.2). Tal abordagem recebe os perfis obtidos pela etapa de Mineração do FCD, de duas janelas adjacentes (por exemplo, $\zeta_{\mathbf{w}}$ e $\zeta_{\mathbf{w}+1}$) buscando encontrar as cinco transições externas possíveis. Dessa forma, de acordo com a variação de objetos nos perfis, sendo A um perfil de $\zeta_{\mathbf{w}}$ e B um perfil de $\zeta_{\mathbf{w}+1}$, tais perfis podem *sobreviver* ($A \rightarrow B$), se *dividir* ($A \overset{\subset}{\rightarrow} \{B_1, ..., B_k\}$), ser *absorvidos* ($A \overset{\subset}{\rightarrow} B$), *desaparecer* ($A \rightarrow \odot$) ou *surgir* ($\odot \rightarrow B$). É importante ressaltar que, após serem realizados testes com o algoritmo detector de *MONIC* (Spiliopoulou et al., 2006), verificou-se que o mesmo não realizava a detecção de surgimento de perfis, a qual foi incluída na adaptação utilizada no *f-DOPE*. Assim, o procedimento dessa fase ocorre de acordo com o Algoritmo 5.4 que foi adaptado de (Spiliopoulou et al., 2006).

Input

A função *MONITORING* recebe como entrada os agrupamentos $\zeta_{\mathbf{w}}$ e $\zeta_{\mathbf{w}+1}$ pertencentes, respectivamente, as janelas $w_{\mathbf{w}}$ e $w_{\mathbf{w+1}}$. Além disso, a função recebe os limiares τ_{match} e τ_{split}. Tais limiares são responsáveis pela detecção da evolução de conceitos.

Output

Como saída, a função *MONITORING* retorna as listas *deads, splits, absorptions, survivals, arisings* as quais representam as evoluções dos perfis em relação aos agrupamentos $\zeta_{\mathbf{w}}$ e $\zeta_{\mathbf{w}+1}$.

Monitoramento

Primeiramente, considere a janela de eventos $w_{\mathbf{w}}$, onde existe o perfil $A \in \zeta_{\mathbf{w}}$. Para detectar uma das transições para este perfil, é necessário verificar se A também existe por meio da averiguação dos grupos em $\zeta_{\mathbf{w}+1}$. Nesse sentido, deve-se investigar duas possibilidades, *sobreposição* e *equivalência* de perfis (Spiliopoulou et al., 2006).

Definição 5 *(*Sobreposição de perfil*) Assuma $A \in \zeta_{\mathbf{w}}$ e $B \in \zeta_{\mathbf{w}+1}$ sendo dois perfis, onde $\zeta_{\mathbf{w}}$ e $\zeta_{\mathbf{w}+1}$ são agrupamentos obtidos, respectivamente, nas janelas $\mathbf{w}, \mathbf{w}+1$. Uma "sobreposição" entre os grupos A e B é baseado nos objetos que formam a intersecção entre tais grupos de acordo com a Equação 5.4.*

$$sobreposicao(A, B) = \frac{A \cap B}{A} \tag{5.4}$$

Definição 6 *(*Equivalência de perfil*) Assuma τ_{match} como um valor percentual mínimo aceitável. $B \in \zeta_{\mathbf{w}+1}$ é considerado uma "equivalência" para $A \in \zeta_{\mathbf{w}}$ em $\zeta_{\mathbf{w}+1}$ se e somente se B for o perfil com o maior valor de "sobreposição" para A e sobreposicao(A, B) $\geqslant \tau_{match}$.*

Algoritmo 5.4: Fase de Monitoramento do *framework f-DOPE* adaptado de Spiliopoulou et al. (2006).

```
1: function MONITORING(ζ_w, ζ_{w+1}, τ_match, τ_split)
2:     deads = absorptionsSurvivals = splits = absorptions = survivals = ∅
3:     arisings = {B ∈ ζ_{w+1}}                                   ▷ ⊙ → B
4:     for A ∈ ζ_w do
5:         splitCandidates = splitUnion = ∅
6:         survivalCandidate = None
7:         survivalCandidateMcell = 0.0
8:         for B ∈ ζ_{w+1} do
9:             Mcell = sobreposicao(A, B)                          ▷ Pela Def. 5
10:            if Mcell ⩾ τ_match and Mcell > survivalCandidateMcell then
11:                survivalcandidate = B
12:                survivalCandidateMcell = Mcell
13:            else if Mcell ⩾ τ_split then
14:                splitCandidates += B
15:                splitUnion = {SplitUnion ∪B}
16:            end if
17:        end for
18:        if survivalCandidate == None e splitCandidates == ∅ then
19:            deads += A                                          ▷ A → ⊙
20:        else if survivalCandidate != None then
21:            absorptionsSurvivals += (A, survivalCandidate)
22:        else if splitCandidates != ∅ then
23:            if sobreposicao(X, splitUnion) ⩾ τ_match then
24:                for B ∈ splitCandidates do                      ▷ A ⊂→ splitCandidates
25:                    splits += (A, B)
26:                    arisings −= B
27:                end for
28:            else if survivalCandidate != None then
29:                absorptionsSurvivals += (A, survivalCandidate)
30:            else
31:                deads += A                                      ▷ A → ⊙
32:            end if
33:        end if
34:    end for
35:    for B ∈ ζ_{w+1} do
36:        absorptionCandidates = makeList(absorptionsSurvivals, B)
37:        if tamanho(absorptionCandidates) > 1 then
38:            for A ∈ absorptionCandidates do
39:                absorptions += (A, B)                           ▷ A ⊂→ B
40:                absorptionsSurvivals −= (A, B)
41:                arisings −= B
42:            end for
43:        else if absorptionCandidates == A then
44:            survivals += (A, B)                                 ▷ A → B
45:            absorptionsSurvivals −= B
46:            arisings −= B
47:        end if
48:    end for
49:    return deads, splits, absorptions, survivals, arisings
50: end function
```

Diferentes valores podem ser escolhidos para os limiares τ_{split} e τ_{match}. Contudo, recomenda-se que τ_{match} deva ser ao menos $0,5$ visando garantir que no máximo um perfil possa ser "equivalente". Por exemplo, τ_{match} pode ser um valor do intervalo $[0,5;\ 1.0]$, onde $0,5$ indica que um perfil deve possuir ao menos metade dos objetos do perfil ao qual esta sendo comparado. Além disso, o valor de τ_{split} deve ser menor que τ_{match}, pois caso não exista um equivalência para $A \in \zeta_{\mathbf{w}}$ uma divisão pode ter acontecido e os objetos de A estão distribuídos em perfis de $\zeta_{\mathbf{w}+1}$. Assim, a sobreposição não pode ser menor que τ_{split} o que evita um resultado incorreto no monitoramento. Pelas Def. 5 e Def. 6 é possível realizar um monitoramento entre perfis de duas janelas de eventos adjacentes visando identificar a existência de mudanças de conceitos.

A Figura 5.3 apresenta exemplos das cinco evoluções que podem ocorrer nesta fase. Se existir somente uma "equivalência" para um perfil $A \in \zeta_{\mathbf{w}}$ na janela $w_{\mathbf{w}+1}$, o perfil A vai sobreviver (linhas 43-47) (ver Figura 5.3 - a). Uma divisão pode ocorrer quando não for identificada uma "equivalência" e o valor de "sobreposição" for $\geqslant \tau_{split}$ (linhas 13-16) (ver Figura 5.3 - b). Por outro lado, se existirem mais "equivalências", o perfil A é absorvido (linhas 38-42) (ver Figura 5.3 - c). O perfil A vai desaparecer se nenhuma das variações acima forem detectadas (ver Figura 5.3 - d). Por fim, no início do Algoritmo 5.4, todos os perfis da janela $w_{\mathbf{w}+1}$ são considerados como perfis que estão surgindo (linha 3). Porém, quando algum destes perfis for indicado como uma sobrevivência, divisão ou absorção de um perfil pertencente a janela $w_{\mathbf{w}}$, tal perfil será removido da lista de perfis que estão surgindo (linhas 26, 41 e 46) (ver Figura 5.3 - e). Experimentos foram realizados para a definição dos melhores valores de τ_{match} e τ_{split} no cenário abordado por esta pesquisa. Tais resultados são descritos no Capítulo 6 e comparados com resultados por meio da aplicação da abordagem empregadas na literatura que mais se assemelha ao *f-DOPE*.

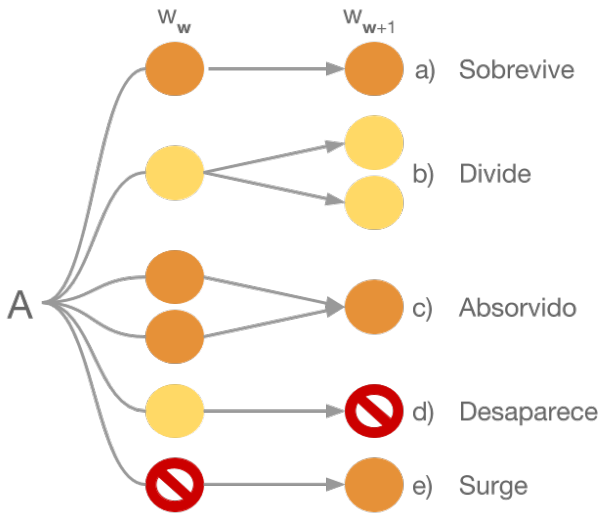


Figura 5.3: Possíveis evoluções de conceitos que podem ocorrer com os perfis de uso na fase de Monitoramento do *framework f-DOPE*.

Esta fase permite a detecção do mudanças de conceitos (Gama, 2010). A abordagem aqui adotada foi aplicada em estudos de caso a fim de validar sua utilização (ver Apêndice E). Além disso, sem a realização desta fase, e a identificação das possíveis variações, não seria possível identificar as

mudanças nos comportamentos de uso de aplicativos, assim como segmentar tais comportamentos, o qual é proposto para a última fase do *framework f-DOPE*.

A complexidade computacional do Algoritmo 5.4 está principalmente relacionada a computação da *sobreposição* para cada par de grupos. Assim, caso ocorram melhoramentos, como o calculo da sobreposição sendo realizada em uma matriz pré-definida para posterior recuperação da *Mcell* compatível, a complexidade será $O(k^2)$ para $k = max(|\zeta_{\mathbf{w}}|, |\zeta_{\mathbf{w}+1}|)$, onde $|\zeta_{\mathbf{w}}|$ é a quantidade de grupos em $\zeta_{\mathbf{w}}$.

5.3.2 Fase de Segmentação

A última fase, chamada de Segmentação, também é responsável pela detecção de novidades. Porém, nesta fase busca-se acompanhar as mudanças nos comportamentos de uso de aplicativos pela investigação das evoluções dos perfis ao longo do tempo. Propõe-se uma nova maneira de investigação das mudanças nos comportamentos de uso de aplicativos em relação às mudanças e evoluções dos conceitos encontrados. Aqui pretende-se adquirir conhecimento visando ajudar na compreensão de todos possíveis comportamentos apresentados pelos dispositivos monitorados. Além disso, é possível segmentar dispositivos com comportamentos similares e identificar comportamentos que possam ser importantes na tomada de decisão das empresas fabricantes de dispositivos móveis. O algoritmo 5.5 demonstra como esta fase é realizada.

Input

Na última fase do *framework f-DOPE* a função *SEGMENTATION* recebe como entrada os agrupamentos $\zeta_{\mathbf{w}}$ e $\zeta_{\mathbf{w}+1}$, assim como as evoluções dos perfis em relação a tais agrupamentos: *absorptions*, *survivals*, *deads*, *splits* e também a lista de *outliers* $O_{\mathbf{w}+1}$.

Output

Ao final da execução desta fase a função *SEGMENTATION* retorna as listas que representam os comportamentos (*L - Leal*, *C - Mudou* e *M - Ausente*) que foram detectados com base: i) nas evoluções dos perfis e ii) nas mudanças de perfis apresentadas pelos dispositivos, ambas são detectadas por meio dos padrões de uso de aplicativos dos dispositivos.

Segmentação

O comportamento *L* ocorre quando um dispositivo pertencente a um perfil $B \in \zeta_{\mathbf{w}+1}$, o qual representa o mesmo perfil $A \in \zeta_{\mathbf{w}}$, ao qual este mesmo dispositivo estava agrupado ($A \rightarrow B$ ou $A \overset{\subseteq}{\rightarrow} B$) (linhas 12, 17 e 31). Por sua vez, o comportamento *C* ocorre quando um dispositivo que pertence a um perfil $B \in \zeta_{\mathbf{w}+1}$, que não representa o mesmo perfil $A \in \zeta_{\mathbf{w}}$, ao qual este dispositivo estava agrupado ($A \overset{\subseteq}{\rightarrow} \{B_1, ..., B_k\}$ ou $A \rightarrow \odot$) (linhas 11, 16, 23, 30 e 35). O comportamento

O ocorre quando um dispositivo não apresenta suporte para para os conjunto de itens frequentes gerados na fase de Associação quando $\zeta_{\mathbf{w}+1}$ foi obtido (identificados por meio do Algoritmo 5.3). Por fim, o comportamento M acontece quando um dispositivo foi considerado *desaparecido*, ou seja não apresentou utilização de aplicativos quando $\zeta_{\mathbf{w}+1}$ foi obtido. Desta forma, tal dispositivo não é agrupado nos perfis de $\zeta_{\mathbf{w}+1}$ e também não pode ter sido identificado como *outlier* O (linha 34).

Algoritmo 5.5: Fase de segmentação do *framework f-DOPE*.

```
1: function SEGMENTATION(ζ_w, ζ_{w+1}, absorptions, survivals, deads, splits, O_{w+1})
2:     L = C = M = ∅                          ▷ Variáveis para armazenar dispositivos conforme ações
3:     for A ∈ ζ_w do
4:         lc = ∅                             ▷ Varável auxiliar para perfis que representam ações leais (L)
5:         if A ∈ A' ∀ (A', B') ∈ absorptions then          ▷ Verifica se A foi absorvido
6:             for (A', B') ∈ absorptions do
7:                 if A == A' then
8:                     lc += B'                              ▷ B' absorveu A
9:                 end if
10:            end for
11:            C += i | i ∈ A ∧ i ∉ B' ∀ B' ∈ lc     ▷ Dispositivos em A que mudaram de perfil
12:            L += i | i ∈ A ∧ i ∈ B' ∀ B' ∈ lc     ▷ Dispositivos em A com mesmo perfil
13:        else if A ∈ A' ∀ (A', B') ∈ survivals then       ▷ Verifica se A sobreviveu
14:            for (A', B') ∈ survivals do
15:                if A == A' then
16:                    C += i | i ∈ A ∧ i ∉ B'    ▷ Dispositivos em A que mudaram de perfil
17:                    L += i | i ∈ A ∧ i ∈ B'    ▷ Dispositivos em A com mesmo perfil
18:                end if
19:            end for
20:        else if A ∈ deads then                          ▷ Verifica se A desapareceu
21:            for A' ∈ deads do
22:                if A == A' then
23:                    C += i | i ∈ A             ▷ Dispositivos em A que mudaram de perfil
24:                end if
25:            end for
26:        else if A ∈ A' ∀ (A', B') ∈ splits then         ▷ Verifica se A se dividiu
27:            for (A', B') ∈ splits do
28:                lc += B'
29:            end for
30:            C += i | i ∈ A ∧ i ∉ B' ∀ B' ∈ lc     ▷ Dispositivos que mudaram de perfil
31:            L += i | i ∈ A ∧ i ∈ B ∀ B ∈ lc       ▷ Dispositivos com mesmo perfil
32:        end if
33:    end for
34:    M += i | i ∉ B ∀ B ∈ ζ_{w+1} ∧ i ∉ O_{w+1}      ▷ Dispositivos desaparecidos
35:    C += i | i ∉ C ∧ i ∉ O_{w+1} ∧ i ∉ M ∧ i ∉ L
36:    return L, C, M
37: end function
```

A complexidade computacional do Algoritmo 5.5 está principalmente relacionada ao número de grupos da janelas analisadas similar ao que ocorre com o algoritmo 5.4. Desta forma, a complexidade seria $O(k^2)$ para $k = max(|\zeta_{\mathbf{w}}|, |\zeta_{\mathbf{w}+1}|)$.

5.3.3 Saída da Etapa 2

Ao final desta etapa e consequentemente do *frameowork f-DOPE* tem-se como resultado os padrões de uso (*itemsets*), os perfis de uso (ζ), as mudanças de conceitos (*deads*, *splits*, *absorptions*, *survivals*, *arisings*) e os comportamentos (*L*, *C*, *M*, *O*). Além destas saídas, propõe-se a criação de *ciclos comportamentais*, os quais são uma representação temporal baseada em comportamentos apresentados ao longo da etapa de Acompanhamento do FCD pelos dispositivos. Tais comportamentos podem ser similares mesmo entre usuários de perfis diferentes (ver Figura 5.4). Isso é possível, dado aos tipos de comportamentos projetados (*L*, *C*, *M* e *O*).

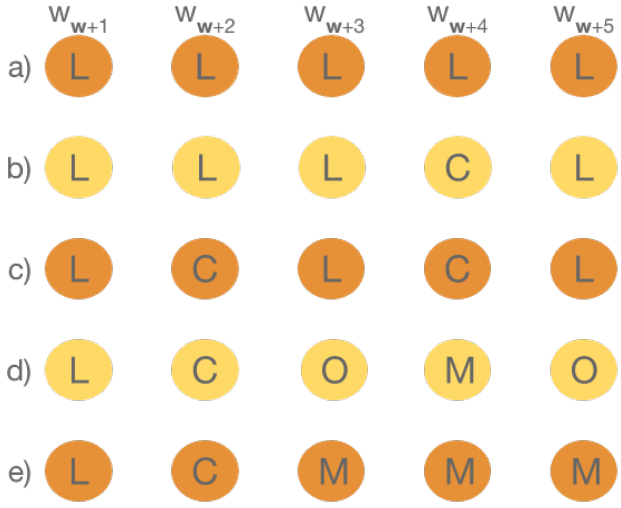


Figura 5.4: Alguns *ciclos comportamentais* que podem ocorrer na fase de Segmentação do *framework f-DOPE*.

A Figura 5.4 apresenta exemplos de *ciclos comportamentais* que são possíveis comportamentos de um ou mais dispositivos. Por exemplo, tais *ciclos* podem demonstrar a continuidade de um dispositivo em um único perfil, a mudança para diferentes perfis ou o desaparecimento de tal dispositivo do FCD. Como pode-se observar, dada a execução de *F-DOPE* em uma janela de eventos inicial **w** e também em uma ou mais janelas adjacentes ($\mathbf{w}+1$, $\mathbf{w}+2$, ..., $\mathbf{w}+5$, ...) é possível estabelecer *ciclos* que indicam o comportamento temporal dos dispositivos. Em resumo, entende-se que *clientes leais* (por exemplo, que apresentam comportamentos *L*) (ver Figura 5.4 - a, b) são aqueles dispositivos que, em sua maioria, são agrupados em perfis de uso de aplicativos similares ao longo de várias janelas de eventos. Por outro lado, *clientes atípicos* são aqueles que mudam de perfil, em muitas janelas de eventos, ou então, são *outliers* (ver Figura 5.4 - c, d). De fato, é difícil agrupar todos os *clientes atípicos* em um único grupo, principalmente porque seus perfis mudam de janela em janela. Por fim, estes clientes também podem ser caracterizados pela mudança esporádica de perfil de uso entre duas janelas adjacentes. Estes consumidores não necessariamente apresentam os mesmos perfis ao longo do FCD, podendo indicar alguma dificuldade na adaptação de um novo produto ou uma insatisfação existente. Neste sentido, comportamentos *C*, *M*, *O* são minorias e muitos consumidores que apresentam comportamentos *M* em uma janela, acabam não voltando a

apresentar eventos de uso de aplicativos podendo ser considerados como clientes que abandonaram a marca (ver Figura 5.4 - e). Após a execução de *f-DOPE* em uma quantidade mínima aceitável de janelas eventos, o qual deve ser decidido dado a capacidade computacional e também ao FCD, podem ser investigados todos os possíveis *ciclos comportamentais*, visando segmentar aqueles *ciclos* que indiquem comportamentos indesejados que fabricantes de dispositivos móveis estejam buscando. Por exemplo, é possível acompanhar um determinado tipo de perfil para investigar um tipo específico de cliente (ex: gamers) ou então buscar clientes que mudam de comportamento ($\neg\ L$) para entender suas dificuldades. Por fim, experimentos com uso dos *ciclos comportamentais* são apresentados no Capítulo 6. Resultados da execução do *f-DOPE* são comparados com resultados da abordagem da literatura que mais se assemelha ao *f-DOPE* visando demonstrar que a aplicação do *f-DOPE* melhora a detecção de perfis e comportamentos. Por fim, além da elaboração de tais *ciclos*, foi desenvolvido um sistema de monitoramento de comportamento de usuários visando possibilitar a visualização dos comportamentos ao longo do tempo, o qual é demonstrado no Apêndice F.

5.4 Considerações Finais do Capítulo

Nesta pesquisa, a identificação e o monitoramento de perfis e comportamentos de uso de aplicativos em dispositivos móveis são abordados, onde consumidores, por meio de seus dispositivos, apresentam diferentes padrões de uso de aplicativos bem como de comportamentos ao longo de um FCD. Assim, o *framework* proposto é dividido em duas etapas principais:

- Na primeira etapa, algoritmos de Mineração de Dados e Aprendizado não supervisionado são aplicadas com o objetivo de:

  (a) Pré-processar e sumarizar o FCD.

  (b) Obter padrões de uso de aplicativos por meio da tarefa de Regras de Associação;

  (c) Agrupar dispositivos de acordo com suas similaridades, calculadas com base nos padrões obtidos, por meio da tarefa de Agrupamento.

  A aplicação destes algoritmos visa extrair os melhores padrões dos dados analisados, considerando a enorme quantidade de eventos que o uso de aplicativos produz. A primeira etapa tem como resultado a caracterização de diferentes perfis de uso em cada janela de eventos.

- Na segunda etapa, técnicas de Detecção de Novidade são empregadas ao longo das janelas de eventos, visando:

  (d) O monitoramento dos perfis de uso (conceitos) por meio da investigação de mudanças e evoluções destes conceitos (perfis) identificados;

  (e) O monitoramento de comportamentos buscando rastrear modificações de tais comportamentos ao longo das janelas de eventos;

Em muitos cenários de FCDs do mundo real, é necessário investigar e acompanhar as variações, não só dos perfis, mas também dos indivíduos inseridos nestes perfis. Neste sentido, a segunda etapa visa também a investigação e monitoramento dos diferentes comportamentos identificados permitindo a segmentação de *ciclos comportamentais* desejados dado a necessidade de uma parte interessada.

Em geral, dado o nosso melhor conhecimento, o *framework f-DOPE* desenvolvido ao longo desta pesquisa é diferente do que foi encontrado na literatura. É possível a combinação de FCD de eventos de uso de aplicativos em dispositivos móveis e a segmentação de comportamentos baseados em padrões de uso destes aplicativos, tudo sobre várias janelas de eventos em um longo período de tempo. São explorados:

1. uma grande quantidade de eventos em FCD, um grande número de dispositivos e um grande número de aplicativos
2. padrões de uso de aplicativos
3. perfis de uso
4. mudança e evolução de perfis e comportamentos
5. a segmentação e o monitoramento do comportamentos

Assim, no presente capítulo foi apresentado o *framework* desenvolvido, chamado *f-DOPE* incluindo: a metodologia utilizada e a arquitetura desenvolvida que foi elaborada visando explorar os pontos fortes e satisfazer as limitações identificadas na literatura. Do ponto de vista da ciência da computação, é investigada uma aplicação de mineração e monitoramento de FCDs que pode ter impacto sobre a indústria de dispositivos móveis.

Por fim o *f-DOPE* é aplicado em experimentos que estão descritos no Capítulo 6. Estes experimentos foram realizados visando demonstrar a qualidade de cada uma das fases do *f-DOPE*, bem como comprovar sua qualidade na identificação e monitoramento de perfis e comportamentos em FCD de uso de aplicativos em dispositivos móveis.

6. RESULTADOS EXPERIMENTAIS

Este Capítulo tem como objetivo apresentar os experimentos realizados com o *framework f-DOPE*, introduzido no Capítulo 5, e compará-lo com a metodologia empregada na literatura que mais se aproxima de tal *framework*. Os experimentos exploram um FCD real de uso de aplicativos em dispositivos móveis, o qual é descrito na Seção 6.1. Na Seção 6.2 é apresentado o plano de experimento. Os resultados dos experimentos com o *f-DOPE* são apresentados na Seção 6.3. Por sua vez, a Seção 6.4 descreve os resultados dos experimentos aplicando a metodologia da literatura. A avaliação dos resultados referente a ambas execuções é apresentada na Seção 6.5. Na Seção 6.6 é realizada uma predição de comportamentos baseada na saída do *f-DOPE* em comparação com a predição de comportamentos baseada na saída da abordagem da literatura, bem como uma avaliação estatística. Por fim, a Seção 6.7 apresenta as considerações finais do Capítulo.

6.1 FCD de uso de aplicativos em dispositivos móveis *DS03*

Esta Seção descreve o terceiro FCD capturado da empresa patrocinadora desta pesquisa. Tal FCD é diferente dos FCD apresentados no Apêndice A, os quais foram utilizados nos estudos de caso ao longo desta pesquisa. Este FCD contém dados capturados durante o período de 140 dias. Tal período teve início em 05 de Junho finalizando em 22 de Outubro de 2017. O FCD contêm 807.584.951 eventos de uso de aplicativos oriundos de 21.392 dispositivos móveis[1] pertencentes ao mesmo modelo de dispositivo e ao mesmo país de origem, onde 81.557 aplicativos distintos foram utilizados em tais dispositivos. Por fim, cada evento de uso de aplicativo do FCD chega com a tupla ε (ε_i, ε_p, ε_d, ε_{et}) conforme descrito no Capítulo 1, Seção 1.3.

A Tabela 6.1 resume todos os eventos do FCD capturado. O número de dispositivos ativos diminui à medida que o período de tempo aumenta. Isso é causado pelo fim do envio de eventos de uso do aplicativo por alguns dispositivos. Neste sentido, considera-se um dispositivo ativo quando ele envia pelo menos um evento de uso de aplicativos em alguma janela de tempo (um dia, um mês, dois meses, três meses e quatro meses). Por exemplo, de 21.392 dispositivos, 3 deles apresentaram eventos de uso do aplicativo em um único dia e, portanto, não foram contabilizados como dispositivos ativos para janelas maiores que 1 dia. Esse comportamento provavelmente representa uma rotatividade, que ocorre quando os clientes param de usar ou alteram seus dispositivos móveis conforme mencionado no Capítulo 1, Seção 1.1.

É possível verificar uma quantidade significativa de eventos (807.584.951). No total, este FCD armazena ocupa cerca de 100 GB de dados (sem compactação). É importante lembrar que aplicativos móveis nativos não precisam ser abertos para gerar eventos de uso (ver Apêndice A). Nesse sentido, as informações apresentadas na Tabela 6.1 não consideram eventos de uso deste tipo de aplicativo pois tais eventos podem ser gerados sem a ação de um usuário no dispositivo.

[1] Não foram adicionados novos dispositivos com o passar do tempo.

Informação	Quantidade
Número de eventos	807.584.951
Aplicativos únicos	81.557
Dispositivos únicos	21.392
Tempo de duração	140 dias
Dispositivos ativos (> 1 dia)	21.388
Dispositivos ativos (> 1 mês)	21.217
Dispositivos ativos (> 2 meses)	21.025
Dispositivos ativos (> 3 meses)	20.770
Dispositivos ativos (> 4 meses)	20.379

Tabela 6.1: Visão geral do FCD de uso de dispositivos móveis *DS03* utilizado nos experimentos finais.

6.2 Plano de Experimento

Esta seção apresenta o plano de experimento, o qual divide-se em três etapas. O objetivo final deste experimento é comparar os resultados com aplicação do *f-DOPE* e da abordagem da literatura, em diferentes aspectos. Ao longo dos experimentos são comparados resultados como: distribuição dos perfis, variação dos perfis ao longo do FCD e tipos de comportamentos. Mais ainda, é realizada uma comparação da previsão de comportamentos de uso de aplicativos pelos dispositivos ao final de 10 janelas de eventos. Tal predição explora a saída do *f-DOPE* e a saída da abordagem da literatura e estas predições são comparadas. Esta comparação tem como principal objetivo demonstrar que o *f-DOPE* é sensível quanto a detecção e o monitoramento de perfis e comportamentos.

Para os experimentos optou-se pela utilização de uma janela de eventos de Marcação (ver Seção 3.2.2) para a captura de eventos ao longo do FCD. Como todos os eventos de uso de aplicativos possuem um tempo final do uso (ε_{et}), a utilização deste tipo de janela permite que todos dispositivos sejam monitorados por um mesmo período de tempo em vez de uma mesma quantidade de eventos. A utilização dos demais tipos de janelas existentes é possível. Contudo, a utilização de uma janela onde o tamanho é determinado de acordo com a quantidade de eventos pode fazer com que um dispositivo com poucas atividades (Figura 1.1 (a) - dispositivo 3) seja monitorado por mais tempo que um dispositivo que possui muitas atividades (Figura 1.1 (a) - dispositivo 1).

Além disso, foi adotado como 7 o número de dias que compõem a janela de marcação. A escolha de 7 dias têm como base: i) o tempo total em que os dados são armazenados pela empresa patrocinadora desta pesquisa e ii) uma quantidade mínima desejável de janelas buscando conseguir monitorar perfis e comportamentos encontrados. Por exemplo, utilizando uma janela de 7 dias (uma

semana) 10 janelas podem ser utilizadas para identificar e monitorar os perfis e os comportamentos gerando *ciclos comportamentais*, não ultrapassando o tempo em que os dados são normalmente armazenados pelas empresas fabricantes de dispositivos móveis. Além disso, uma semana representa um período de tempo médio em comparação com outras possibilidades de tamanho. Uma janela pequena (ex: 1 dia) pode não ser suficiente para identificar todas as atividades realizadas normalmente em um dispositivo uma vez que o uso de aplicativos pode mudar a cada dia. Por outro lado, uma janela grande (ex: 1 mês) pode ser acima do necessário visto que com mais tempo consequentemente existe uma quantidade maior de atividades em cada dispositivo.

Na Seção 6.2.1 é apresentada a configuração do *f-DOPE*. A Seção 6.2.2 descreve a configuração da metodologia adotada na literatura. Por fim, a Seção 6.2.3 diz respeito à avaliação dos experimentos realizados.

6.2.1 Configuração do *f-DOPE*

O *framework* proposto é executado conforme apresentação realizada no Capítulo 5 e tem a seguinte configuração para cada uma das suas fases:

- ***Fase de Absorção***: Optou-se como aplicativos *mais utilizados* aqueles usados por 1% ou mais dispositivos analisados ($\tau_{most} = 0{,}01$). Decidiu-se pela remoção dos aplicativos que não são *remanescentes* para cada dispositivo com o uso dos limiares $\tau_{rem} = 0,01$ e $minTime = 70$. Estipulou-se como aplicativos *populares* aqueles utilizados por ao menos 10% dos dispositivos ($\tau_{pop} = 0,10$), sendo tais aplicativos discretizados com a técnica *IP* conforme descrito no Capítulo 2, Seção 2.1.2.

- ***Fase de Associação***: Optou-se pela existência de no máximo 4 itens nos conjuntos de itens gerados ($minLen = 4$). Decidiu-se por limiares de *suporte* ($minSup = 0,01$), *all-confidence* ($minAllConf =$ média de todos valores de *all-confidence* obtidos), *confiança* ($minConf = 0,10$) e *lift* ($minLift = 1$).

- ***Fase de Caracterização***: Realização do cálculo da matriz de distância entre os dispositivos com base nos conjuntos de itens encontrados na fase de Associação de acordo com a Equação 5.3. Para a tarefa de Agrupamento, estipulou-se o uso do algoritmo *Ward* em conjunto com a medida de avaliação *GAP* dados os resultados dos estudos de caso do Apendice D.

- ***Fase de Monitoramento***: Utilizou-se a abordagem por *enumeração* para o monitoramento dos perfis. Os limiares τ_{match} e τ_{split} foram variados a cada $0,05$ entre os intervalos de $[0,50; 0,90]$ para τ_{match} e $[0,10; 0,40]$ para τ_{split} a fim de encontrar a combinação que apresentasse melhores resultados.

- ***Fase de Segmentação***: Optou-se por utilizar a verificação dos *ciclos comportamentais* buscando encontrar os comportamentos existentes. Tais ciclos são ações (L, C, O, M) com

base nas mudanças de conceitos encontradas na fase de Monitoramento. Este conjunto de *ciclos comportamentais* também é utilizado para validar a previsão de comportamentos na Seção 6.6.

6.2.2 Configuração da Metodologia Empregada na Literatura

Dada a revisão sistemática da literatura descrita no Capítulo 4, os trabalhos de Spiliopoulou et al. (2006), Oliveira e Gama (2010c) e Pereira e Mendes-Moreira (2016) são os que mais se assemelham ao objetivo buscado por esta pesquisa. Tais trabalhos, em alguns casos, utilizam as mesmas técnicas e algoritmos. Assim, foi realizada a seguinte configuração para execução de uma abordagem da literatura:

- ***FCD***: Primeiramente, para que pudesse haver uma comparação e uma equivalência de uma abordagem da literatura com *framework f-DOPE*, optou-se em utilizar o FCD *DS03* após parte da execução das etapas de pré-processamento e sumarização, o qual é realizado pela fase de Absorção do *f-DOPE*. Assim, os atributos não foram discretizados e transformados, permanecendo com sua formatação inicial onde tais dados são contínuos. Com o grande volume de eventos, o formato do FCD e as evidências encontradas pelos estudos de caso percebeu-se que essa seria a melhor abordagem a ser aplicada.

- ***Fase de Caracterização*** Os trabalhos da literatura que mais se assemelham a este trabalho empregam diferentes algoritmos de agrupamento. Contudo, *X-Means* (Pereira e Mendes-Moreira, 2016) e *bisecting K-Means* (Spiliopoulou et al., 2006) se assemelham em sua forma de execução. Além disso, ambos algoritmos aplicam uma medida que visa avaliar o melhor número de grupos a serem obtidos juntamente à execução do agrupamento. Assim, optou-se por utilizar o algoritmo *X-Means* o qual aplica a medida *BIC* como forma de avaliação do melhor número de grupos, onde a distância euclidiana é utilizada no cálculo das distâncias ao longo da execução do algoritmo *X-Means*.

- ***Fase de Monitoramento***: Os trabalhos da literatura que mais se assemelham a este trabalho empregam as duas abordagens encontradas para o monitoramento de perfis: *enumeração* (Spiliopoulou et al., 2006; Oliveira e Gama, 2010c) e *compreensão* (Oliveira e Gama, 2010c; Pereira e Mendes-Moreira, 2016). Como ambas abordagens são utilizadas e não existe uma dominante, dado que Oliveira e Gama (2010c) utilizam as duas abordagens, optou-se por utilizar a mesma abordagem aplicada no *f-DOPE*. Tal abordagem faz mais sentido no cenário abordado por investigar os objetos agrupados, enquanto que a abordagem por *compreensão* avalia as características dos perfis formados. Assim, os limiares τ_{match} e τ_{split} também foram variados a cada $0,05$ entre os intervalos de $[0,50;\ 0,90]$ para τ_{match} e $[0,10;\ 0,40]$ para τ_{split} a fim de encontrar a melhor configuração.

- ***Fase de Segmentação***: Dado o objetivo de prever o comportamento dos dispositivos optou-se em empregar a verificação dos *ciclos comportamentais* da mesma forma como é realizado para o *f-DOPE*. Tais *ciclos* não alteram as resultados anteriores. Pelo contrário, são representações de tais resultados. Além disso, ambos resultados podem ser comparados com essas opções e os *ciclos comportamentais* gerados são utilizados para validar a previsão de comportamentos na Seção 6.6.

6.2.3 Avaliação dos Resultados

Para a avaliação dos resultados, considerando a aplicação do FCD *DS03* e de acordo com as configurações descritas acima, são observadas as seguintes métricas:

- ***Distribuição dos perfis***: É verificado o número de perfis obtidos a cada janela, observando-se a distribuição dos objetos nos perfis obtidos.
- ***Variação dos perfis ao longo do FCD***: É avaliada a evolução dos perfis ao longo do tempo e também verifica-se a quantidade de mudanças de conceitos detectadas.
- ***Tipos de comportamentos***: São verificados a quantidade e os tipos de *ciclos comportamentais* encontrados.
- ***Conjunto de métricas aplicadas na exploração da saída do f-DOPE em comparação com a saída do X-Means***: Métricas preditivas baseadas em matrizes de confusão são investigadas para verificação do desempenho das predições realizadas com base nas saídas do *f-DOPE* em comparação ao *X-Means*. Além disso, uma avaliação estatística é aplicada em tal comparação.

6.3 Execução do *f-DOPE*

Nesta Seção são apresentados os resultados da execução de experimentos por meio da aplicação do *framework* proposto, conforme configuração descrita na Seção 6.2.1, no FCD *DS03*.

6.3.1 Fase de Absorção

Inicialmente, o FCD *DS03* contém informações de um total de 81.557 aplicativos, sendo 21.078 da primeira janela. É possível notar a existência de um número substancial de aplicativos utilizados por apenas um único dispositivo, enquanto outros aplicativos foram abertos em apenas alguns dispositivos. Como pode ser visto na Figura 6.1, aproximadamente 58% (12.203) dos aplicativos foram utilizados por apenas um dispositivo e quase 91% (<= 1 e <= 10) de aplicativos foram

abertos no máximo em 10 dispositivos. De fato, apenas 76 aplicativos (<= 10.000, <= 20.000 e <= 25.000) são utilizados em mais de 10.000 dispositivos. Isso reforça a necessidade de selecionar apenas aplicativos utilizados por um número significativo de dispositivos (Def. 1), os quais apresentam uma quantidade mínima e necessária de atividades e para executar uma análise mais realista do uso do aplicativos (Li et al., 2015).

Foi investigado o FCD *DS03* visando verificar a distribuição do número de aplicativos utilizados pelos dispositivos. Da mesma maneira que se havia verificado ao longo dos estudos de caso, tal distribuição se apresentou muito similar (ver Figura 6.1). Neste caso, fica evidente que é necessário definir um conjunto de aplicativos mais significativos. Para validar tal necessidade, verificou-se a Contribuição Cumulativa (*CC* - Equação 6.1) baseada em todos os aplicativos encontrados para cada janela. A Figura 6.2 mostra os valores de *CC* para duas análises distintas, i) a linha vermelha tracejada mostra os principais aplicativos com base no número total de dispositivos exclusivos e ii) a linha verde sólida mostra os aplicativos baseado no seu tempo total de uso considerando todos os dispositivos (S_p). O primeiro aplicativo na linha vermelha tracejada é o aplicativo usado pelo maior número de dispositivos únicos, e sua contribuição $\sum_{\varepsilon \epsilon S} \varepsilon_d$ é de aproximadamente 10% em relação a todo uso de aplicativos em D ($\sum_{\varepsilon \epsilon D} \varepsilon_d$). Além disso, o primeiro aplicativo na linha verde sólida é o aplicativo com o maior tempo de uso, e sua contribuição é de cerca de 25%.

$$CC = \sum_{S} \left(\left(\frac{\sum_{\varepsilon \epsilon S} \varepsilon_d}{\sum_{\varepsilon \epsilon D} \varepsilon_d} \right) \times 100 \right) \tag{6.1}$$

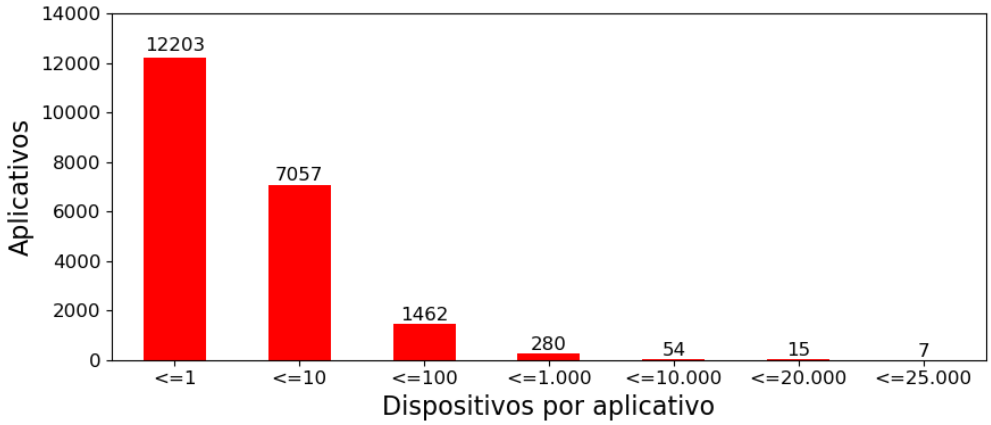


Figura 6.1: Distribuição do número de aplicativos utilizados por dispositivos únicos na primeira semana do FCD *DS03*.

Assim, conforme a Def.1 apresentada no Capítulo 5, Seção 5.2.1 é válida a possibilidade de indicação dos aplicativos *mais utilizados* com base no número de clientes únicos a cada janela de eventos. O *CC* apresentado na Figura 6.2 indica que a seleção dos 113 aplicativos mais usados fornece cobertura de mais de 90% do tempo total de uso de todos os aplicativos nessa janela (linha lilas tracejada). Deste modo, é possível afirmar que a escolha de aplicativos pelo número de dispositivos únicos que os utilizam pode definir os aplicativos *mais utilizados*. Portanto, considera-se aqueles aplicativos utilizados por 1% ou mais dispositivos ($\tau_{most} = 0{,}01$). A Tabela 6.2 apresenta

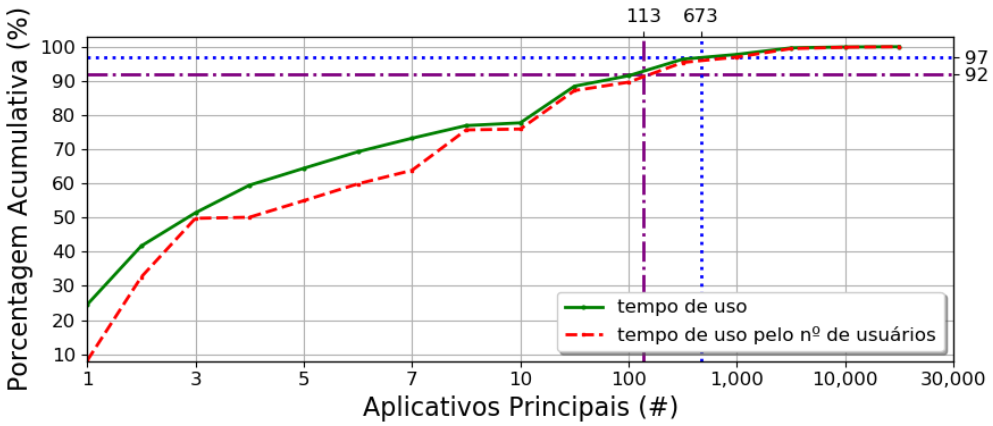


Figura 6.2: Métricas para definição dos aplicativos *mais utilizados* por janela de eventos. O cruzamento das linhas tracejadas violetas destaca a porcentagem de *CC* para os 113 aplicativos mais utilizados entre os dispositivos, enquanto que o cruzamento das linhas pontilhadas azuis indica a porcentagem de *CC* para os 673 aplicativos mais utilizados em tais dispositivos.

o total de aplicativos encontrados para as 10 primeiras janelas do FCD, bem como a porcentagem destes aplicativos em relação a todos os aplicativos detectados ao longo de tais janela. Além disso, a Tabela 6.2 apresenta o número total de dispositivos e eventos para cada semana.

O tempo de uso dos aplicativos *mais utilizados* varia de 1,03 segundos a 133 horas para uma determinada janela de tempo. Além disso, cada usuário apresenta, em média, 23 aplicativos diferentes representando um número bastante alto de acordo com Annie (2017). Nesse sentido, os aplicativos *mais utilizados* são ainda mais explorados. De acordo com a Def. 2 apresentada no Capítulo 5, Seção 5.2.1, quando um aplicativo possuir um tempo total de utilização menor que 1% de *UTUT* ($\tau_{rem} = 0,01$) para tal dispositivo, ou então ser utilizado por menos de 70 segundos (*minTime* $= 70$), o mesmo é desconsiderado para tal dispositivo. Com a aplicação desse filtro, o número médio de aplicativos utilizados por dispositivos chega a 15 por janela, indicando a conveniência da realização desta filtragem.

Alguns dos aplicativos restantes são utilizados por mais de 40% dos dispositivos enquanto outros são abertos em mais de 90% dos dispositivos. Esses aplicativos podem ser utilizados por diferentes frequências de tempo e precisam ser tratados de forma diferente. Assim, optou-se por reconhecer como *populares*, apenas os aplicativos que são utilizados por 10% ou mais dispositivos ($\tau_{pop} = 0,10$). O número dos aplicativos *populares*, selecionados neste experimento também é apresentado na Tabela 6.2. É possível verificar que existem aplicativos muito utilizados, o que vai ao encontro com a Def. 3 apresentada no Capítulo 5, Seção 5.2.1.

Além desta opção, os aplicativos *populares* foram discretizados pela técnica *IP* (Han et al., 2011). Uma amostra dos aplicativos *populares* discretizada na primeira janela é apresentado na Tabela 6.3. Ao final da discretização, em vez de transformar cada intervalo de um aplicativo popular em um novo atributo, o que aumentaria a dimensionalidade do FCD, decidiu-se por resumir os dados usando o rótulo de intervalo de cada dispositivo como o valor para cada aplicativo. Por exemplo, o aplicativo *Whatsapp*, o qual foi dividido em cinco intervalos (1_5, 2_5, ..., 5_5), sendo

Janela	Dispositivos		Aplicativos						Eventos
dd/mm	Total	%	Total	%	Mais utilizados	%	Populares	%	
05/06-11/06	21.392	100,00%	21.078	25,84%	110	0,52%	12	0,06%	41.900.761
12/06-18/06	21.237	99,28%	21.451	26,30%	111	0,52%	14	0,07%	49.012.797
19/06-25/06	21.130	98,78%	21.099	25,87%	110	0,52%	12	0,06%	42.663.529
26/06-02/07	21.089	98,58%	21.197	25,99%	111	0,52%	12	0,06%	40.137.897
03/07-09/07	21.018	98,25%	20.611	25,27%	107	0,52%	12	0,06%	39.659.524
10/07-16/07	20.974	98,05%	21.051	25,81%	112	0,53%	12	0,06%	40.342.428
17/07-23/07	20.909	97,74%	21.333	26,16%	114	0,53%	12	0,06%	39.824.520
24/07-30/07	20.840	97,42%	21.565	26,44%	113	0,52%	12	0,06%	39.702.367
31/07-06/08	20.780	97,14%	21.337	26,16%	113	0,53%	12	0,06%	39.599.677
07/08-13/08	20.736	96,93%	21.282	26,09%	114	0,54%	12	0,06%	44.394.108
μ	21.011	98,22%	21.200	25,99%	112	0,53%	12	0,06%	41.723.761
σ	197	0,92%	252	0,31%	2	0,01%	1	0,00%	2.864.687

Tabela 6.2: Um resumo das primeiras dez semanas do FCD *DS03*. O número total e a porcentagem de dispositivos, aplicativos, aplicativos *mais utilizados*, aplicativos *populares*, assim como o número de eventos para cada semana. No fim, a média e o desvio-padrão de cada elemento.

que 1_5 representa o intervalo com o menor tempo total de utilização e 5_5 representa o intervalo com maior tempo total de utulização. Tais representações mostram, no primeiro valor o intervalo em que o tempo total de uso do aplicativo é inserido, (1, 2, ..., 5), enquanto que o segundo valor representa o total de intervalos que o aplicativo foi discretizado (5). Os aplicativos mais *utilizados* não discretizados têm um intervalo exclusivo (1_1) que representa o uso do aplicativo, mas sem discretização em intervalos. Por fim, os dispositivos, que não estão incluídos em intervalos, possuem um tempo total de utilização para o aplicativo em questão igual a zero.

Pkg do aplicativo	Dispositivos %	IP intervalos
whatsapp	98%	5
facebook.katana	83%	4
android.apps.photos	60%	6
facebook.orca	18%	5
netflix.mediaclient	11%	4

Tabela 6.3: O número de intervalos de alguns dos aplicativos *populares* discretizados pela técnica *IP* na primeira janela do FCD *DS03*.

Discussão

Os resultados descritos nesta seção referem-se à execução do Algoritmo 5.1 apresentada no Capítulo 5, Seção 5.2.1. Visando avaliar os aplicativos *mais utilizados*, investigou-se tais aplicativos

com base em a) o número de dispositivos em que o aplicativo foi usado e b) o tempo total de uso do aplicativo em todos os dispositivos que o usam. No primeiro caso, é possível identificar a aceitação de um novo aplicativo ou o declínio de um aplicativo antigo. Por exemplo, se o número de dispositivos em que um aplicativo *p* for utilizado representar 20% dos dispositivos, e esse número aumentar significativamente em outras janelas, provavelmente *p* está ganhando aceitação. Assim, as partes interessadas podem usar esse tipo de conhecimento para diferentes fins. Podem, por exemplo, sugerir esses aplicativos quando o número de dispositivos que usam tais aplicativo aumenta significativamente. Ou então, sugerir a remoção da memória quando o número de dispositivos que usam um aplicativo diminuir. O segundo caso permite observar as alterações que ocorrem com um aplicativo em janelas posteriores. No entanto, é importante observar que um aplicativo pode ser aberto e usado muitas vezes em um único dia por apenas um dispositivo.

O tempo total de uso de um aplicativo é a soma do tempo de uso de eventos de uso de aplicativo independentes em vários dispositivos. Pode-se observar a ampla variação do tempo total de uso de um aplicativo em outras janelas, mas essa alteração pode não indicar precisamente um aumento ou uma diminuição da aceitação do aplicativo. No entanto, a avaliação do tempo de uso do aplicativo para um único dispositivo pode ser realizada para oferecer suporte na identificação de perfis de consumidores. Assim, foram utilizados os 110 aplicativos *mais utilizados* da primeira semana e os 111 da segunda semana para um breve comparativo. As Figuras 6.3 e 6.4 mostram *treemaps* dos aplicativos mais *utilizados* para tais semanas, respectivamente. Estas *treemaps* são compostas de quadrados com diferentes tamanhos e cores. Cada quadrado representa um dos aplicativos *mais utilizados*. A cor e sua intensidade indicam o número de dispositivos em que o aplicativo foi usado, variando de muitos dispositivos (amarelo) a poucos dispositivos (roxo). O tamanho do quadrado representa o tempo total de uso de cada aplicativo, em que os grandes quadrados indicam aplicativos amplamente utilizados e os menores, aplicativos menos utilizados.

Como pode ser visto nas Figuras 6.3 e 6.4, o número de dispositivos para cada aplicativo varia de acordo com a janela de tempo. Por exemplo, o YouTube foi usado por 18.745 dispositivos na primeira janela, enquanto 17.181 dispositivos o usaram na segunda. Observa-se que alguns aplicativos diminuem o número de dispositivos de uma janela para outra, principalmente os 10 principais aplicativos, como o Chrome e o Instagram. Contudo, vários aplicativos, a maioria deles entre os 10 principais, aumentaram o número de dispositivos na segunda semana (por exemplo, CartolaFC, Gmail, Tinder e outros). Para fornecer uma visão melhor sobre a variação de uso e facilitar a identificação de ícones, é possível verificar na Tabela 6.4 detalhes sobre alguns dos principais aplicativos e a variação de posição para cada tipo de investigação. Percebe-se também que o tempo total de uso também muda de uma janela para outra da mesma forma que aconteceu com o número de dispositivos. Por exemplo, o WhatsApp foi usado por 8.954.554.852 minutos na primeira semana (Figura 6.3) e por 6.933.872.036 minutos na segunda (Figura 6.4). Mesmo que muitos aplicativos tenham diminuído o tempo total de uso, vários aplicativos aumentaram seu tempo de uso, como Deskclock, Tinder, Aliexpress e outros. Mesmo com as alterações no tempo de uso total, encontramos os mesmos 10 principais aplicativos nas duas janelas de tempo. No entanto, dos

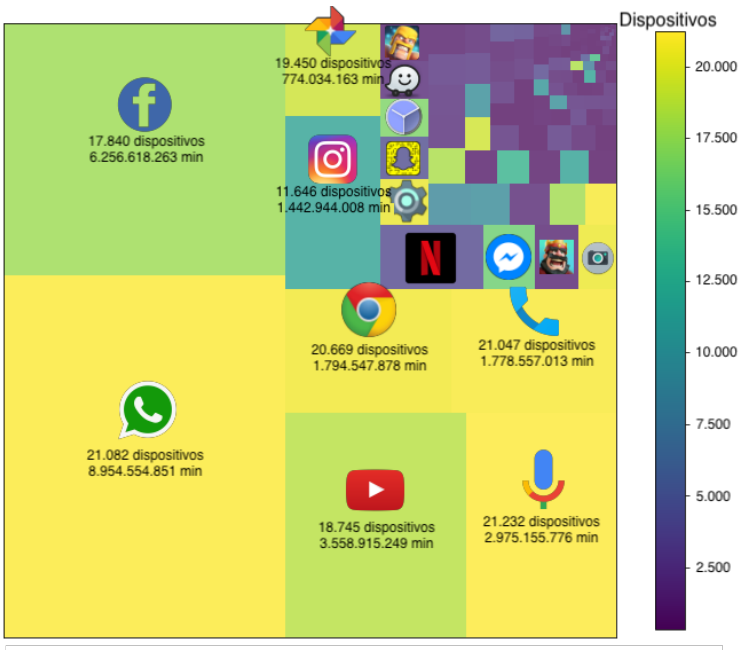


Figura 6.3: *TreeMap* dos aplicativos *mais utilizados* na primeira janela do FCD *DS03*. O tempo total de uso de todos aplicativos é 26.806.132.530 minutos e o total de dispositivos é 21.392.

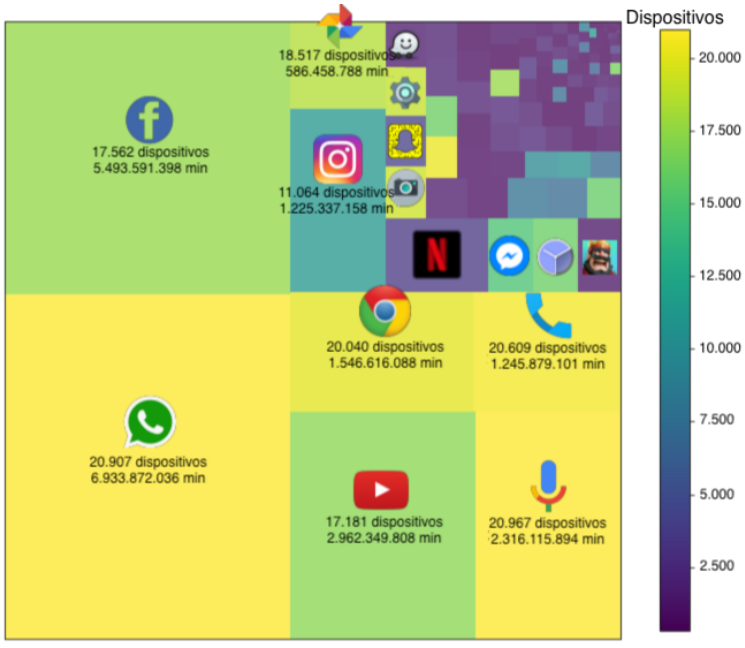


Figura 6.4: *TreeMap* dos aplicativos *mais utilizados* na segunda janela do FCD *DS03*. O tempo total de uso de todos aplicativos é 33.051.975.705 minutos e o total de dispositivos é 21.237.

11 aplicativos *mais utilizados* na primeira semana, vários aplicativos mudaram a posição na segunda semana. Por exemplo, o aplicativo 11 da primeira semana (Clash Royale) foi superado pelo aplicativo 12 (Desk Clock) na segunda semana. É importante destacar que a maioria dos principais aplicativos são os mesmos em ambas as janelas de tempo. De fato, alguns aplicativos podem desaparecer em algumas janelas devido à diminuição do uso, o que é um comportamento típico em cenários de FCD.

Dado os resultados obtidos na fase de Absorção do *f-DOPE* é possível verificar que os mesmos vão ao encontro de trabalhos encontrados na literatura. Neste caso, o trabalho de Li et al.

Nome do *Pkg*	Nome do App	Dispositivos (a)		Tempo de Uso (b)		Ícone
		1° janela	2° janela	1° janela	2° janela	
whatsapp.com	WhatsApp	2	2	1	1	
com.facebook.katana	Facebook	14	10	2	2	
com.google.android.youtube	Youtube	10	11	3	3	
com.google.android.googlequicksearchbox	Google Search	1	1	4	4	
com.android.chrome	Google Chrome	5	5	5	5	
com.android.dialer	Dialer	3	3	6	6	
com.instagram.android	Instagram	20	18	7	7	
com.google.android.apps.photos	Google Photos	9	8	8	8	
com.netflix.mediaclient	Netflix	31	31	9	9	
com.facebook.orca	Facebook Messenger	16	16	10	10	
com.supercell.clashroyale	Clash Royale	67	56	11	12	
com.android.camera	Camera	6	6	12	13	
com.snapchat.android	Snapchat	30	29	14	14	
com.google.android.deskclock	Desk Clock	15	14	15	11	
br.com.mobits.cartolafc	CartolaFC	45	37	25	25	
com.google.android.gm	Gmail	21	20	32	32	
com.tinder	Tinder	79	68	45	44	
com.nianticlabs.pokemongo	Pokemon Go	127	114	48	45	
com.alibaba.aliexpresshd	Aliexpress	71	76	66	55	

Tabela 6.4: Alguns dos aplicativos *mais utilizados* encontrados ao longo dos experimentos. Suas identificações de pacotes, ícone e posição em relação ao número de dispositivos únicos e tempo total de uso para as duas primeiras semanas do FCD *DS03*.

(2015) apresenta resultados similares no que diz respeito a definição de aplicativos *mais populares*. Além disso, Ferreira et al. (2014) indicam a necessidade de definir um tempo mínimo de utilização de aplicativos, o que vai ao encontro com o que é descrito por Annie (2017). Por outro lado, Li et al. (2015) e Xu et al. (2011) citam os aplicativos altamente utilizados, os quais são chamados de *populares* nesta pesquisa, indicando o tratamento diferenciado para tais aplicativos. Por fim, estudos de caso foram realizados (ver Apêndice B) aplicando diferentes técnicas para encontrar os melhores parâmetros para esta fase. Assim, dado os resultados do experimentos realizados e o cenário desta pesquisa é possível afirmar que os processos de pré-processamento e sumarização presentes para esta fase do *f-DOPE* são realmente importantes para FCD de uso de aplicativos em dispositivos móveis.

6.3.2 Fase de Associação

Com o objetivo de obter diferentes padrões de uso de aplicativos, foram realizados experimentos com Algoritmo 5.2 apresentado no Capítulo 5, Seção 5.2.2. Ademais, optou-se pelo

algoritmo *Apriori*, introduzido por Agrawal et al. (1994), o qual se mostrou eficiente para lidar com problemas de Regras de Associação de Mineração no cenário abordado desde o início desta pesquisa.

No início desta fase, a estrutura de dados sumarizada ω, a qual é gerada ao final da fase anterior, para cada janela de tempo, é transformada em um conjunto de transações. Neste caso, a Tabela 6.5 apresenta alguns exemplos das transações geradas na primeira janela, onde n representa o número de dispositivos observados em tal janela. Com relação à função *ITEMSET-GEN* do Algoritmo 5.2, passo (a) da geração de regras de associação (ver Seção 2.2.1), decidiu-se gerar no máximo conjuntos de itens de tamanho 4 ($maxLen = 4$). Tal escolha foi motivada pelos estudos de caso realizados ao longo da pesquisa (ver Apêndice C) e pelo fato de que conjuntos de itens muito grandes tem um custo computacional elevado com pouquíssimo benefício para a geração dos padrões. Além disso, conjuntos de itens maiores geram muitas regras semelhantes, nas quais os itens são ordenados de maneira diferente, contendo o mesmo suporte. No mesmo sentido, optou-se pelos limiares $minSup = 0,01$ e *minAllConf* como a média de *all-confidence* de todos os conjuntos de itens gerados. Na Tabela 6.6 são apresentados, o número de conjuntos de itens e a quantidade de conjunto com valores de *all-confidence* maior que o *minAllConf* encontrados na primeira janela do FCD *DS03*.

TID	Itens
1	whatsapp=3_5, facebook.katana=2_4, android.apps.maps=0, ...
2	whatsapp=1_5, facebook.katana=0, android.apps.maps=0, ...
3	whatsapp=2_5, facebook.katana=1_4, android.apps.maps=1_1, ...
n	...

Tabela 6.5: Parte de algumas das transações de uso de aplicativos da primeira janela de eventos do FCD *DS03* utilizadas como entrada para o Algoritmo 5.2.

Para a obtenção das regras, função *RULES-GEN*, passo (b) da geração de regras, utilizou-se o limiar *minSup* igual ao escolhido no passo anterior e estipulou-se, *minConf* como 0,10 e *minLift* = 1. Na Tabela 6.7 são apresentadas algumas regras geradas na primeira janela do FCD e seus valores de *suporte*, *confiança* e *lift*. Durante os estudos de caso (ver Apêndice C) verificou-se redundância de algumas regras que não melhoraram os resultados encontrados. Dado tal situação, são investigados os conjuntos de itens que geraram as regras finais. Assim, na função ***apriori-generatingItemsets*** do algoritmo *Apriori*, os conjuntos de itens que geraram as regras com valor de *lift* maior que *minLift* são capturados para a continuação do *f-DOPE*. A Tabela 6.6 mostra o total de regras obtidas para cada janela de eventos do FCD *DS03*, bem como o total de regras com *lift* maior que o *minLift*. Além disso, o número total de conjuntos de itens finais para cada semana também é apresentado na Tabela 6.6. Ademais, são apresentados na Tabela 6.8 alguns dos conjuntos de itens finais e seus valores de *suporte* e *all-confidence* encontrados na primeira janela do FCD *DS03*.

Janela	Conjuntos de itens			Regras			Conjuntos Finais	
dd/mm	Total	> minAllConf	%	Total	> minLift	%	Total	%
05/06-11/06	2.842	758	26,67%	1.774	1.403	79,09%	635	22,34%
12/06-18/06	2.650	688	25,96%	1.576	1.215	77,09%	569	21,47%
19/06-25/06	2.586	687	26,57%	1.579	1.258	79,67%	568	21,96%
26/06-02/07	2.578	683	26,49%	1.543	1.180	76,47%	553	21,45%
03/07-09/07	2.421	683	28,21%	1.508	1.167	77,39%	559	23,09%
10/07-16/07	2.424	684	28,22%	1.502	1.152	76,70%	562	23,18%
17/07-23/07	2.580	675	26,16%	1.542	1.204	78,08%	563	21,82%
24/07-30/07	2.284	649	28,42%	1.414	1.093	77,30%	534	23,38%
31/07-06/08	2.571	680	26,45%	1.557	1.253	80,48%	571	22,21%
07/08-13/08	2.388	674	28,22%	1.503	1.166	77,58%	553	23,16%
μ	2.532	686	27,14%	1.550	1.209	77,98%	567	22,41%
σ	150	26	0,94%	88	79	1,26%	25	0,70%

Tabela 6.6: Resumo do processo de mineração de regras de associação com o Algoritmo 5.2. O total e a porcentagem de conjunto de itens, dos conjuntos de itens selecionados pelo limiar de *all-confidence*, das regras, das regras selecionadas pelo limiar de *lift* e dos conjuntos de itens finais para cada janela de eventos do FCD *DS03*. No fim, a média e o desvio-padrão de cada elemento.

X	Y	Suporte	Confiança	Lift
clashroyale=1_1	clashofclans=1_1	0,014	0,303	9,268
facebook.katana=3_4	facebook.orca=1_5	0,011	0,201	1,314
⋱	⋱	⋱	⋱	⋱
chrome=2_4	youtube=2_4	0,015	0,143	1,314

Tabela 6.7: Algumas regras geradas pela função *ITENSET-GEN* do Algoritmo 5.2 na primeira janela do FCD *DS03*.

Conjunto de Itens	Suporte	all-confidence
facebook.katana=2_4, whatsapp=2_5	0,118	0,345
chrome=1_4, android.youtube=1_4	0,300	0,529
⋱	⋱	⋱
chrome=1_4, android.dialer=1_4, facebook.orca=1_5	0,046	0,080

Tabela 6.8: Alguns dos conjuntos de itens gerados ao final da execução do Algoritmo 5.2 na primeira janela do FCD *DS03*.

Discussão

Os resultados descritos nesta Seção referem-se à execução do Algoritmo 5.2 apresentado no Capítulo 5, Seção 5.2.2. A fase de Associação tem por objetivo a execução do algoritmo de Mineração de Regras de Associação visando obter diferentes padrões de uso de aplicativos móveis,

os quais são utilizados como base para a definição de similaridade de uso de aplicativos entre os dispositivos para uma identificação de perfis de uso.

Primeiramente, a definição do tamanho máximo dos conjuntos de itens a serem gerados (*maxLen*) é primordial. Conjunto de itens muito grandes podem fazer com que o processo seja lento sem que haja benefícios, uma vez que se está trabalhando com grandes conjuntos de transações. A escolha de *minSup* elevado pode fazer com que conjuntos de itens, envolvendo itens raros, sejam perdidos. Por outro lado, se o valor de *minSup* for muito baixo, o número de conjunto de itens candidatos pode ser muito grande (Tan et al., 2006; Han et al., 2011). A definição de *minSup* = $0,01$ corresponde ao mesmo valor mínimo indicado para a seleção dos aplicativos *mais utilizados* (1%). Para as definições de limiares para *minAllConf* e *minSup* não existe na literatura fórmulas para calcular tais valores. Mesmo alguns trabalhos avaliando essas e outras medidas (Tew et al., 2014; Tan et al., 2006; Han et al., 2011), cada cenário se comporta de uma maneira diferente, o que dificulta ambas as avaliações destes valores. Um dos trabalhos que investiga mineração de regras de associação mas não visa identificar e monitorar perfis, é o trabalho de Tseng e Hsu (2014). Em tal trabalho são encontradas padrões similares aos achados nesta fase em um cenário muito similar com o desta pesquisa. Por fim, foram realizados estudos de caso (ver Apêndice C) visando encontrar os melhores parâmetros para esta fase. Desta forma, com os resultados aqui descritos fica evidente a existência de correlação no uso de aplicativos e que esta verificação é significante para a obtenção de perfis de uso, o que é verificado na Seção 6.3.3.

6.3.3 Fase de Caracterização

Para a fase de Caracterização realizada de acordo com o Algoritmo 5.3 apresentado no Capítulo 5, Seção 5.2.3, os conjuntos de itens obtidos pela fase anterior são usados como uma base de conhecimento para encontrar perfis de uso de aplicativos. Tal escolha ocorreu a partir dos resultados dos estudo de caso descritos no Apêndice D. Em tais estudos foram investigados diferentes algoritmos e formas de abordar os eventos dos FCD de uso de aplicativos em dispositivos móveis. Contudo, nos estudos em que foram aplicadas categorias de aplicativos, com e sem o pré-processamento do FCD, os resultados não foram promissores. Mesmo utilizando algoritmos e medidas comumente utilizados pelos trabalhos relacionados (por exemplo, *K-means e Ward*), não foi possível identificar mais de 3 perfis bem como variações de tais perfis ao longo do tempo com tais abordagens. Somente com a adição das fases de Absorção de Associação foi alcançado uma quantidade de perfis promissores assim como as variações de tais perfis ao longo do tempo, o que é esperado dado a mudança na distribuição dos dados.

Para os experimentos aqui descritos, inicialmente criou-se uma matriz de distância entre os dispositivos Δ, que é baseada na Equação 5.3 descrita no Capítulo 5, Seção 5.2.3. Após o cálculo e a estruturação de Δ, a tarefa de Agrupamento é executada. Para tal tarefa optou-se pelo algoritmo Ward (Jardine e Sibson, 1971) e a medida de avaliação GAP (Tibshirani et al., 2001) que juntos apresentaram o melhor resultado para o cenário abordado. Além disso, é importante lembrar que os

dispositivos que somente apresentam suporte para conjuntos de itens que foram descartados durante a execução da fase de Associação (ver Seção 6.3.2), são considerados *outliers*. Esses dispositivos representam em média 0,78% dos dispositivos analisados. A Tabela 6.9 apresenta o número de perfis definidos, bem como o número total e a porcentagem de dispositivos *outliers* para cada janela do FCD *DS03* com a aplicação do *f-DOPE*.

Janela	*f-DOPE*			X-Means Perfis
dd/mm	*Perfis*	Outliers	%	
05/06-11/06	6	117	0,56%	8
12/06-18/06	5	161	0,75%	8
19/06-25/06	5	163	0,77%	8
26/06-02/07	4	171	0,81%	8
03/07-09/07	7	169	0,82%	8
10/07-16/07	5	158	0,75%	8
17/07-23/07	4	159	0,75%	8
24/07-30/07	7	191	0,89%	8
31/07-06/08	5	171	0,80%	8
07/08-13/08	5	199	0,94%	8
μ	5	166	0,78%	8
σ	1	21	0,10%	0

Tabela 6.9: Quantidade de perfis identificados em cada janela do FCD *DS03* para os os experimentos realizados com o *f-DOPE* e com a metodologia da literatura *X-Means*, assim como o número total e a porcentagem de *outliers* encontrados pelo *f-DOPE*. No fim, a média e o desvio-padrão de cada elemento.

Ao final desta fase, todos os dispositivos foram mapeados para um dos k perfis formados. Além disso, é possível dizer que cada perfil representa padrões de uso de aplicativo específicos, os quais são baseados nos conjuntos de itens encontrados pela fase de Associação (ver Seção 6.3.2). Por fim, uma avaliação dos resultados da aplicação da primeira etapa do *f-DOPE* é realizada nas Seções 6.5.1 e 6.5.2. Em resumo, tal etapa permite a identificação de muitos perfis, possibilitando que tais perfis e padrões de uso possam ser monitorados ao longo do tempo.

6.3.4 Fase de Monitoramento

Depois de obter os perfis por meio da etapa anterior, é possível investigar as mudanças de conceitos de uma janela para outra usando a abordagem *enumeration* (Spiliopoulou et al., 2006). Foram realizados experimentos variando τ_{match} de $0,50$ a $0,90$ e τ_{split} de $0,10$ a $0,40$, ambos a cada $0,05$. Os resultados das mudanças de conceito entre a primeira e a segunda janela do FCD *DS03* dado os perfis obtidos, sendo 6 para ζ_w e 5 para ζ_{w+1}, são apresentadas na Tabela 6.10.

A Figura 6.5 apresenta o número de perfis que surgiram ao longo do FCD *DS03*. Tal verificação foi possível pela adaptação da abordagem de *enumeração* conforme apresentado no Capítulo 5, Secção 5.3.1, Algoritmo 5.4. Assim, dado os resultados obtidos para as duas primeiras

τ_{match}	τ_{split}	$A \overset{\subset}{\to} B$	$A \to B$	$A \to \odot$	$A \overset{\subset}{\to} \{B_1, ..., B_k\}$	$\odot \to B$	τ_{match}	τ_{split}	$A \overset{\subset}{\to} B$	$A \to B$	$A \to \odot$	$A \overset{\subset}{\to} \{B_1, ..., B_k\}$	$\odot \to B$
0,50	0,10	0	1	0	5	0	0,70	0,30	0	0	5	1	0
0,50	0,15	0	1	0	5	0	0,70	0,35	0	0	5	1	0
0,50	0,20	0	1	1	4	0	0,70	0,40	0	0	6	0	0
0,50	0,25	0	1	2	3	0	0,75	0,10	0	0	0	6	0
0,50	0,30	0	1	4	1	0	0,75	0,15	0	0	1	5	0
0,50	0,35	0	1	4	1	0	0,75	0,20	0	0	5	1	0
0,50	0,40	0	1	5	0	0	0,75	0,25	0	0	5	1	0
0,55	0,10	0	0	0	6	0	0,75	0,30	0	0	5	1	0
0,55	0,15	0	0	0	6	0	0,75	0,35	0	0	5	1	0
0,55	0,20	0	0	1	5	0	0,75	0,40	0	0	6	0	0
0,55	0,25	0	0	3	3	0	0,80	0,10	0	0	1	5	0
0,55	0,30	0	0	5	1	0	0,80	0,15	0	0	4	2	0
0,55	0,35	0	0	5	1	0	0,80	0,20	0	0	6	0	0
0,55	0,40	0	0	6	0	0	0,80	0,25	0	0	6	0	0
0,60	0,10	0	0	0	6	0	0,80	0,30	0	0	6	0	0
0,60	0,15	0	0	0	6	0	0,80	0,35	0	0	6	0	0
0,60	0,20	0	0	1	5	0	0,80	0,40	0	0	6	0	0
0,60	0,25	0	0	3	3	0	0,85	0,10	0	0	2	4	0
0,60	0,30	0	0	5	1	0	0,85	0,15	0	0	5	1	0
0,60	0,35	0	0	5	1	0	0,85	0,20	0	0	6	0	0
0,60	0,40	0	0	6	0	0	0,85	0,25	0	0	6	0	0
0,65	0,10	0	0	0	6	0	0,85	0,30	0	0	6	0	0
0,65	0,15	0	0	0	6	0	0,85	0,35	0	0	6	0	0
0,65	0,20	0	0	3	3	0	0,85	0,40	0	0	6	0	0
0,65	0,25	0	0	5	1	0	0,90	0,10	0	0	2	4	0
0,65	0,30	0	0	5	1	0	0,90	0,15	0	0	6	0	0
0,65	0,35	0	0	5	1	0	0,90	0,20	0	0	6	0	0
0,65	0,40	0	0	6	0	0	0,90	0,25	0	0	6	0	0
0,70	0,10	0	0	0	6	0	0,90	0,30	0	0	6	0	0
0,70	0,15	0	0	0	6	0	0,90	0,35	0	0	6	0	0
0,70	0,20	0	0	4	2	0	0,90	0,40	0	0	6	0	0
0,70	0,25	0	0	5	1	0							

Tabela 6.10: Quantidade de variações encontradas na primeira janela do FCD *DS03* com a execução do *f-DOPE* dadas as combinações de τ_{match} e τ_{split}.

semanas na Tabela 6.10 e o gráfico presente na Figura 6.5, pode-se observar que valores altos para τ_{match} geralmente indicam o surgimento ($\odot \to B$) de todos os perfis na janela (lembrando que na primeira janela não é possível ter surgimentos). Portanto, não há perfis que se dividam ($A \overset{\subset}{\to} \{B_1, ..., B_k\}$) ou sobrevivam ($A \to B$) de uma janela para outra. Além disso, valores altos de τ_{split} quando τ_{match} é no máximo $0,60$ também indicam o surgimento de todos os perfis em uma nova janela. Por exemplo, para valores de $\tau_{match} = 0,50$ e $\tau_{split} = 0,15$, identificou-se uma sobrevivência e 5 divisões indicando mudanças de conceitos. Alguns perfis podem agregar mais objetos, enquanto outros podem perder tais objetos, porém mantendo as mesmas propriedades ou não. Assim, quando

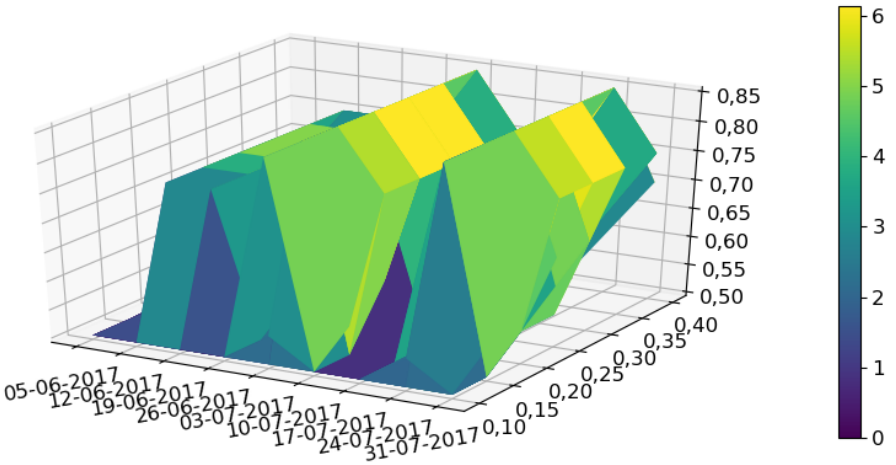


Figura 6.5: Número de perfis que surgem dada a variação dos limiares τ_{match} e τ_{split} ao longo das janelas do FCD *DS03* com a execução do *f-DOPE*.

vários objetos alterarem de perfis, tais alterações podem indicar que as características dos conceitos inicialmente aprendidos também mudaram. Por fim, uma comparação dos resultados aqui descritos, com a aplicação da metodologia da literatura é realizada na Seção 6.5.3.

6.3.5 Fase de Segmentação

Pela investigação das mudanças nos comportamentos dos dispositivos, de acordo com os perfis e as mudanças de conceitos detectadas, diferentes *ciclos comportamentais* foram obtidos. Assim, para cada variação nos valores de τ_{match} e τ_{split} diferentes comportamentos (L, C) podem ocorrer gerando ao final *ciclos comportamentais* distintos. Contudo, é importante lembrar que para o *f-DOPE* é possível a identificação de dois outros comportamentos (O e M). Nesse sentido, alguns dispositivos apresentam o mesmo *ciclo comportamental*, enquanto outros podem ter um *ciclo comportamental* único. *Ciclos* apresentados por poucos dispositivos, por exemplo por 2 dispositivos, representam dispositivos que mudam seu comportamento constantemente. Enquanto que *ciclos* identificados para muitos dispositivos representam dispositivos que possuem um comportamento mais estável e normalmente L. A Tabela 6.11 apresenta o número de *ciclos comportamentais* encontrados dada a variação dos valores dos limiares τ_{match} e τ_{split}, enquanto que a Figura 6.6 apresenta um gráfico com a quantidade de *ciclos comportamentais* (eixo y) de acordo com a combinação de tais limiares (eixo x). Percebe-se que conforme τ_{match} e τ_{split} variam diferentes mudanças de conceitos são identificadas e consequentemente um número diferente de *ciclos comportamentais* são encontrados. A linha azul indica o número de *ciclos comportamentais* identificados com o *f-DOPE*, enquanto a linha vermelha tracejada representa o número de *ciclos comportamentais* obtidos para o *X-Means*. Por fim, a comparação dos resultados aqui com a aplicação de metodologias da literatura é descrita na Seção 6.5.3.

τ_{match}	τ_{split}	*f-DOPE*	*X-Means*	τ_{match}	τ_{split}	*f-DOPE*	*X-Means*
0,50	0,10	1214	1170	0,70	0,30	708	719
0,50	0,15	1519	1176	0,70	0,35	498	504
0,50	0,20	1413	1182	0,70	0,40	351	483
0,50	0,25	1447	1187	0,75	0,10	980	1033
0,50	0,30	1000	1194	0,75	0,15	1076	1089
0,50	0,35	787	1196	0,75	0,20	912	945
0,50	0,40	652	1195	0,75	0,25	741	678
0,55	0,10	1041	1144	0,75	0,30	669	438
0,55	0,15	1429	1154	0,75	0,35	493	346
0,55	0,20	1265	1169	0,75	0,40	351	327
0,55	0,25	1213	1178	0,80	0,10	1051	1050
0,55	0,30	743	1190	0,80	0,15	1071	859
0,55	0,35	552	1167	0,80	0,20	659	758
0,55	0,40	392	1159	0,80	0,25	468	508
0,60	0,10	993	1194	0,80	0,30	468	233
0,60	0,15	1441	1132	0,80	0,35	351	166
0,60	0,20	1086	1182	0,80	0,40	351	156
0,60	0,25	991	1186	0,85	0,10	1244	806
0,60	0,30	713	1141	0,85	0,15	691	408
0,60	0,35	498	1041	0,85	0,20	523	309
0,60	0,40	351	1031	0,85	0,25	398	274
0,65	0,10	993	1043	0,85	0,30	398	174
0,65	0,15	1088	1112	0,85	0,35	351	123
0,65	0,20	1026	1149	0,85	0,40	351	123
0,65	0,25	765	1135	0,90	0,10	1174	34
0,65	0,30	713	997	0,90	0,15	592	206
0,65	0,35	498	883	0,90	0,20	512	187
0,65	0,40	351	875	0,90	0,25	351	187
0,70	0,10	1004	994	0,90	0,30	351	134
0,70	0,15	971	1102	0,90	0,35	351	120
0,70	0,20	917	1053	0,90	0,40	351	120
0,70	0,25	765	978				

Tabela 6.11: Quantidade de *ciclos comportamentais* encontrados com as execuções de *f-DOPE* e *X-Means* dadas as combinações de τ_{match} e τ_{split} no FCD *DS03*.

6.4 Execução da Metodologia Empregada na Literatura

Esta Seção descreve os resultados da realização de experimentos por meio da aplicação de metodologias empregadas na literatura (ver Seção 6.2.2) no FCD *DS03*. É importante ressaltar

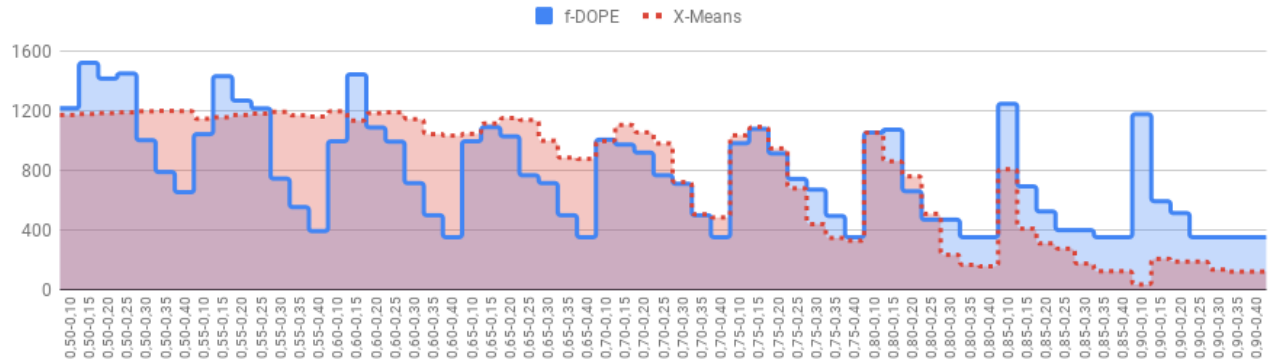


Figura 6.6: Comparação da variação de *ciclos comportamentais* obtidos ao final da execução do *f-DOPE* e *X-Means* para cada combinação dos limiares τ_{match} e τ_{split}.

que os passos de discretização e transformação de atributos em valores categóricos realizados na execução do *f-DOPE* conforme Seção 6.3.1 não são aplicados para a execução dos experimentos aqui descritos. Neste caso, o FCD *DS03* contêm os aplicativos *mais utilizados* (Def. 1) que também são remanescentes (Def 2) sumarizados com os valores contínuos dos atributos (tempo total de uso dos aplicativos) seguindo as opções de configuração descritas na Seção 6.2.2.

6.4.1 Fase de Caracterização

O algoritmo de agrupamento *X-Means* foi aplicado visando identificar perfis de uso de aplicativos a cada janela do FCD *DS03*. Além disso, utilizou-se a distância *Euclidiana* para calcular as distância entre os objetos pertencentes a tal FCD. Pereira e Mendes-Moreira (2016) não relatam as demais configurações para execução do *X-Means* em seus experimentos, somente citando o trabalho introdutório de tal algoritmo. Assim, foi seguida a configuração do *X-Means* utilizada pela ferramenta *Weka* [2] (Witten et al., 2016) com duas modificações. O parâmetro *maxNumClusters*, que define o número máximo de grupos possíveis foi definido igualmente ao valor utilizado no *f-DOPE* ($\sqrt{N}$) (MacQueen et al., 1967)). Outra modificação foi realizada no atributo *maxIterations*, a qual indica quantas execuções do algoritmo são realizadas. Por padrão tal parâmetro é 1 e foi modificado para 2. Dado tais alterações, a execução de tal abordagem resultou na detecção dos perfis que também são apresentado na Tabela 6.9.

Em resumo, todos os dispositivos foram mapeados para um dos k perfis obtidos pelo *X-Means*. Assim como para o *f-DOPE*, cada perfil representa padrões de uso de aplicativo específicos. Contudo, neste caso, os padrões são baseados diretamente no tempo total de uso dos aplicativos utilizados. Por fim, a comparação desses resultados em relação ao *f-DOPE* é descrito nas Seções 6.5.1 e 6.5.2.

[2] https://www.cs.waikato.ac.nz/

6.4.2 Fase de Monitoramento

Depois de obter os perfis por meio da aplicação do algoritmo *X-Means*, é necessário investigar as mudanças de conceitos de uma janela para outra usando a abordagem *enumeration* (Spiliopoulou et al., 2006). Os experimentos foram realizados igualmente ao *f-DOPE* variando τ_{match} de $0,50$ a $0,90$ e τ_{split} de $0,10$ a $0,40$, ambos a cada $0,05$. Os resultados das mudanças de conceito entre a primeira e a segunda janela dado os perfis obtidos com o algoritmo *X-Means* (8 para ambas janelas ζ_w e ζ_{w+1}) são apresentadas na Tabela 6.12.

τ_{match}	τ_{split}	$\mathbf{A} \overset{\subset}{\rightarrow} \mathbf{B}$	$\mathbf{A} \rightarrow \mathbf{B}$	$\mathbf{A} \rightarrow \odot$	$\mathbf{A} \overset{\subset}{\rightarrow} \{\mathbf{B_1}, ..., \mathbf{B_k}\}$	$\odot \rightarrow \mathbf{B}$	τ_{match}	τ_{split}	$\mathbf{A} \overset{\subset}{\rightarrow} \mathbf{B}$	$\mathbf{A} \rightarrow \mathbf{B}$	$\mathbf{A} \rightarrow \odot$	$\mathbf{A} \overset{\subset}{\rightarrow} \{\mathbf{B_1}, ..., \mathbf{B_k}\}$	$\odot \rightarrow \mathbf{B}$
0,50	0,10	2	5	0	1	0	0,70	0,30	0	2	6	0	0
0,50	0,15	2	5	0	1	0	0,70	0,35	0	2	6	0	0
0,50	0,20	2	5	0	1	0	0,70	0,40	0	2	6	0	0
0,50	0,25	2	5	0	1	0	0,75	0,10	0	2	0	6	0
0,50	0,30	2	5	1	0	0	0,75	0,15	0	2	1	5	0
0,50	0,35	2	5	1	0	0	0,75	0,20	0	2	5	1	0
0,50	0,40	2	5	1	0	0	0,75	0,25	0	2	5	1	0
0,55	0,10	2	5	0	1	0	0,75	0,30	0	2	6	0	0
0,55	0,15	2	5	0	1	0	0,75	0,35	0	2	6	0	0
0,55	0,20	2	5	0	1	0	0,75	0,40	0	2	6	0	0
0,55	0,25	2	5	0	1	0	0,80	0,10	0	0	3	5	0
0,55	0,30	2	5	1	0	0	0,80	0,15	0	0	5	3	0
0,55	0,35	2	5	1	0	0	0,80	0,20	0	0	7	1	0
0,55	0,40	2	5	1	0	0	0,80	0,25	0	0	7	1	0
0,60	0,10	0	6	0	2	0	0,80	0,30	0	0	8	0	0
0,60	0,15	0	6	0	2	0	0,80	0,35	0	0	8	0	0
0,60	0,20	0	6	1	1	0	0,80	0,40	0	0	8	0	0
0,60	0,25	0	6	1	1	0	0,85	0,10	0	0	4	4	0
0,60	0,30	0	6	2	0	0	0,85	0,15	0	0	6	2	0
0,60	0,35	0	6	2	0	0	0,85	0,20	0	0	7	1	0
0,60	0,40	0	6	2	0	0	0,85	0,25	0	0	7	1	0
0,65	0,10	0	4	0	4	0	0,85	0,30	0	0	8	0	0
0,65	0,15	0	4	0	4	0	0,85	0,35	0	0	8	0	0
0,65	0,20	0	4	3	1	0	0,85	0,40	0	0	8	0	0
0,65	0,25	0	4	3	1	0	0,90	0,10	0	0	6	2	0
0,65	0,30	0	4	4	0	0	0,90	0,15	0	0	7	1	0
0,65	0,35	0	4	4	0	0	0,90	0,20	0	0	7	1	0
0,65	0,40	0	4	4	0	0	0,90	0,25	0	0	7	1	0
0,70	0,10	0	2	0	6	0	0,90	0,30	0	0	8	0	0
0,70	0,15	0	2	0	6	0	0,90	0,35	0	0	8	0	0
0,70	0,20	0	2	5	1	0	0,90	0,40	0	0	8	0	0
0,70	0,25	0	2	5	1	0							

Tabela 6.12: Quantidade de variações encontradas na primeira janela do FCD *DS03* com a execução do *X-Means* dadas as combinações de τ_{match} e τ_{split}.

Dado os resultados obtidos e o gráfico presente na Figura 6.7, o qual apresenta o a quantidade de perfis que surgem ao longo do tempo, observa-se que conforme o valor de τ_{match} aumenta o número de desaparecimento ($A \rightarrow \odot$) de perfis também aumenta. Além disso, a sobrevivência ($A \rightarrow B$) de perfis só é detectada com valores mais baixos de para tal limiar. Valores muito altos de τ_{split} fazem com que a divisões ($A \overset{\subset}{\rightarrow} \{B_1, ..., B_k\}$) não seja detectadas. Por exemplo, os limiares $\tau_{match} = 0,65$ e $\tau_{split} = 0,10$ indicam 6 sobrevivências e 2 divisões (lembrando que na primeira janela não é possível ter surgimentos). Por fim, a avaliação dos resultados aqui descritos em comparação com o *f-DOPE* é descrita na Seção 6.5.3.

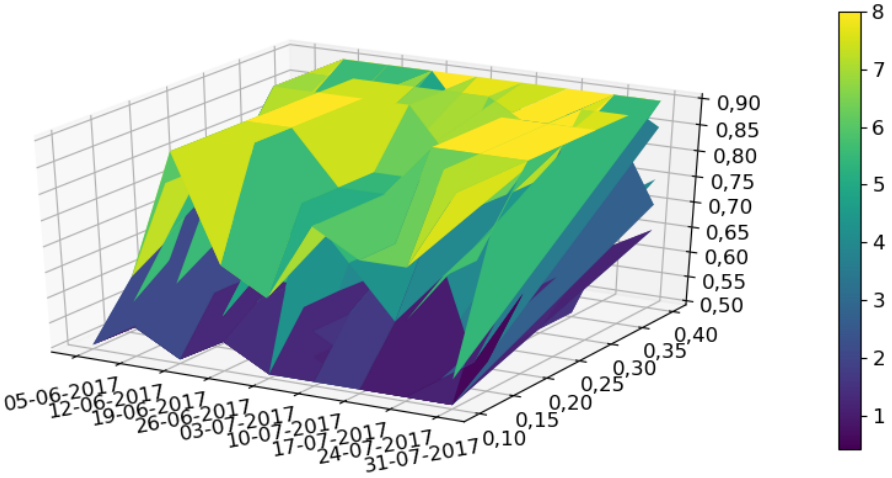


Figura 6.7: Número de perfis que surgem dada a variação dos limiares τ_{match} e τ_{split} na primeira janela do FCD *DS03* com execução do *X-Means*.

6.4.3 Fase de Segmentação

Com a identificação de perfis com o algoritmo *X-Means* também foram encontrados diferentes *ciclos comportamentais*. Da mesma forma, para cada variação nos valores de τ_{match} e τ_{split} diferentes ações (L, C) poderiam ocorrer. Contudo, é importante lembrar que para a abordagem da literatura é somente possível a identificação de uma outra ação M, pois com essa abordagem a ação O (*outliers*) não é detectada. Aqui, alguns dispositivos também apresentaram o mesmo *ciclo comportamental* e outros apresentaram um único *ciclo comportamental*. Além disso, a quantidade de dispositivos com o mesmo *ciclo comportamental* segue o padrão que foi apresentado pelo *f-DOPE*. *Ciclos comportamentais* com poucos dispositivos representando muitas mudanças de comportamento e *ciclos comportamentais* com muitos dispositivos indicando comportamentos L. A Tabela 6.11 também apresenta a quantidade de *ciclos comportamentais* encontrados dada a variação nos valores de τ_{match} e τ_{split} para esta abordagem, enquanto que a Figura 6.7 também mostra um gráfico que apresenta a variação na quantidade de *ciclos comportamentais* dado as variações de tais limiares para o *X-Means*. Por fim, os resultados aqui obtidos são comparados com os resultados da aplicação do *f-DOPE* na Seção 6.5.3.

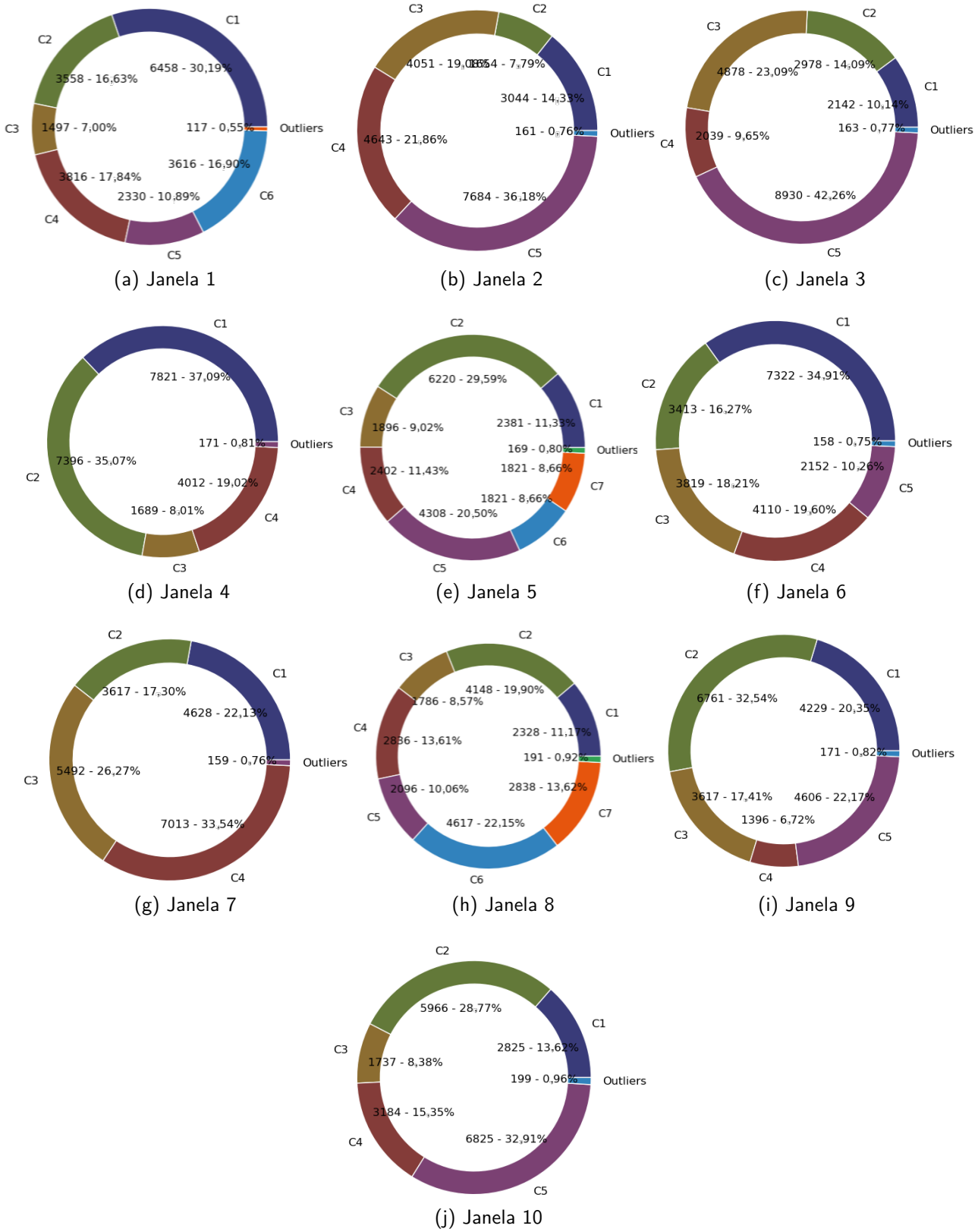
6.5 Avaliação dos Resultados

6.5.1 Distribuição dos perfis

Em relação aos resultados obtidos com os experimentos realizados nas Seções 6.3.1, 6.3.2 e 6.3.3 com a execução do *f-DOPE*, pode-se observar que foram encontrados em média 5 perfis e 166 *outliers*, que representam em média 0,78% dos dispositivos (ver Tabela 6.9), podendo-se dizer que os *outliers* são um grupo identificado pela ausência de suporte aos padrões de uso de aplicativos encontrados. Também na Tabela 6.9, coluna *X-Means*, é possível perceber que a abordagem da literatura sempre encontrou a mesma quantidade de perfis (8) (ver Seção 6.4.1). Deste modo, verificou-se também a distribuição dos dispositivos nos perfis obtidos em ambas as execuções. Para o *f-DOPE* observa-se na Figura 6.8 que cada perfil possui um número diferente de objetos. Além disso, quase sempre é verificado um perfil predominante, mas com menos de 50% dos objetos, e outros perfis menores. Por outro lado, com o a execução do *X-Means* frequentemente são encontrados grupos muito pequenos, as vezes com menos de 2% dos objetos (ver Figura 6.9 (a-d, f-j). Em comparação, tais perfis tem quase a mesma proporção dos *outliers* encontrados pelo *f-DOPE*.

6.5.2 Variação dos perfis ao longo do FCD

Neste experimento, em cada janela menos dispositivos do FCD *DS03* são analisados e as quantidades de aplicativos bem como de eventos também mudam ao longo do tempo. Assim, é esperado deste cenário que o número de perfis possa ser diferentes em algumas janelas. Em relação aos resultados obtidos com os experimentos realizados nas Seções 6.3.1, 6.3.2 e 6.3.3, onde foi aplicado o *f-DOPE*, foram encontrados diferentes números de perfis ao longo de algumas das janelas de eventos (ver Tabela 6.9, coluna *f-DOPE*) o que representa o surgimento de novos perfis ou a evolução de perfis conhecidos. Por outro lado, em relação ao experimentos realizados na Seção 6.4.1, com a execução da metodologia da literatura, foi sempre obtido o mesmo número de perfis (8) (ver Tabela 6.9, coluna *X-Means*), não existindo alterações. Como existe uma grande variação na distribuição dos dados a cada janela, esperava-se que isso fosse refletido em um número diferente de perfis ao menos em algumas semanas, o que não ocorreu com o *X-Means*. Desta forma, pode-se dizer que a realização da discretização por *IP* satisfaz os requisitos necessários para a tarefa de discretização não supervisionada no cenário abordado. Além disso, o uso de conjunto de itens ajuda com problema de regras redundantes ou muito similares. Mais ainda, com base em tais conjuntos, os quais permitem a identificação de *outliers* e a definição de similaridade entre dispositivos, é possível a identificação de perfis de uso, onde todos os achados refletem a mudança na distribuição do FCD.



(a) Janela 1 (b) Janela 2 (c) Janela 3

(d) Janela 4 (e) Janela 5 (f) Janela 6

(g) Janela 7 (h) Janela 8 (i) Janela 9

(j) Janela 10

Figura 6.8: Distribuição dos dispositivos nos perfis identificados em cada janela (a-j) do FCD *DS03* pelo *framework* proposto.

Dadas as evoluções que ocorrem com os perfis identificados, conforme aprestado nas Seções 6.3.4 e 6.4.2, é possível verificar que a aplicação do *f-DOPE*, bem como do *X-Means*, com

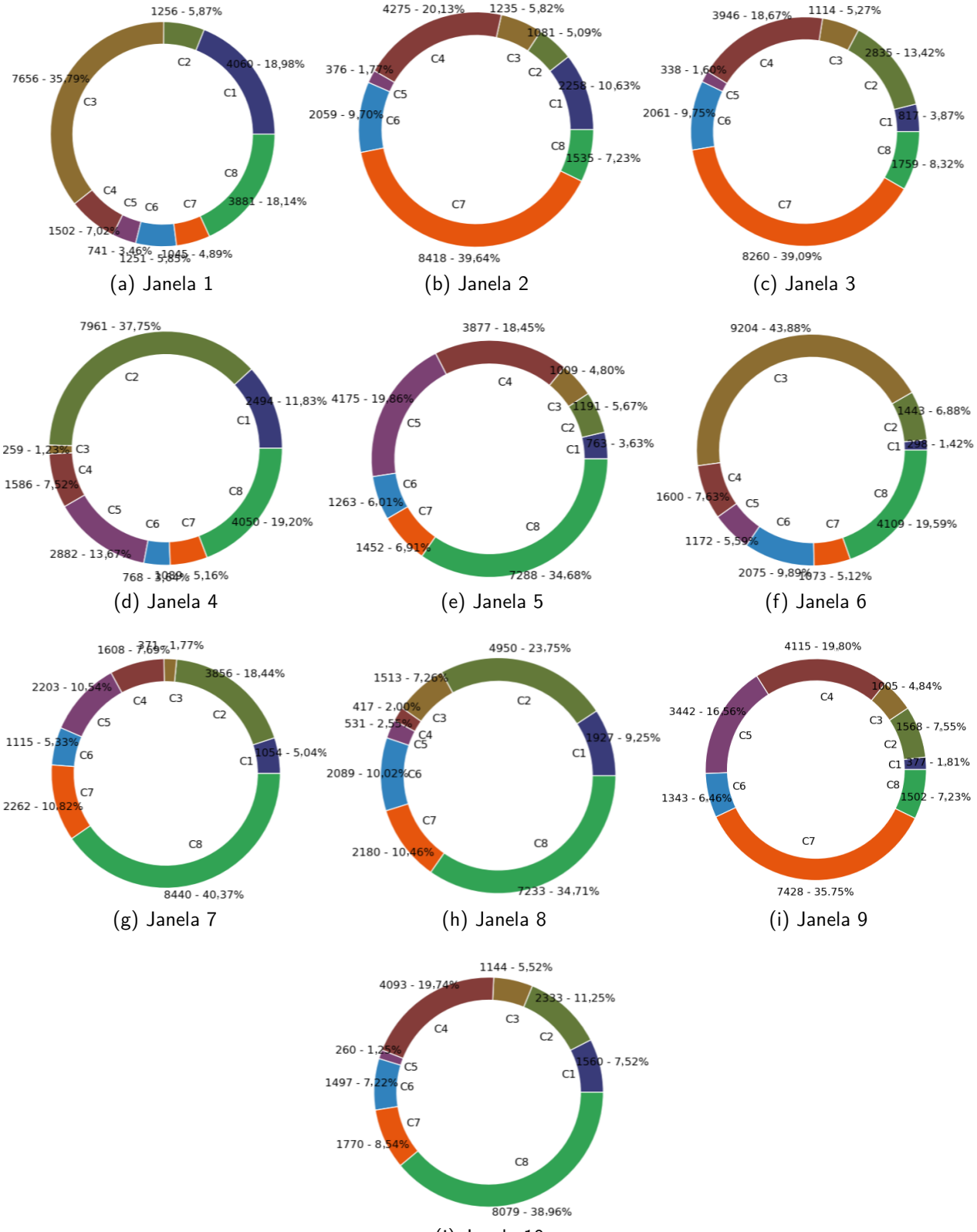


Figura 6.9: Distribuição dos dispositivos nos perfis identificados em cada janela (a-j) do FCD *DS03* com *X-Means*.

a abordagem de *enumeração* encontra diferentes variações de perfis dados os valores de τ_{match} e de τ_{split}. Contudo, não fica evidente uma melhor configuração para ambas as execuções. Assim, na

Seção 6.5.3 são discutidos os *ciclos comportamentais* encontrado para ambas abordagens com a variação de tais limiares visando o melhor comparativo.

6.5.3 Tipos de comportamentos

Dados o monitoramento dos perfis obtidos pela aplicação do *f-DOPE*, descrita na Seção 6.3.4, e a geração dos *ciclos comportamentais*, detalhada na Seção 6.3.5, fica evidente que a variação do τ_{split} provoca maiores mudanças na obtenção dos *ciclos comportamentais* (ver Tabela 6.11). Um número de *ciclos comportamentais* pequeno indica em que os dispositivos apresentam muitos comportamentos C. Isso mostra que não existe muita sobrevivência dos perfis ao longo do tempo e que eles desaparecem ou se dividem na maioria dos casos. Assim, muitos dispositivos apresentam comportamentos C e consequentemente o mesmo *ciclo comportamental*. Por outro lado, um número mais elevado de *ciclos comportamentais* aponta para dispositivos apresentando mais comportamentos L. Neste caso, existem sobrevivências de perfis ao longo do tempo e por isso tais dispositivos se mantêm em seus perfis, o que é esperado no cenário abordado conforme descrito na caracterização do problema no Capítulo 1, Seção 1.3. Assim, comportamentos C não são tão frequentes e quando acontecem possuem ocorrência em janelas distintas e para um número pequeno de dispositivos.

Com ralação ao monitoramento dos perfis identificado com a execução do (X-Means), relatada na Seção 6.4.2, e a concepção dos *ciclos comportamentais*, relatada na Seção 6.4.3, é possível perceber que a variação de τ_{split} não afeta a quantidade de *ciclos comportamentais* obtidos da mesma maneira que afetou na execução do *f-DOPE*. Esta verificação pode estar associada ao número de perfis identificados, uma vez que para o *X-Means* o número de perfis foi sempre igual, enquanto que para o *f-DOPE* este número mudou ao longo das janelas. Contudo, valores elevados de τ_{match} fazem com que o número de *ciclos comportamentais* tenha uma queda indicando a existência de muitos comportamentos C.

Por fim, é possível verificar que a combinação de valores de $\tau_{match} = 0,50$ e com $\tau_{split} = [0,10; 0,25]$ apresentam resultados mais similares para a execução do *f-DOPE* (ver Figura 6.6). Neste intervalo de valores a variação do número de *ciclos comportamentais* é menor do que no restante das possibilidades mantendo o padrão de aumento e diminuição de tal número. Da mesma forma, não é possível afirmar os valores de τ_{match} e de τ_{split} que apresentam resultados melhores para a execução com *X-Means*. Estima-se que valores mais elevados, como $\tau_{match} = [0,60; 0,75]$ e $\tau_{split} < 0,25$ possam ser melhores dadas as variações ocorridas com esse limiares conforme apresenta a Figura 6.6. No intervalo de valores mencionado se inicia uma variação do número de *ciclos comportamentais* que tende a seguir o mesmo padrão que ocorre em todas as possibilidades com o *f-DOPE*, enquanto que valores mais elevados de τ_{match} mostram que o número de *ciclos comportamentais* varia demasiadamente. Em resumo, a variação destes limiares depende da detecção dos perfis e do comportamento dos objetos ao longo do tempo, não sendo possível afirmar a melhor combinação para tais limiares somente com a verificação do número de *ciclos comportamentais* obtidos. Assim,

tais valores devem ser abordados na exploração da saída do *f-DOPE* em comparação com a saída do *X-Means* visando a predição de comportamentos, a qual é apresentada na Seção 6.6.

6.6 Explorando a saída do *f-DOPE* ao comparar com a saída do *X-Means*

Nesta Seção, é discutido a predição de comportamentos para décima janela de eventos do FCD *DS03*. O objetivo é comparar a execução do *framework f-DOPE* em relação ao *X-Means* visando apontar qual mantém uma identificação de perfis e comportamentos mais harmônico ao longo do tempo. Para este fim, após a identificação de perfis e comportamentos para nove janelas para ambas execuções, é aplicado um algoritmo classificador com o objetivo de prever o comportamento dos dispositivos da décima janela de eventos. Neste caso, busca-se que a identificação de perfis e comportamentos pelo *f-DOPE* consiga melhores resultados.

Para este experimento, utilizou-se as variações de τ_{match} e τ_{split}, onde para cada combinação de valores, foram gerados diferentes *ciclos comportamentais* considerando todos os 21.392 dispositivos analisados para ambas as abordagens. Um exemplo da saída de cada execução é dado na Tabela 6.13, onde são mostrados três *ciclos comportamentais* de exemplo. Na primeira janela não há identificação de comportamento pois é necessário uma janela anterior para a execução da Detecção de Novidade. Com a aplicação do *f-DOPE* os *ciclos comportamentais* obtidos são compostos dos comportamentos (*L*, *C*, *M* e *O*) da segunda até a penúltima janela analisada enquanto que, na abordagem da literatura, não existem os *outliers* (*O*).

O classificador utiliza como referência um rótulo normalmente dado por um supervisor (Tan et al., 2006). Neste caso, o rótulo utilizado é o comportamento identificado na última janela de eventos (Jan. 10). Assim, procurou-se trabalhar com apenas 2 classes pois a aplicação do *f-DOPE* e da abordagem da literatura geram classes diferentes. Desta forma, a classe majoritária *L* continuou existindo, enquanto que as demais (*C*, *M* ou *O*) foram transformadas em ¬L como mostra o exemplo na terceira linha da Tabela 6.13.

Jan. 2	Jan. 3	Jan. 4	Jan. 5	Jan. 6	Jan. 7	Jan. 8	Jan. 9	Jan. 10
L	L	L	L	L	L	L	L	L
L	L	L	M	O	L	L	L	L
L	C	L	M	L	L	L	M	¬L

Tabela 6.13: Exemplos de *ciclos comportamentais* utilizados no avaliação da predição de comportamentos.

Os conjuntos de *ciclos comportamentais*, para as 69 combinações de limares, foram um a um utilizados como conjunto de dados para o algoritmo J48, o qual é uma árvore de decisão. Este algoritmo foi utilizado na ferramenta Weka (Witten et al., 2016). As únicas alterações da configuração padrão deste algoritmo em tal ferramenta foram as mudanças dos atributos *minNumObj*

e *unpruned*. O atributo *minNumObj* é utilizado para definir o mínimo de instâncias por folha durante a geração da árvore de decisão. Tal atributo foi modificado de 2 para 1. O atributo *unpruned* por padrão é falso determinando a diminuição da árvore por meio da remoção de galhos sem afetar muito o modelo criado. Tal atributo foi modificado para verdadeiro para que tal poda da árvore não fosse realizada.

Para a avaliação dos modelos foi aplicada um validação cruzada (Tan et al., 2006). Neste tipo de validação o conjunto de dados é dividido em partes (*folds*) nas mesmas proporções do conjunto de dados e cada objeto participa o mesmo número de vezes da fase de treinamento e uma vez do teste (Han et al., 2011). Neste caso estipulou-se 10 partes para a validação cruzada (Witten et al., 2016). As métricas de avaliação utilizadas para avaliar os modelos criados são baseadas em matrizes de confusão pois tais matrizes ajudam na análise de classificadores (Han et al., 2011; Witten et al., 2016). Em uma matriz de confusão mxm, m são as classes buscadas. Neste caso, em um exemplo com duas classes, uma positiva e outra negativa, alguns dos termos básicos são, *verdadeiros positivos* (*vp*) que são os objetos positivos classificados corretamente, *verdadeiros negativos* (*vn*) que são os objetos negativos classificados corretamente, *falsos positivos* (*fp*) que são os objetos negativos classificados como positivos e *falsos negativos* (*fn*) que são os objetos positivos classificados como negativos (Han et al., 2011). Baseado em tais possibilidades algumas métricas podem ser calculadas. As métricas utilizadas são *Acurácia* (Equação 6.2), que mede quão frequente um classificador está correto, *Sensibilidade* (Equação 6.3) (também conhecida como *Taxa de Verdadeiros Positivos* ou *Recall*), que mede quão frequente um classificados acerta uma determinada classe considerada positiva, *Taxa de Falsos Positivos* - (*FPR*) (Equação 6.4), que mede quão frequente um classificador diz que objetos considerados negativos são classificados como positivos, *Precisão* (Equação 6.5), que mede quão frequentemente um classificador acerta uma classe positiva e *F-Measure* (também conhecida como *F-score* e *F1 score*) (Equação 6.6), a qual é uma média harmônica entre *Sensibilidade* e *Precisão*.

$$\text{Acurácia} = \frac{vp + vn}{vp + vn + fp + fn} \tag{6.2}$$

$$Sensibilidade = \frac{vp}{vp + fn} \tag{6.3}$$

$$FPR = \frac{fp}{vn + fp} \tag{6.4}$$

$$\text{Precisão} = \frac{vp}{vp + fp} \tag{6.5}$$

$$F\text{-}Measure = 2 \times \frac{Precisão \times Sensibilidade}{Precisão + Sensibilidade} \tag{6.6}$$

Dadas as configurações acima mencionadas e as medidas descritas foram realizadas as predições. Por questões de espaço, todos os 69 resultados obtidos com a aplicação do *f-DOPE* são apresentados na Tabela G.1 que está no Apêndice G. Pelo mesmo motivo, todos os 69 resultados com a aplicação do *X-Means* se encontram na Tabela H.1 que está no Apêndice H. Uma vez que clientes tendem a manter um padrão no uso de aplicativos (L) e a alteração ($\neg L$) no uso de tais aplicativos não é frequente (ver contextualização realizada no Capítulo 1, Seção 1.3), existe um desbalanceamento das classes, onde são esperados mais comportamentos L. Neste caso, para algumas combinações de limiares, tanto com a execução do *f-DOPE* como na aplicação do *X-Means*, não foram gerados comportamentos possíveis de serem classificados corretamente. Em alguns casos foram geradas muitas mudanças de conceitos, resultando em muitos ou somente comportamentos $\neg L$ na última semana. Assim, foram selecionados para serem analisados somente os resultados que possuem, na última semana, mais comportamentos L do que $\neg L$, onde a distribuição da classe $\neg L \leq 40\%$, seguindo o que se espera no cenário abordado. As Tabelas 6.14 e 6.15 apresentam tais resultados, respectivamente obtidos pelas execuções do *f-DOPE* e *X-Means* indicando na primeira coluna os limiares, na segunda coluna o percentual de distribuição das classes, na terceira coluna a medida de *Sensibilidade*, na quarta coluna a medida *FPR*, na quinta coluna a medida *Precisão* e na sexta coluna a medida *F-Measure*, todas separadas pelas classes possíveis ($\neg L$ e L), enquanto que da sétima à décima coluna estão, respectivamente, *Acurácia*, e as medidas ponderadas de *Sensibilidade*, *FPR*, *Precisão* e *F-Measure*.

Limiares		*Distribuição*		*Sensibilidade*		*FPR*		*Precisão*		*F-Measure*		*Acurácia*	*Sensibilidade*	*FPR*	*Precisão*	*F-Measure*
τ_{match}	τ_{split}	$\neg L$	L	$\neg L$	L	$\neg L$	L	$\neg L$	L	$\neg L$	L		*Ponderada*	*Ponderada*	*Ponderada*	*Ponderada*
0,50	**0,10**	**20,67%**	**79,33%**	**0,191**	**0,994**	**0,006**	**0,809**	**0,896**	**0,825**	**0,315**	**0,902**	**0,828**	**0,828**	**0,643**	**0,840**	**0,781**
0,50	0,15	28,73%	71,27%	0,148	0,988	0,012	0,852	0,837	0,742	0,251	0,848	0,747	0,747	0,611	0,770	0,677
0,55	**0,10**	**11,68%**	**88,32%**	**0,333**	**0,996**	**0,004**	**0,667**	**0,915**	**0,919**	**0,489**	**0,956**	**0,919**	**0,919**	**0,589**	**0,918**	**0,901**
0,55	0,15	39,58%	60,42%	0,194	0,930	0,070	0,806	0,645	0,638	0,298	0,757	0,639	0,639	0,515	0,641	0,575
0,60	**0,10**	**11,68%**	**88,32%**	**0,333**	**0,996**	**0,004**	**0,667**	**0,920**	**0,919**	**0,489**	**0,956**	**0,919**	**0,919**	**0,590**	**0,919**	**0,901**
0,60	0,15	39,58%	60,42%	0,253	0,900	0,100	0,747	0,624	0,648	0,360	0,753	0,644	0,644	0,491	0,638	0,598
0,65	**0,10**	**11,68%**	**88,32%**	**0,333**	**0,996**	**0,004**	**0,667**	**0,920**	**0,919**	**0,489**	**0,956**	**0,919**	**0,919**	**0,590**	**0,919**	**0,901**
0,65	0,15	39,58%	60,42%	0,187	0,942	0,058	0,813	0,678	0,639	0,239	0,761	0,643	0,643	0,514	0,654	0,576
0,70	**0,10**	**11,68%**	**88,32%**	**0,331**	**0,996**	**0,004**	**0,669**	**0,917**	**0,918**	**0,487**	**0,956**	**0,918**	**0,918**	**0,591**	**0,918**	**0,901**
0,70	0,15	39,58%	60,42%	0,181	0,944	0,056	0,819	0,677	0,637	0,285	0,761	0,642	0,642	0,517	0,653	0,573
0,75	**0,10**	**11,68%**	**88,32%**	**0,333**	**0,995**	**0,005**	**0,667**	**0,903**	**0,919**	**0,487**	**0,955**	**0,918**	**0,918**	**0,590**	**0,917**	**0,901**
0,75	0,15	39,58%	60,42%	0,163	0,950	0,050	0,837	0,682	0,643	0,263	0,761	0,639	0,639	0,525	0,653	0,654
0,80	**0,10**	**11,68%**	**88,32%**	**0,333**	**0,995**	**0,005**	**0,667**	**0,903**	**0,919**	**0,486**	**0,955**	**0,918**	**0,918**	**0,590**	**0,917**	**0,901**
0,85	**0,10**	**11,68%**	**88,32%**	**0,331**	**0,996**	**0,004**	**0,669**	**0,913**	**0,918**	**0,486**	**0,956**	**0,918**	**0,918**	**0,591**	**0,918**	**0,901**
0,90	**0,10**	**11,68%**	**88,32%**	**0,332**	**0,995**	**0,005**	**0,668**	**0,898**	**0,918**	**0,485**	**0,955**	**0,918**	**0,918**	**0,591**	**0,916**	**0,900**

Tabela 6.14: Resultados das medidas de avaliação após execução do classificador nos *ciclos comportamentais* obtidos pelo *f-DOPE* dadas as combinações de τ_{match} e τ_{split}. Em negrito os melhores resultados de acordo com os critérios adotados.

Uma vez que a medida de *Acurácia* pode apresentar bons resultados, os quais na verdade podem ser uma simples casualidade dado o desbalanceamento de classes, deve-se analisar todas as medidas indicadas. Tal casualidade pode ocorrer, quando existem poucas instâncias de uma classe, por exemplo, da classe negativa, e o modelo classifica todas as instâncias como da classe positiva, a qual é a majoritária. Assim, em relação aos resultados selecionados para ambas execuções,

Limiares		Distribuição		Sensibilidade		FPR		Precisão		F-Measure		Acurácia	Sensibilidade	FPR	Precisão	F-Measure
τ_{match}	τ_{split}	$\neg L$	L	$\neg L$	L	$\neg L$	L	$\neg L$	L	$\neg L$	L		Ponderada	Ponderada	Ponderada	Ponderada
0,50	**0,10**	**29,29%**	**70,71%**	**0,190**	**0,957**	**0,043**	**0,810**	**0,644**	**0,740**	**0,293**	**0,835**	**0,732**	**0,732**	**0,586**	**0,712**	**0,676**
0,50	0,15	29,29%	70,71%	0,183	0,957	0,043	0,817	0,639	0,739	0,285	0,834	0,730	0,730	0,590	0,710	0,673
0,50	0,20	29,29%	70,71%	0,165	0,968	0,032	0,835	0,680	0,737	0,266	0,837	0,733	0,733	0,600	0,720	0,669
0,50	0,25	32,89%	67,11%	0,251	0,911	0,089	0,749	0,581	0,713	0,351	0,800	0,694	0,694	0,532	0,669	0,652
0,55	**0,10**	**29,29%**	**70,71%**	**0,190**	**0,957**	**0,043**	**0,810**	**0,648**	**0,741**	**0,294**	**0,835**	**0,733**	**0,733**	**0,585**	**0,714**	**0,677**
0,55	0,15	29,29%	70,71%	0,182	0,957	0,043	0,818	0,636	0,738	0,283	0,834	0,730	0,730	0,591	0,708	0,672
0,55	0,20	29,29%	70,71%	0,160	0,970	0,030	0,840	0,688	0,736	0,260	0,837	0,733	0,733	0,603	0,722	0,668
0,55	0,25	32,89%	67,11%	0,230	0,921	0,079	0,770	0,587	0,709	0,330	0,801	0,693	0,693	0,543	0,669	0,646
0,60	**0,10**	**29,29%**	**70,71%**	**0,194**	**0,956**	**0,044**	**0,806**	**0,645**	**0,741**	**0,299**	**0,835**	**0,733**	**0,733**	**0,583**	**0,713**	**0,678**
0,60	0,15	29,29%	70,71%	0,177	0,956	0,044	0,823	0,625	0,737	0,276	0,832	0,728	0,728	0,595	0,704	0,670
0,60	0,20	29,29%	70,71%	0,168	0,961	0,039	0,832	0,641	0,736	0,266	0,834	0,729	0,729	0,600	0,708	0,667
0,60	0,25	32,89%	67,11%	0,246	0,917	0,083	0,754	0,593	0,713	0,348	0,802	0,696	0,696	0,533	0,673	0,653
0,65	**0,10**	**23,53%**	**76,47%**	**0,148**	**0,986**	**0,014**	**0,852**	**0,762**	**0,786**	**0,248**	**0,875**	**0,785**	**0,785**	**0,652**	**0,780**	**0,725**
0,65	0,15	23,53%	76,47%	0,144	0,987	0,013	0,856	0,777	0,786	0,244	0,875	0,785	0,785	0,654	0,784	0,724
0,70	**0,10**	**22,77%**	**77,23%**	**0,156**	**0,986**	**0,014**	**0,844**	**0,763**	**0,798**	**0,259**	**0,882**	**0,797**	**0,797**	**0,655**	**0,790**	**0,740**
0,70	0,15	22,77%	77,23%	0,144	0,991	0,009	0,856	0,820	0,797	0,245	0,833	0,798	0,798	0,663	0,802	0,738
0,75	0,10	22,77%	77,23%	0,153	0,986	0,014	0,847	0,759	0,798	0,255	0,882	0,796	0,796	0,657	0,789	0,739
0,75	**0,15**	**22,77%**	**77,23%**	**0,150**	**0,990**	**0,010**	**0,850**	**0,821**	**0,798**	**0,254**	**0,884**	**0,799**	**0,799**	**0,659**	**0,803**	**0,740**

Tabela 6.15: Resultados das medidas de avaliação após execução do classificador nos *ciclos comportamentais* obtidos pelo *X-Means* dadas as combinações de τ_{match} e τ_{split}. Em negrito os melhores resultados de acordo com os critérios adotados.

constatou-se que, quanto maior o valor de τ_{match} mais comportamentos $\neg L$ são obtidos na última semana. Consequentemente, isso também ocorre nas demais semanas, mostrando que valores muito altos para este limiar fazem com que todos conceitos mudem ao longo do tempo e que os dispositivos não continuem em seus perfis, o que não é esperado pelo cenário abordado.

Em continuidade, para cada uma das execuções foram apontados os melhores resultados (em negrito), dentre os selecionados anteriormente (ver Tabelas 6.14 e 6.15), visando a exploração de suas matrizes de confusão, de acordo com os seguintes critérios:

- O melhor resultado para cada possível valor de τ_{match};
  - O resultado com valor mais alto da medida *F-Measure*;
  - Existindo empate, o resultado com valor mais alto da medida *Acurácia*;

Dados os critérios acima, selecionou-se os melhores resultados de classificação na saída do *f-DOPE*, que são $\tau_{match} = [0,55; 0,60; 0,65; 0,70; 0,75; 0,80; 0,85; 0,90]$, todos em combinação $\tau_{split} = 0,10$. É possível observar que para o *f-DOPE*, os melhores resultados são muito similares (ver Tabela 6.14). Por exemplo, para a combinação de $\tau_{match} = 0,60$ e $\tau_{split} = 0,10$ o classificador tem uma uma *Acurácia* de 0,919 e uma *Sensibilidade ponderada* de 0,919, sendo 0,333 para a classe $\neg L$ e 0,996 para a classe L. Além disso, com estes limiares a *Precisão ponderada* é de 0,919 sendo 0,920 para a classe $\neg L$ e 0,919 para a classe L. Para o *FPR ponderado* 0,590, onde a classe $\neg L$ apresenta 0,004 e a classe L 0,667. Por fim, a *F-Measure ponderada* fica em 0,919, sendo 0,489 para a classe $\neg L$ e 0,956 para a classe L. Os quatro melhores resultados seguindo os critérios definidos, onde limiares mais baixos foram priorizados por serem aqueles menos afetados pelo desbalanceamento

de classes, tiveram suas matrizes de confusão (ver Figura 6.16) analisadas. Como nas métricas analisadas as matrizes de confusão de tais resultados também são semelhantes. Em resumo, são classificadas poucas instâncias como *fn* e algumas como *fp*, os quais são os erros do classificador. Por outro lado, muitas instâncias são classificadas como *vn* e *vp* indicando os acertos do classificador. Assim, é realizada uma previsão correta de instâncias da classe majoritária, bem como da classe rara.

Real \ Predito	$\neg L$	L
$\neg L$	vn = 833	fp = 1.666
L	fn = 77	vp = 18.816

(a) Matriz de Confusão $\tau_{match} = 0,55$ e $\tau_{split} = 0,10$

Real \ Predito	$\neg L$	L
$\neg L$	vn = 832	fp = 1.667
L	fn = 72	vp = 18.821

(b) Matriz de Confusão $\tau_{match} = 0,60$ e $\tau_{split} = 0,10$

Real \ Predito	$\neg L$	L
$\neg L$	vn = 832	fp = 1.667
L	fn = 72	vp = 18.821

(c) Matriz de Confusão $\tau_{match} = 0,65$ e $\tau_{split} = 0,10$

Real \ Predito	$\neg L$	L
$\neg L$	vn = 828	fp = 1.671
L	fn = 75	vp = 18.818

(d) Matriz de Confusão $\tau_{match} = 0,70$ e $\tau_{split} = 0,10$

Tabela 6.16: Matrizes de confusão para os quatro melhores resultados de predição de comportamento com a aplicação do *f-DOPE*.

No mesmo sentido, para a classificação na saída da metodologia da literatura os melhores resultados (ver Tabela 6.15) são $\tau_{match} = [0,50;\ 0,55;\ 0,60;\ 0,65;\ 0,70;\ 0.75]$, quase todos em combinação $\tau_{split} = 0,10$, exceto $\tau_{match} = 0,75$ que apresenta melhor resultado com $\tau_{split} = 0,15$. Como pode-se observar, valores de τ_{split} maiores que $0,20$ indicam valores muito altos para a medida de *Sensibilidade* apontado para a existência de mais comportamentos $\neg L$ encontrados, não indo de encontro com cenário abordado. Os quatro melhores resultados, seguindo os critérios definidos, tiveram suas matrizes de confusão (ver Figura 6.17) analisadas. O número de instâncias classificadas como *fn*, assim como classificadas como *fp* é maior se comparado ao resultados das matrizes na Tabela 6.16. Mesmo com muitas instâncias sendo classificadas como *vn* e *vp* o número de *vp* é menor comparado aos resultados das matrizes na Tabela 6.16. Mesmo assim, pode-se dizer que é realizada uma previsão correta de instâncias da classe majoritária, bem como da classe rara.

Dentre os melhores resultados para abordagem da literatura percebe-se que as diferenças também são pequenas. Por exemplo, para o resultado de $\tau_{match} = 0,75$ e $\tau_{split} = 0,10$, o classificador tem uma *Acurácia* de 0,799, uma *Sensibilidade ponderada* de 0,796, sendo 0,153 para a classe $\neg L$ e 0,986 para a classe L. Além disso, com estes limiares a *Precisão ponderada* é de 0,789 sendo 0,759 para a classe $\neg L$ e 0,798 para a classe L. Para o *FPR ponderada* 0,657, a classe $\neg L$ apresenta 0,014

Real \ Predito	$\neg L$	L
$\neg L$	vn = 1.217	fp = 5.049
L	fn = 669	vp = 14.457

(a) Matriz de Confusão $\tau_{match} = 0,60$ e $\tau_{split} = 0,10$

Real \ Predito	$\neg L$	L
$\neg L$	vn = 756	fp = 4.360
L	fn = 236	vp = 16.040

(b) Matriz de Confusão $\tau_{match} = 0,65$ e $\tau_{split} = 0,10$

Real \ Predito	$\neg L$	L
$\neg L$	vn = 759	fp = 4.111
L	fn = 236	vp = 16.286

(c) Matriz de Confusão $\tau_{match} = 0,70$ e $\tau_{split} = 0,10$

Real \ Predito	$\neg L$	L
$\neg L$	vn = 731	fp = 4.139
L	fn = 159	vp = 16.363

(d) Matriz de Confusão $\tau_{match} = 0,75$ e $\tau_{split} = 0,15$

Tabela 6.17: Matrizes de confusão para os quatro melhores resultados de predição de comportamento com a aplicação do *X-Means*.

e a classe L 0,847. Enquanto que a *F-Measure ponderada* fica em 0,793, sendo 0,225 para a classe $\neg L$ e 0,882 para a classe L.

Com base nos resultados de cada algoritmo, buscou-se mostrar que o desempenho preditivo com base nas saídas do *f-DOPE* é melhor em comparação ao desempenho preditivo baseado nas saídas do *X-Means*. Para esta avaliação utilizou-se os valores da medida *F-Measure* dos resultados selecionados para ambas abordagens. Contudo, como alguns dos valores possíveis de τ_{match} foram selecionados somente para a classificação baseada nas saídas do *f-DOPE* ([0,80; 0,85; 0,90]), estes foram desprezados. Assim, os valores selecionados estão descritos na Tabela 6.18. Nesta comparação a predição com base nas saídas do *f-DOPE* obteve 6 êxitos contra 1 da predição realizada com base nas em saídas do *X-Means*. Além disso, ao comparar todos os resultados das predições com base em ambas as saídas (ver Tabelas G.1 e H.1), removendo os resultados onde só existe uma uma classe possível ([0,55;0,30], [0,55;0,35], [0,55;0,40], [0,60;0,30]), [0,60;0,35], [0,60;0,40], [0,65;0,25], [0,65;0,30], [0,65;0,35], [0,65;0,40], [0,70;0,20], [0,70;0,25], [0,70;0,30], [0,70;0,35], [0,70;0,40], [0,80;0,20], [0,80;0,25], [0,80;0,30], [0,80;0,35], [0,80;0,40], [0,85;0,15], [0,85;0,20], [0,85;0,25], [0,85;0,30], [0,85;0,35], [0,85;0,40], [0,90;0,15], [0,90;0,20], [0,90;0,25], [0,90;0,30], [0,90;0,35], [0,90;0,40]) a predição com base nas saídas do *f-DOPE* obtem 16 êxitos contra 10 da predição baseada nas saídas do *X-Means*. Dessa forma, há evidências suficientes em ambos resultados para afirmar que a predição com base nas saídas do *f-DOPE* é significativamente melhor.

Pode-se dizer que a aplicação do *f-DOPE* mantém um nível de identificação de perfis e comportamentos que reflete o cenário abordado onde existem mais comportamento L. Com os conjuntos de *ciclos comportamentais* obtidos pelo *f-DOPE* o classificador consegue ter uma melhor previsão para comportamentos da última semana indicando que o conjunto de *ciclos comportamentais* exemplifica melhor o que pode ocorrer no futuro. Isso demonstra que os comportamentos

Base de dados		F-Measure ponderado	
τ_{match}	τ_{split}	*f-DOPE*	*X-Means*
0,50	0,10	**0,781**	0,676
0,55	0,10	**0,901**	0,677
0,60	0,10	**0,901**	0,678
0,65	0,10	**0,901**	0,725
0,70	0,10	**0,901**	0,740
0,75	0,10	**0,901**	0,739
0,75	0,15	0,654	**0,740**
Êxitos		6	1

Tabela 6.18: Comparação dos resultados de *F-Measure* obtidos na predição de comportamentos com as saídas do *f-DOPE* e *X-Means* para os sete melhores resultados de τ_{match} encontrados em ambas execuções.

identificados são consistentes o que pode ajudar na identificação de comportamentos desejados pelas partes interessadas.

6.7 Considerações Finais do Capítulo

Neste capítulo foram apresentados diferentes resultados para avaliar o comportamento do *framework f-DOPE* em comparação com o *X-Means*, o qual é uma abordagem utilizada por trabalhos apresentados no Capítulo 4. Inicialmente foram apresentados experimentos baseados em um FCD real de uma fabricante de dispositivo móveis internacional. Com base em tal FCD foram realizados experimentos para a fase de Mineração do FCD do *f-DOPE* que resultou na obtenção de perfis de uso ao longo de várias semanas do FCD. Por outro lado, parte dos experimentos inicias, como seleção dos *aplicativos mais utilizados* e *remanescentes* foi utilizado para a geração de perfis com o *X-Means* visando manter o mesmo conjunto de dados em ambos os casos.

Após a identificação dos perfis, a segunda etapa do *f-DOPE* foi avaliada em comparação ao *X-Means* visando o monitoramento de perfis e comportamentos. Neste caso, experimentos foram realizados visando a obtenção de *ciclos comportamentais*. Por fim, foi realizada a predição de comportamentos na décima janela de eventos do FCD utilizado buscando encontrar os melhores resultados para diferentes combinações dos limiares τ_{match} e τ_{split}, os quais são fundamentais na geração dos *ciclos comportamentais*.

Pode-se verificar que os experimentos oriundos dos *ciclos comportamentais* gerados pelo *f-DOPE* apresentam melhores resultados apoiando a hipótese de que *f-DOPE* é capaz de identificar

perfis e ajudar no monitoramento de perfis e comportamentos para o domínio de aplicação abordado. Além disso, os resultados mostram que a predição realizada é significativamente melhor com base nos *ciclos comportamentais* gerados pela aplicação do *f-DOPE*.

Desta forma, acredita-se que o *f-DOPE* é mais sensível na identificação de perfis e comportamento, assim como na sensibilidade a mudanças na distribuição do FCD. Tais resultados motivam o prosseguimento da pesquisa e a exploração de novas possibilidades para tal *f-DOPE*. Como o *f-DOPE* é configurável os valores descritos nos experimentos realizados podem ser utilizados como estimativas para contribuições futuras. Por fim, os resultados obtidos ao longo desta pesquisa foram descritos em artigo científicos submetidos para conferências e revistas conforme descrito no Capítulo 7, Seção 7.4. Mais ainda, no Capítulo 7 são apresentadas as considerações finais desta pesquisa.

7. CONSIDERAÇÕES FINAIS

Este Capítulo visa apresentar as principais conclusões obtidas ao longo desta pesquisa. São apresentadas as contribuições da pesquisa (ver Seção 7.1), as limitações existentes (ver Seção 7.2) e os trabalhos futuros que podem ser abordados (ver Seção 7.3). Por fim são apresentadas as publicações realizadas no decorrer do presente trabalho (ver Seção 7.4).

7.1 Contribuições

Por meio do estudo realizado é possível observar que existe pouca informação relacionada a identificação de perfis de usuários, bem como a investigação comportamental do uso de aplicações em dispositivos móveis. A maioria dos trabalhos encontrados, relacionados a identificação de perfis de uso, são análises de dados em *batch* que, normalmente, não utilizam informações sobre uso de aplicativos. Desta forma, o maior desafio está em identificar e analisar os diferentes padrões de uso de aplicativos em dispositivos móveis em um cenário de FCD, monitorando as mudanças de comportamentos de usuários no decorrer do tempo, as quais podem apontar comportamentos que podem ser importante para várias partes interessadas.

Conforme apresentado no Capítulo 4, por meio de uma revisão sistemática da literatura, foram encontrados muitos trabalhos propostos visando a tarefa de identificação e/ou a tarefa de monitoramento. Dentre as limitações existentes em tais trabalhos, é notável a falta de exploração de cenários de FCD ao longo de Janelas de Eventos, a ausência de processos de monitoramento de comportamentos e a pequena quantidade de clientes e/ou aplicativos analisados, mesmo em cenários estacionários. Assim, o problema de identificação e monitoramento de perfis e comportamentos que visa ajudar empresas fabricantes de dispositivos móveis pode ser visto como um problema em aberto.

Portanto, além de acrescentar na qualidade e na otimização de processos aplicados à FCD a principal contribuição desta tese é o desenvolvimento de um novo e efetivo *framework*, chamado *f-DOPE*, que busca identificar e monitorar perfis e comportamento com base em FCD de uso de aplicativos em dispositivos móveis. Tais perfis e comportamentos podem ser úteis para diferentes partes interessadas como empresas fabricantes deste tipo de dispositivo. Especificamente, o *f-DOPE*, apresentado no Capítulo 5, foi projetado em dua etapas, chamadas de Mineração do FCD e Acompanhamento do FCD, onde são aplicadas técnicas de Mineração de Dados, tarefas de Aprendizado de Máquina não supervisionado e técnicas de Detecção de Novidade.

Assim, as contribuições desta tese são:

- **Uma revisão sistemática da literatura.** Foi realizada uma revisão da literatura (ver Capítulo 4) visando encontrar trabalhos relacionados ao problema abordado por esta pesquisa. Foram encontrados vários estudos que foram investigados e detalhados buscando encontrar o estado da arte do problema abordado.

- **Uma nova abordagem para capturar, analisar, pre-processar e sumarizar diferentes tipos de dados do uso de aplicativos móveis.** Uma das principais contribuições está em detectar padrões de uso de aplicativos em dispositivos móveis buscando identificar as diferentes formas de utilização destes aplicativos em tais dispositivos. Assim, é aplicado a Tarefa de Mineração de Regras de associação. Além disso, é apresentado no Capítulo 6, Seção 6.3.1 resultados da primeira fase do *framework f-DOPE* que demonstram a necessidade de tal etapa no cenário abordado.

- **Uma nova maneira aplicar padrões de uso de aplicativos na identificação de perfis de consumidores.** Outra importante contribuição está em criar perfis com base em padrões de uso de aplicativos em FCD. Este objetivo foi alcançado por meio de um cálculo de similaridade, baseado em padrões de uso de aplicativos, com a aplicação da medida *Jaccard* modificada conforme Equação 5.3. São apresentados no Capítulo 6, Seções 6.3.2 e 6.3.3 resultados da primeira etapa do *framework f-DOPE* que demonstram a eficácia de tal etapa em relação a tal contribuição.

- **Um novo processo para monitorar perfis e também comportamentos de consumidores ao longo de FCD de uso de aplicativos móveis**. Mais uma contribuição é o desenvolvimento de uma etapa do *framework f-DOPE* onde perfis e comportamentos de uso de aplicativos são monitorados, no qual mudanças de conceitos e mudanças de perfis são investigadas. Nesse sentido, a representação de perfis por *enumeração*, com a adaptação de tal abordagem e a busca por mudanças de perfis e comportamentos gerando diferentes *ciclos comportamentais* são aplicados para alcançar tal contribuição. Os resultados da segunda etapa do *framework f-DOPE* que demonstram a importância de tal etapa são apresentados no Capítulo 6, Seção 6.3.

Por fim, em uma visão de Ciência da Computação, foi desenvolvido uma aplicação para mineração e monitoramento de FCD de uso de aplicativos que podem afetar a indústria de dispositivos móveis.

7.2 Limitações

A revisão sistemática realizada durante esta pesquisa tem como limitação o fato de que a mesma foi realizada por apenas dois pesquisadores. Embora esta pesquisa tenha seguido um protocolo e tenha sido rigorosamente conduzida, é de suma importância a discussão entre pesquisadores adicionais interessados nesta área, principalmente em relação aos julgamentos a serem alcançados sobre os estudos descobertos.

Outra limitação, é referente a dificuldade de se obter conjuntos de dados públicos. Com a existências de tais conjuntos seria possível realizar mais comparações com outras abordagens da literatura.

Em geral, o *framework f-DOPE* pode ser calibrado em diferentes aspectos. Entre os parâmetros que podem ser calibrados estão, τ_{most}, τ_{rem}, *minTime*, τ_{pop}, *minLen*, *minSup*, *minAllConf*, *minConf*, *minLift*, τ_{match} e τ_{split}. Além de tais parâmetros, podem ser abordadas outros tipos de janela de eventos, de tamanho de tais janela e os algoritmos para: i) discretização, ii) regras de associação, iii) agrupamento, iv) medida de agrupamento, v) detecção de novidade e vi) monitoramento de comportamentos. A busca pelas melhores abordagens para o cenário de FCD não é uma tarefa fácil. Por meio de experimentos foi possível verificar que algumas vezes o desempenho na obtenção dos *ciclos comportamentais* foge do cenário abordado, sendo encontrados ciclos somente com comportamentos $\neg L$ quando os parâmetros não são bem calibrados.

O *framework f-DOPE* é aplicado em valores de tempo de uso de aplicativos e sua utilização em outros cenários similares pode ser possível. Contudo, a inclusão de novos tipos de valores depende de adaptações a serem realizadas, principalmente na fase de Absorção de tal *framework*.

7.3 Trabalhos Futuros

Como trabalhos futuros, pretende-se ampliar o número de testes com o *framework f-DOPE*, buscando-se avaliar a possibilidade de alcançar melhores resultados na segmentação de *ciclos comportamentais* com um sistema de previsão de rotatividade que realize previsões imediatas. Outro ponto a ser explorado futuramente, diz respeito ao uso de novos métodos para mineração de regras de associação para dados de dispositivos móveis (Hsu, 2017). Por fim, planeja-se aumentar a gama de recursos usados no *f-DOPE*, como dados de tráfego na internet, nível de armazenamento, nível de bateria e sensores.

7.4 Publicações

Além de obter resultados interessantes e de qualidade por meio da aplicação do *framework* proposto, esse estudo também resultou na escrita de cinco artigos:

- O primeiro artigo (Machado e Ruiz, 2017) foi submetido, aceito e apresentado presencialmente na conferência IWCMC 2017 (*The 13th International Wireless Communications and Mobile Computing Conference*), que possui *Qualis A2*. Tal conferência é reconhecida pelo foco na área de dispositivos móveis. No artigo aceito foram apresentados a primeira versão do *f-DOPE*, bem como alguns resultados preliminares.

- Outro artigo, com resultados dos estudos de caso aparentados nos Apêndices (Machado e Ruiz, 2018), foi submetido e aceito na conferência SBBD 2018 (*The 33rd Brazilian Symposium on Databases*), que possui *Qualis B2*. Nesta conferência tal trabalho foi presencialmente apresentado e avaliado no WTDBD (*Thesis and Dissertations Workshop*).

- Os três FCD abordados durante esta pesquisa serviram como inspiração para a simulação de um FCD de uso de aplicativos em dispositivos móveis, visando ser o primeiro grande conjunto de dados de eventos de uso de aplicativos disponível publicamente. Tal simulação foi cuidadosamente realizada buscando manter os padrões analisados no cenário real abordado por esta pesquisa. Este artigo foi submetido para a conferência CIKM (*ACM International Conference on Information and Knowledge Management*), que possui *Qualis A1*, onde foi rejeitado estando agora em processo de revisão para ser submetido em breve para uma revista a ser definida.
- A revisão sistemática da literatura foi submetida para a revista *ARI* (*Artificial Intelligence Review*) que possui *Qualis A1*. Tal artigo ainda estava em revisão até o final da escrita desta tese.
- O *framework f-DOPE* desenvolvido ao longo desta pesquisa e os experimentos apresentados pelo Capítulo 6 foram descritos em um artigo que será submetido para a revista *TMC* (IEEE Transactions on Mobile Computing) que possui *Qualis A1*.

REFERÊNCIAS BIBLIOGRÁFICAS

Ackermann, M. R., Märtens, M., Raupach, C., Swierkot, K., Lammersen, C. e Sohler, C. (2012). Streamkm++: A clustering algorithm for data streams. *Journal of Experimental Algorithmics (JEA)*, vol. 17, pp. 2–4.

Aggarwal, C. C. (2003). A framework for diagnosing changes in evolving data streams. In: *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pp. 575–586, San Diego, California. ACM.

Aggarwal, C. C. (2007). *Data streams: models and algorithms*. Springer Science & Business Media, USA.

Aggarwal, C. C. (2013). An introduction to outlier analysis. In: *Outlier Analysis*, vol. 1, cap. 1, pp. 1–40. Springer, New York, USA.

Agrawal, R., Imieliński, T. e Swami, A. (1993). Mining association rules between sets of items in large databases. In: *ACM sigmod record*, pp. 207–216, New York, NY, USA. ACM.

Agrawal, R., Srikant, R. et al. (1994). Fast algorithms for mining association rules. In: *Proceedings of the 20th international conference on Very large data bases, VLDB*, pp. 487–499, San Francisco, CA, USA. ACM.

Almana, A. M., Aksoy, M. S. e Alzahrani, R. (2014). A survey on data mining techniques in customer churn analysis for telecom industry. *International Journal of Engineering Research and Applications*, vol. 45, pp. 165–171.

Alpaydin, E. (2014). *Introduction to machine learning*. MIT press, London, England.

Annie, A. (2017). Spotlight on consumer app usage part 1. Recuperado de: https://goo.gl/JUv8XP, Fevereiro 2019.

Atlas, D. (2016). Get the new mobile web intelligence report for q3 2016. Recuperado de: https://goo.gl/QW4FVU, Fevereiro 2019.

Babcock, B., Babu, S., Datar, M., Motwani, R. e Widom, J. (2002a). Models and issues in data stream systems. In: *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 1–16, Madison, Wisconsin. ACM.

Babcock, B., Datar, M. e Motwani, R. (2002b). Sampling from a moving window over streaming data. In: *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 633–634, San Francisco, California. Society for Industrial and Applied Mathematics.

Backiel, A., Baesens, B. e Claeskens, G. (2016). Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. *Journal of the Operational Research Society*, vol. 67, pp. 1135–1145.

Bahmani, B., Mohammadi, G., Mohammadi, M. e Tavakkoli-Moghaddam, R. (2013). Customer churn prediction using a hybrid method and censored data. *Management Science Letters*, vol. 3, pp. 1345–1352.

Ballea, B., Casasa, B., Catarineua, A. e Gavaldaa, R. (2013). The architecture of a churn prediction system based on stream mining. In: *Artificial Intelligence Research and Development: Proceedings of the 16th International Conference of the Catalan Association for Artificial Intelligence*, pp. 157–166, Catalonia, Spain. IOS Press.

Blondel, V. D., Decuyper, A. e Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science*, vol. 4, pp. 1–55.

Blondel, V. D., Esch, M., Chan, C., Clérot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z. e Ziemlicki, C. (2012). Data for development: the d4d challenge on mobile phone data. *ArXiv preprint arXiv:1210.0137*.

Böhmer, M., Hecht, B., Schöning, J., Krüger, A. e Bauer, G. (2011). Falling asleep with angry birds, facebook and kindle: a large scale study on mobile application usage. In: *Proceedings of the 13th international conference on Human computer interaction with mobile devices and services*, pp. 47–56, Stockholm, Sweden. ACM.

Brzeziński, D. (2010). Mining data streams with concept drift. (Dissertação de Mestrado), Poznan University of Technology, Poznań.

Chandola, V., Banerjee, A. e Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, vol. 41, pp. 15:1–15:58.

Chu, B.-H., Tsai, M.-S. e Ho, C.-S. (2007). Toward a hybrid data mining model for customer retention. *Knowledge-Based Systems*, vol. 20, pp. 703–718.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A. e Joshi, A. (2008). Social ties and their relevance to churn in mobile telecom networks. In: *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, pp. 668–677, Nantes, France. ACM.

Dawson, R. (2011). How significant is a boxplot outlier. *Journal of Statistics Education*, vol. 19, pp. 1–12.

de Montjoye, Y.-A., Smoreda, Z., Trinquart, R., Ziemlicki, C. e Blondel, V. D. (2014). D4d-senegal: the second mobile phone data for development challenge. *ArXiv preprint arXiv:1407.4885*.

Do, T. M. T., Blom, J. e Gatica-Perez, D. (2011). Smartphone usage in the wild: a large-scale analysis of applications and context. In: *Proceedings of the 13th international conference on multimodal interfaces*, pp. 353–360, Canary Islands, Spain. ACM.

Ester, M., Kriegel, H.-P., Sander, J. e Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, vol. 96, pp. 226–231, Portland, Oregon.

Fabbri, S., Silva, C., Hernandes, E., Octaviano, F., Di Thommazo, A. e Belgamo, A. (2016). Improvements in the start tool to better support the systematic review process. In: *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*, pp. 21:1–21:5, Limerick, Ireland. ACM.

Fan, W. e Bifet, A. (2013). Mining big data: current status, and forecast to the future. *ACM SIGKDD Explorations Newsletter*, vol. 14, pp. 1–5.

Faria, E. R., Gama, J. e Carvalho, A. C. (2013). Novelty detection algorithm for data streams multi-class problems. In: *Proceedings of the 28th annual ACM symposium on applied computing*, pp. 795–800, Coimbra, Portugal. ACM.

Farid, D. M., Zhang, L., Hossain, A., Rahman, C. M., Strachan, R., Sexton, G. e Dahal, K. (2013). An adaptive ensemble classifier for mining concept drifting data streams. *Expert Systems with Applications*, vol. 40, pp. 5895–5906.

Ferreira, D., Goncalves, J., Kostakos, V., Barkhuus, L. e Dey, A. K. (2014). Contextual experience sampling of mobile application micro-usage. In: *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, pp. 91–100, Toronto, ON, Canada. ACM.

Gaber, M. M. (2012). Advances in data stream mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, pp. 79–85.

Gama, J. (2010). *Knowledge discovery from data streams*. CRC Press, Boca Raton.

Gama, J. (2012). A survey on learning from data streams: current and future trends. *Progress in Artificial Intelligence*, vol. 1, pp. 45–55.

Gama, J. e Gaber, M. M. (2007). *Learning from data streams*. Springer, Hobart, TAS.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. e Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)*, vol. 46, pp. 44:1–44:37.

Gan, G., Ma, C. e Wu, J. (2007). *Data clustering: theory, algorithms, and applications*. SIAM, Philadelphia, PA.

Giannella, C., Han, J., Pei, J., Yan, X. e Yu, P. S. (2003). Mining frequent patterns in data streams at multiple time granularities. *Next generation data mining*, vol. 212, pp. 191–212.

Hajiha, A., Radfar, R. e Malayeri, S. S. (2011). Data mining application for customer segmentation based on loyalty: An iranian food industry case study. In: *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 504–508, Singapore. IEEE.

Hamka, F., Bouwman, H., De Reuver, M. e Kroesen, M. (2014). Mobile customer segmentation based on smartphone measurement. *Telematics and Informatics*, vol. 31, pp. 220–227.

Han, J., Pei, J. e Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier, USA.

Hsu, F.-M., Lu, L.-P. e Lin, C.-M. (2012). Segmenting customers by transaction data with concept hierarchy. *Expert Systems with Applications*, vol. 39, pp. 6221–6228.

Hsu, K.-W. (2017). Effectively mining time-constrained sequential patterns of smartphone application usage. In: *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, pp. 39:1–39:8, Beppu, Japan. ACM.

Jagadish, H. (2015). Big data and science: Myths and reality. *Big Data Research*, vol. 2, pp. 49–52.

Jain, A. K. e Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc., USA.

Jain, N. e Srivastava, V. (2013). Data mining techniques: A survey paper. *IJRET: International Journal of Research in Engineering and Technology*, vol. 2, pp. 116–119.

Jardine, N. e Sibson, R. (1971). Mathematical taxonomy. *London etc.: John Wiley*, vol. 3.

Kargupta, H., Bhargava, R., Liu, K., Powers, M., Blair, P., Bushra, S., Dull, J., Sarkar, K., Klein, M., Vasa, M. et al. (2004). Vedas: A mobile and distributed data stream mining system for real-time vehicle monitoring. In: *Proceedings of the 2004 SIAM International Conference on Data Mining*, pp. 300–311, Orlando, USA. SIAM.

Kargupta, H., Park, B.-H., Pittie, S., Liu, L., Kushraj, D. e Sarkar, K. (2002). Mobimine: Monitoring the stock market from a pda. *ACM SIGKDD Explorations Newsletter*, vol. 3, pp. 37–46.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, vol. 33, pp. 1–26.

Kranen, P., Assent, I., Baldauf, C. e Seidl, T. (2011). The clustree: indexing micro-clusters for anytime stream mining. *Knowledge and information systems*, vol. 29, pp. 249–272.

Krishnaswamy, S., Gama, J. e Gaber, M. M. (2012). Mobile data stream mining: From algorithms to applications. In: *IEEE 13th International Conference on Mobile Data Management (MDM)*, pp. 360–363, Bengaluru, Karnataka, India. IEEE.

Lauschke, C. e Ntoutsi, E. (2012). Monitoring user evolution in twitter. In: *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 972–977, Istanbul, Turkey. IEEE.

Lee, U., Lee, J., Ko, M., Lee, C., Kim, Y., Yang, S., Yatani, K., Gweon, G., Chung, K.-M. e Song, J. (2014). Hooked on smartphones: an exploratory study on smartphone overuse among college students. In: *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 2327–2336, Toronto, Ontario, Canada. IOS Press.

Li, G. e Deng, X. (2012). Customer churn prediction of china telecom based on cluster analysis and decision tree algorithm. In: *Proceedings of International Conference on Artificial Intelligence and Computational Intelligence*, pp. 319–327, Berlin, Heidelberg. Springer.

Li, H., Lu, X., Liu, X., Xie, T., Bian, K., Lin, F. X., Mei, Q. e Feng, F. (2015). Characterizing smartphone usage patterns from millions of android users. In: *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pp. 459–472, Tokyo, Japan. ACM.

Liu, H., Hussain, F., Tan, C. L. e Dash, M. (2002). Discretization: An enabling technique. *Data mining and knowledge discovery*, vol. 6, pp. 393–423.

Machado, N. L. R. e Ruiz, D. D. A. (2017). Customer: A novel customer churn prediction method based on mobile application usage. In: *Proceedings of the 13th International conference on Wireless Communications and Mobile Computing Conference (IWCMC), 2017*, pp. 2146–2151, Valencia, Spain. IEEE.

Machado, N. L. R. e Ruiz, D. D. A. (2018). A framework for identification and monitoring of profiles and behaviors of users based on mobile app usage. In: *Proceedings of the 33rd Brazilian Symposium on Databases (SBBD), 2018*, pp. 88–94, Rio de Janeiro, Brazil. SBC.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA. University of California Press.

Markou, M. e Singh, S. (2003). Novelty detection: a review—part 1: statistical approaches. *Signal processing*, vol. 83, pp. 2481–2497.

Meireles, D. J. J. A. (2014). Previsão de churn em telecomunicações. (Dissertação de Mestrado), Universidade do Porto, Porto, Portugal.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill international editions - computer science series. McGraw-Hill Education, New York, USA.

Mohammad, Y., Matsumoto, K. e Hoashi, K. (2017). A dataset for activity recognition in an unmodified kitchen using smart-watch accelerometers. In: *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*, pp. 63–68, Stuttgart, Germany. ACM.

Nilsson, N. J. (1996). *Introduction to machine learning. An early draft of a proposed textbook*. Citeseer, Stanford, CA.

Ntoutsi, E., Spiliopoulou, M. e Theodoridis, Y. (2012). Fingerprint: Summarizing cluster evolution in dynamic environments. *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 8, pp. 27–44.

Ntoutsi, I., Spiliopoulou, M. e Theodoridis, Y. (2009). Tracing cluster transitions for different cluster types. *Control & Cybernetics*, vol. 38, pp. 239–259.

Ntoutsi, I., Spiliopoulou, M. e Theodoridis, Y. (2011). Summarizing cluster evolution in dynamic environments. In: *Proceedings of the International Conference on Computational Science and Its Applications*, pp. 562–577, Berlin, Heidelberg. Springer.

Oliveira, M. e Gama, J. (2010a). Bipartite graphs for monitoring clusters transitions. In: *International Symposium on Intelligent Data Analysis*, pp. 114–124, Berlin, Heidelberg. Springer.

Oliveira, M. e Gama, J. (2010b). Understanding clusters evolution. In: *Proceedings of the Workshop on Ubiquitous Data Mining*, pp. 16–20, Lisbon, Portugal.

Oliveira, M. e Gama, J. (2012). A framework to monitor clusters evolution applied to economy and finance problems. *Intelligent Data Analysis*, vol. 16, pp. 93–111.

Oliveira, M. D. e Gama, J. (2010c). Mec-monitoring clusters' transitions. In: *Proceedings of the Fifth Starting AI Researchers' Symposium (STAIRS)*, pp. 212–224, Lisbon, Portugal. IOS Press.

Park, C. H. e Shim, H. (2010). Detection of an emerging new class using statistical hypothesis testing and density estimation. *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 24, pp. 1–14.

Pelleg, D., Moore, A. W. et al. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In: *International Conference on Machine Learning (ICML)*, vol. 1, pp. 727–734, Haifa, Israel.

Pereira, G. e Mendes-Moreira, J. (2016). Monitoring clusters in the telecom industry. In: *New Advances in Information Systems and Technologies*, vol. 45, pp. 631–640. Springer, Cham.

PhridviRaj, M. e GuruRao, C. (2014). Data mining–past, present and future–a typical survey on data streams. *Procedia Technology*, vol. 12, pp. 255–263.

Portal, S. (2016). Statistics and facts about mobile app usage. Recuperado de: https://goo.gl/W5BDI2. Fevereiro 2019.

Portal, S. (2019). Number of apps available in leading app stores. Recuperado de: https://goo.gl/3FLn4f. Janeiro 2019.

Pyo, S., Kim, E. et al. (2015). Lda-based unified topic modeling for similar tv user grouping and tv program recommendation. *IEEE transactions on cybernetics*, vol. 45, pp. 1476–1490.

Rehman, A. e Raza Ali, A. (2015). Customer churn prediction, segmentation and fraud detection in telecommunication industry. In: *Proceedings of the 4th ASE International Conference on Big Data, Harvard University*, pp. 1–9, USA. Academy of Science and Engineering.

Rizoiu, M.-A., Velcin, J., Bonnevay, S. e Lallich, S. (2015). Cluspath: a temporal-driven clustering to infer typical evolution paths. *Data Mining and Knowledge Discovery*, vol. 30, pp. 1–26.

Shabana, K., Wilson, J. e Chaudhury, S. (2016). A multi-view non-parametric clustering approach to mobile subscriber segmentation. In: *Proceedings of the IEEE 18th Conference on Business Informatics (CBI)*, vol. 1, pp. 173–181, Paris, France. IEEE.

Siddiqui, Z. F., Krempl, G., Spiliopoulou, M., Pena, J. M., Paul, N. e Maestu, F. (2015). Predicting the post-treatment recovery of patients suffering from traumatic brain injury (tbi). *Brain Informatics*, vol. 2, pp. 33–44.

Siddiqui, Z. F., Oliveira, M., Gama, J. e Spiliopoulou, M. (2012). Where are we going? predicting the evolution of individuals. In: *Proceedings of the International Symposium on Intelligent Data Analysis*, pp. 357–368, Helsinki, Finland. Springer.

Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., de Carvalho, A. C. e Gama, J. (2013). Data stream clustering: A survey. *ACM Computing Surveys (CSUR)*, vol. 46, pp. 13:1–13:37.

Simon, H. A. (1996). *The sciences of the artificial*. MIT press, London, England.

Sohn, S. Y. e Kim, Y. (2008). Searching customer patterns of mobile service using clustering and quantitative association rule. *Expert systems with Applications*, vol. 34, pp. 1070–1077.

Spiliopoulou, M., Ntoutsi, E., Theodoridis, Y. e Schult, R. (2013). Monic and followups on modeling and monitoring cluster transitions. In: *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 622–626, Prague, Czech Republic. Springer.

Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y. e Schult, R. (2006). Monic: modeling and monitoring cluster transitions. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 706–711, Philadelphia, PA, USA. ACM.

Tan, P.-N., Steinbach, M., Kumar, V. et al. (2006). *Introduction to data mining*. Pearson Addison Wesley Boston, Boston, MA, USA.

Tew, C., Giraud-Carrier, C., Tanner, K. e Burton, S. (2014). Behavior-based clustering and analysis of interestingness measures for association rule mining. *Data Mining and Knowledge Discovery*, vol. 28, pp. 1004–1045.

Tibshirani, R., Walther, G. e Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, pp. 411–423.

Tseng, W.-R. e Hsu, K.-W. (2014). Smartphone app usage log mining. *International Journal of Computer and Electrical Engineering*, vol. 6, pp. 151–156.

Vendramin, L., Campello, R. J. e Hruschka, E. R. (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, vol. 3, pp. 209–235.

Verkasalo, H. (2009). Contextual patterns in mobile service usage. *Personal and Ubiquitous Computing*, vol. 13, pp. 331–342.

Wagner, D. T., Rice, A. e Beresford, A. R. (2013). Device analyzer: Understanding smartphone usage. In: *Proceedings of the International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*, pp. 195–208, Tokyo, Japan. Springer.

Wei, C.-P. e Chiu, I.-T. (2002). Turning telecommunications call details to churn prediction: a data mining approach. *Expert systems with applications*, vol. 23, pp. 103–112.

Witten, I. H. e Frank, E. (2011). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, Burlington, MA.

Witten, I. H., Frank, E., Hall, M. A. e Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, Burlington, MA.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, pp. 38:1–38:10, London, UK. ACM.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y. et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, vol. 14, pp. 1–37.

Xiong, H., Tan, P.-N. e Kumar, V. (2003). Mining strong affinity association patterns in data sets with skewed support distribution. In: *Proceedings of the Third IEEE International Conference on Data Mining (ICDM)*, pp. 387–394, Melbourne, FL, USA. IEEE.

Xu, Q., Erman, J., Gerber, A., Mao, Z., Pang, J. e Venkataraman, S. (2011). Identifying diverse usage behaviors of smartphone apps. In: *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pp. 329–344, Berlin, Germany. ACM.

Zhang, G. (2007). Customer segmentation based on survival character. In: *International Conference on Wireless Communications, Networking and Mobile Computing (WiCom)*, pp. 3391–3396, United States. IEEE.

Zhu, T., Wang, B., Wu, B. e Zhu, C. (2011). Role defining using behavior-based clustering in telecommunication network. *Expert Systems with Applications*, vol. 38, pp. 3902–3908.

APÊNDICE A – OUTROS FCDS DE USO DE APLICATIVOS UTILIZADOS NA PESQUISA

Neste Apêndice, são apresentados FDCs utilizados ao longo desta pesquisa. Tais FCDs são de uma das maiores fabricantes mundiais de dispositivos móveis, a qual monitora milhões de *smartphones* diariamente. É importante salientar que tal empresa está em conformidade com os requisitos de proteção de dados, como por exemplo o *GDPR* (*General Data Protection Regulation*)[1], que permite a utilização de dados capturados sem violar a ética da pesquisa. Dentre os FCD utilizados, cada um possui dados de um mesmo modelo de dispositivo. Tais modelos foram escolhidos conforme a evolução e o lançamento de novas versões visando a experimentação do *framework* desenvolvido em dados oriundos de modelos atuais e mais utilizados. Por fim, dois FCDs foram abordados durante os estudos de caso, ambos estão descritos abaixo. Além disso, um terceiro FCD é utilizado para os experimentos finais, o qual é descrito na Seção 6.1.

Ambos FCDs possuem eventos de uso de aplicativos em dispositivos móveis do Brasil. Além disso, cada FCD contém eventos gerados por um único modelo de dispositivo.

O primeiro FCD, chamado de *DS01*, possui eventos capturados por um período contínuo de 70 dias, iniciando em 28 de Fevereiro e terminando em 07 de Maio de 2016. O *DS01* contém um total de 173, 590, 133 eventos de uso de aplicativos oriundos de 24, 782 dispositivos móveis[2], onde 44, 608 aplicativos distintos foram utilizados por tais usuários.

O segundo FCD, chamado de *DS02*, contém eventos capturados por um período contínuo de 140 dias, iniciando em 11 de Dezembro de 2016 e terminando em 29 de Abril de 2017. O *DS02* contém um total de 1, 045, 013, 673 eventos de uso de aplicativos proveniente de 34, 552 dispositivos móveis[3], onde 60, 116 aplicativos distintos foram utilizados.

No FCD *DS02*, uma parte do período de tempo foi utilizado para detectar se os usuários mantiveram o uso de seus dispositivos após as 10 semanas de análise. Neste caso, de 19 de Fevereiro até 29 de Abril de 2017, onde 506.013.782 eventos foram capturados, e usuários que deixaram de utilizar seus dispositivos por 40 dias ou mais foram rotulados como usuários que cometeram abandono. Esta rotulação será utilizada como forma de avaliar a segmentação do *framework* proposto. A Tabela A.1 mostra o número de usuários e o total de eventos para cada janela de tempo ao longo dos dois FCD.

Na Tabela A.1, em cada janela de tempo é possível verificar uma grande quantidade de eventos gerados, em média 17.359.013 eventos para o FCD *DS01* e 53.899.989 eventos para o FCD *DS02*. Além disso, o número de usuários é reduzido com o passar do tempo. Esta redução ocorre pela realização de trocas de dispositivos pelos usuários. Alguns usuários trocam seus dispositivos por outros da mesma fabricante (clientes leais), enquanto outros trocam por dispositivos de outras

[1] General Data Protection Regulation - https://gdpr-info.eu/
[2] Não foram adicionados novos dispositivos com o passar do tempo.
[3] Não foram adicionados novos dispositivos com o passar do tempo.

DS01				*DS02*			
Janela	**Dispositivos**		**Eventos**	**Janela**	**Dispositivos**		**Eventos**
	Total	**%**			**Total**	**%**	
28/02 até 05/03	24.782	100,00%	16.775.679	11/12 até 17/12	34.552	100,00%	56.419.520
06/03 até 12/03	24.487	98,81%	18.586.997	18/12 até 24/12	34.188	98,95%	55.211.335
13/03 até 19/03	24.302	98,06%	18.226.072	25/12 até 31/12	33.947	98,25%	54.152.246
20/03 até 26/03	24.108	97,28%	17.683.554	01/01 até 07/01	33.789	97,79%	54.055.848
27/03 até 02/04	23.936	96,59%	17.772.630	08/01 até 14/01	33.644	97,43%	53.412.248
03/04 até 09/04	23.827	96,15%	17.290.124	15/01 até 21/01	33.535	97,06%	53.747.594
10/04 até 16/04	23.726	95,74%	17.111.936	22/01 até 28/01	33.456	96,83%	53.313.103
17/04 até 23/04	23.560	95,07%	16.875.754	29/01 até 04/02	33.349	96,52%	53.148.299
24/04 até 30/04	23.428	94,54%	16.870.939	05/02 até 11/02	33.253	96,24%	52.987.113
01/05 até 07/05	23.314	94,08%	16.396.448	12/02 até 18/02	33.213	96,12%	52.552.585
μ	23.947	96,63%	17.359.013	μ	33.695	97,52%	53.899.989
σ	450	1,82%	658.825	σ	410	1,19%	1.094.304

Tabela A.1: Quantidade de eventos em cada uma das dez semanas dos FCDs *DS01* e *DS02*, bem como o número total e a porcentagem de dispositivos para cada janela de tais FCDs. Abaixo a média e o desvio-padrão de cada elemento.

fabricantes (*churn*). Alguns usuários também podem não apresentar uso de aplicativos em uma determinada semana, o que ocorre por outros motivos, como por exemplo férias.

Os modelos dos dispositivos selecionados funcionam sobre um Sistema Operacional - SO *Android*. Neste tipo de SO é possível a aquisição de diferentes aplicações por meio de lojas de aplicativos *online*, também chamadas de *marketplaces*, como por exemplo a *Google Play Store*. Foram analisados eventos dos FCDs de todos os aplicativos adquiridos em tais lojas, e de quase todos os aplicativos nativos. Aplicativos nativos são aplicativos instalados nos dispositivos antes destes chegarem aos usuários finais. Estes aplicativos geram eventos de uso sem realmente serem utilizados pelos usuários. A Tabela A.2[4] mostra os aplicativos nativos que foram descartados por não serem realmente usados pelos usuários.

Alguns dos aplicativos nativos são adicionados juntamente com o SO, enquanto outros são adicionados por terceiros (ex: operadoras de telecom). A maioria dos aplicativos nativos são denominados *launchers*, os quais são adicionados com o SO.

[4]Somente as partes de interesse do nome de cada aplicativo são apresentados no decorrer deste estudo.

[5]*device_provider* é a anonimização do nome da empresa responsável pelos FCDs, a qual não deve ser mencionada devido a questões de privacidade.

Aplicativos nativos
com.android.systemui
com.google.android.packageinstaller
com.android.packageinstaller
android
com.*device_provider*[5].setup
com.*device_provider*.storageoptimizer

Tabela A.2: Aplicativos nativos desconsiderados.

APÊNDICE B – ESTUDO DE CASO DA FASE DE ABSORÇÃO

Neste Apêndice são apresentados estudos de casos relacionados a fase de Absorção do *framework f-DOPE*. Inicialmente, é apresentado o estudo de caso onde buscou-se a identificação de perfis com uso de categorias de aplicativos. Após, são apresentados os estudos de caso buscando investigar aplicativos *mais utilizados*, *populares* e também o tempo de uso de tais aplicativos. Por fim, são apresentados os estudos de casos para diferentes técnicas de discretização, *frequência igual*, *agrupamento* e *IP*.

Identificação de Perfis por meio da Categorização de Aplicativos

Neste estudo foi encontrado um número significativo de aplicativos para dispositivos móveis a partir de todos os eventos gerados pelo FCDs utilizados nos estudos de caso (ver Apêndice A). Porém, poucos aplicativos são efetivamente utilizados pelos usuários. Contudo, para tentar entender esta utilização e buscar encontrar diferentes perfis que representem padrões de uso de aplicativos, inicialmente tentou-se adicionar categorias a cada um destes aplicativos. Categorias de aplicativos são utilizadas pelos desenvolvedores para disponibilizar estes aplicativos em *marketplaces*. Dessa forma, o número de atributos a serem utilizados foi reduzido (de aplicativos distintos para categorias distintas) assim como sua esparcidade. Para essa avaliação foi utilizado o FCD *DS01*.

Para adicionar categorias aos aplicativos, foi utilizada uma API (Interface de Programação de Aplicação - *Application program interface*) fornecido pela *Google Play Store*. Assim, buscou-se o nome do pacote de cada aplicativo, visando encontrar suas respectivas categorias. É importante notar que tipicamente em *marketplaces* de aplicativos para dispositivos móveis, cada aplicativo pode somente pertencer a uma categoria (ex: Social). Além disso, *marketplaces* distintas possuem diferentes tipos de categorias. Na busca de cada *pkg*, algumas vezes não foi possível encontrar a categoria de um determinado aplicativo. Aplicativos não encontrados podem ser aplicativos, de teste (ex: utilizados por desenvolvedores), que são distribuídas por outros meios (ex: pre-instaladas por fabricantes de dispositivos), que foram removidos da loja *online* ou nativos (Böhmer et al., 2011). Nestes casos, os *pkg* são buscados em outras de *marketplaces*[1]. Quando o *pkg* não é encontrado em nenhuma *marketplace*, a categoria deste aplicativo é classificada como *desconhecida*.

Nesta análise, os milhões de eventos recebidos em cada janela de tempo foram sumarizados para cada dispositivo, por cada aplicativo e então por cada categoria. Por exemplo, um dispositivo possui dados de uso de aplicativos para diferentes categorias como Social e Comunicação. Na categoria Social foram sumarizadas o tempo de uso dos aplicativos *Facebook*, *Instagram*, *Snapchat*, enquanto na categoria Comunicação o tempo de uso dos aplicativos *WhatsApp*, *Viber* e *Telegram*. Uma vez que muitos preditores levam em consideração o tamanho relativo de diferentes atributos, mesmo que as escalas possam ser arbitrárias, foram aplicadas diferentes técnicas de preparação dos dados sumarizados em cada janela do FCD *DS01*. Tais técnicas são *normalização*, *padronização*

[1]Foram rastreadas lojas *online* de aplicativos para dispositivos *Android* até a quinta janela do FCD *DS01*.

e *transformação logarítmica* (Tan et al., 2006). Também foram executados diferentes algoritmos para a tarefa de Agrupamento, como *K-Means*, *WARD*, e *DBSCAN*, combinados com medidas de avaliação como *SWC*, *DUNN*, *CH*, e *Elbow method* (Vendramin et al., 2010). Combinações das técnicas de pré-processamento, dos algoritmos de agrupamento e das medidas de avaliação, foram aplicadas aos FCD das janelas analisadas. Como resultado, para todas as janelas, um número máximo de três grupos foi obtido. A Figura B.1 apresenta os resultados com a execução do algoritmo *K-Means* após a aplicação da técnica de *normalização* e das medidas de avaliação citadas acima na primeira janela do FCD *DS01*.

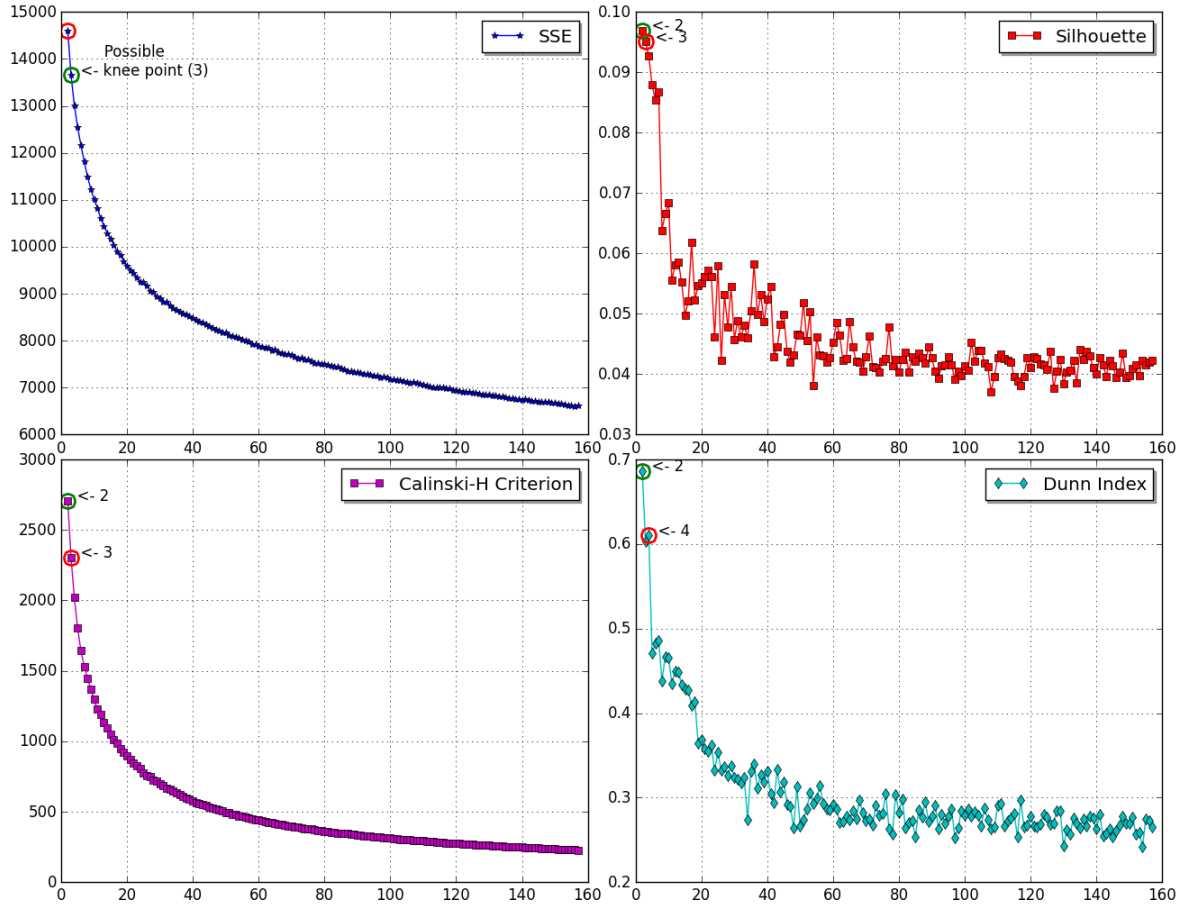


Figura B.1: Resultados das medidas de avaliação com a execução da técnica de *normalização* e do algoritmo *K-Means* na primeira janela do FCD *DS01*.

Na Figura B.1 o parâmetro k do algoritmo *K-Means*, representado no eixo x, foi variado de 2 até $\sqrt{N}$ (MacQueen et al., 1967). Para cada uma das medidas de avaliação, um circulo verde e outro vermelho foram apontados, os quais indicam respectivamente, melhor número de grupos e segundo melhor número de grupos a serem formados. Os resultados obtidos com o uso de categorias de aplicativos não representou um cenário real esperado, uma vez que este estudo trabalha com a hipótese da existência de diversos tipos de perfis que compõem a população de dispositivos, dado o número de aplicativos e modelos de dispositivos existentes. Portanto, como obteve-se, para todas as janelas, um número muito limitado de perfis, tais resultados sugerem que a categorização dos aplicativos não é a melhor abordagem a ser utilizada na descoberta de perfis de uso no cenário proposto. Alguns trabalhos, como (Böhmer et al., 2011; Li et al., 2015; Lee et al., 2014;

Xu et al., 2011) mostram que categorias são criadas e/ou agrupadas, pois diferentes taxonomias são utilizadas pelas *marketplaces*. Além disso, as categorias são principalmente escolhidas pelos desenvolvedores de aplicativos, os quais atribuem tais aplicativos a categorias no momento do envio para as *maketplaces* (Böhmer et al., 2011). Estas situações reforçam o indício de que o uso de categorias de aplicativos não é a melhor maneira para determinar perfis de uso.

Aplicativos Mais Utilizados e Aplicativos Populares

Em continuidade, buscou-se definir as importâncias dos milhares de aplicativos encontrados nos FCDs. Assim, o tempo de uso de cada aplicativo foi examinado buscando estabelecer uma quantidade de uso que seja relevante para determinar que um aplicativo realmente foi utilizado em um dispositivo. Nesse sentido, explorou-se como decidir sistematicamente quais aplicativos deveriam ser consideradas os mais relevantes. Para isso foi utilizado com base o número de dispositivos únicos que fazem uso de cada aplicativo. A Figura B.2 apresenta um histograma e uma função de distribuição acumulativa de dispositivos únicos para todos os aplicativos encontrados durante todo o FCD *DS01*.

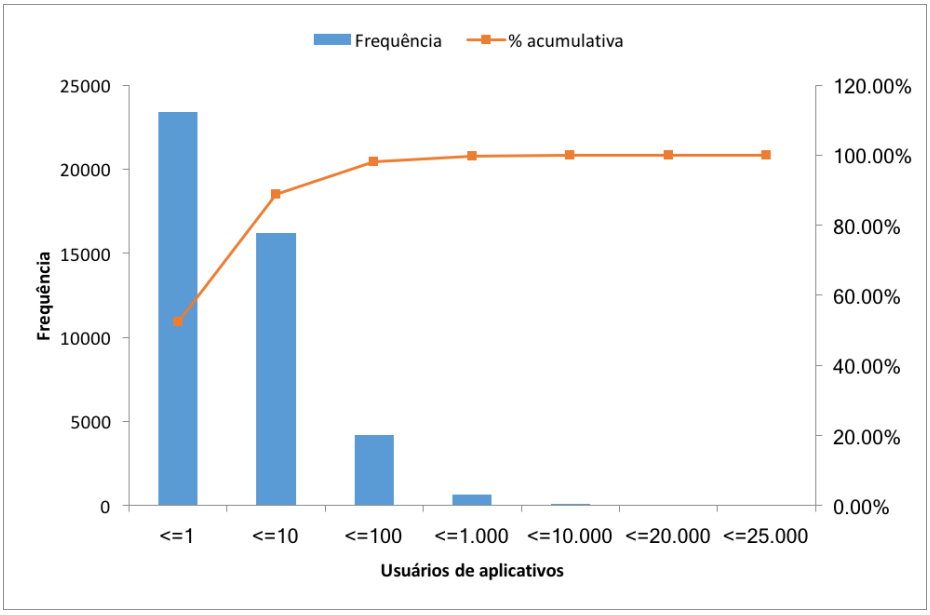


Figura B.2: Distribuição da frequência de aplicativos por usuários únicos considerando todo o período do FCD *DS01*.

É possível observar a existência de um número substancial de aplicativos utilizados em apenas um único dispositivo (Figura B.2 intervalo <= 1). Além disso, é possível identificar que somente cerca de 2% dos aplicativos são utilizados por mais de 100 (0.4%) (Figura B.2 intervalo <= 1000 até <= 25.000) dispositivos únicos. É importante notar que alguns aplicativos são utilizados em mais dispositivos do que outros. Neste sentido, é necessário selecionar somente aplicativos que são utilizados por um número significativo de usuários, os quais possuem atividades mínimas necessárias para ajudar na identificação de perfis de uso (Li et al., 2015). A Figura B.3 apresenta outro histograma e outra função de distribuição acumulativa de usuários únicos para todos os

aplicativos encontrados, desta vez, durante a primeira janela do FCD *DS01*. É possível verificar que função de distribuição acumulativa é muito similar com a apresentada para todo o FCD (Figura B.2). Esta mesma análise foi realizada nas demais janelas de tal FCD, onde foi possível observar que tal distribuição permaneceu sempre similar.

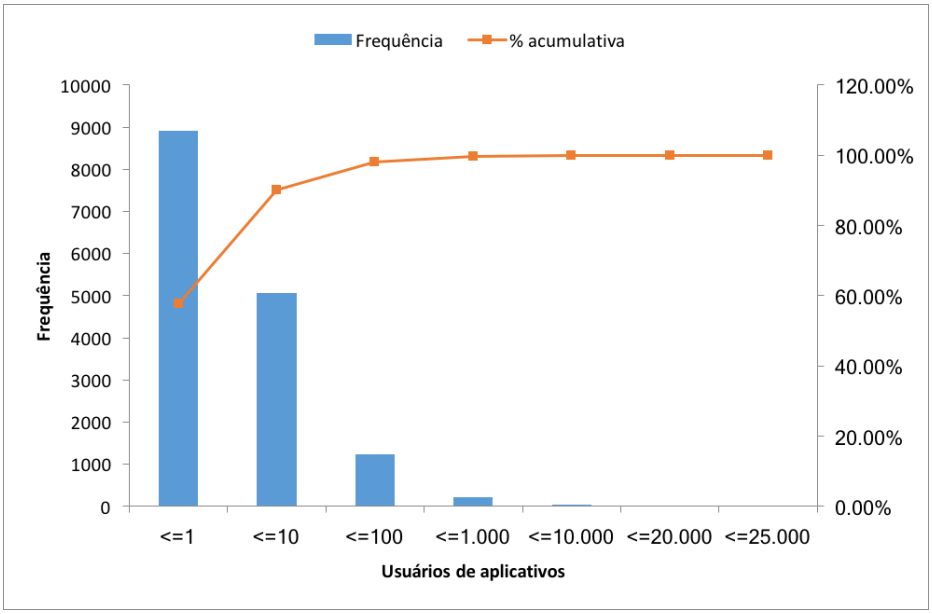


Figura B.3: Distribuição da frequência de aplicativos por usuários únicos considerando a primeira janela do FCD *DS01*.

Visando descartar os aplicativos que não fornecem uma quantidade aceitável de uso por dispositivo e selecionar apenas os aplicativos que possuem uma utilização mínima optou-se por selecionar somente os aplicativos utilizados por 1% ou mais usuários únicos. A definição de 1% como ponto de corte indica que em média 0.99% dos aplicativos encontrados são mantidos, o que representa uma média de 149 aplicativos a cada janela (Ver Tabela B.1). A Tabela B.1 mostra o número de aplicativos *mais utilizados*, bem como a porcentagem destes aplicativos em relação ao total de aplicativos na sua respectiva janela. Além disso, a Tabela B.1 apresenta o total de aplicativos encontrados para cada janela, bem como a porcentagem destes aplicativos em relação a todos os aplicativos detectados ao longo do FCD *DS01*.

A função de distribuição acumulativa apresentada na Figura B.4 mostra que a seleção de 100 ou mais aplicativos fornece uma cobertura de mais de 90% do tempo total de uso apresentado pela primeira janela do FCD *DS01* (linhas azuis). Na Figura B.4 o eixo x representa os aplicativos *mais utilizados* de duas formas. Ordenados em relação ao número total de dispositivos (linha vermelha tracejada). Neste caso, o aplicativo 1 é o aplicativo que possui o maior número de dispositivos únicos. E ordenados em relação ao tempo total de uso (linha verde). Assim, o aplicativo 1 é o aplicativo que apresenta o maior tempo total de uso. O eixo y, por sua vez, apresenta a porcentagem acumulativa do tempo total de uso dos principais aplicativos. Dessa forma, a linha verde demonstra a porcentagem acumulativa do tempo total de uso dos aplicativos iniciando pelo aplicativo mais utilizado em termos de tempo total de uso e a linha vermelha tracejada representa a porcentagem acumulativa de tempo total de uso dos aplicativos mais utilizando iniciando pelo aplicativo utilizado

Janela	Aplicativos					
	Total	**%**	***Mais utilizados***	**%**	***Populares***	**%**
28/02 até 05/03	15.714	35,23%	150	0,95%	33	0,21%
06/03 até 12/03	15.714	35,23%	151	0,96%	33	0,21%
13/03 até 19/03	15.459	34,66%	143	0,93%	34	0,22%
20/03 até 26/03	15.419	34,57%	144	0,93%	32	0,21%
27/03 até 02/04	15.014	33,66%	145	0,97%	33	0,22%
03/04 até 09/04	14.694	32,94%	147	1,00%	32	0,22%
10/04 até 16/04	14.723	33,01%	150	1,02%	34	0,23%
17/04 até 23/04	14.829	33,24%	144	0,97%	33	0,22%
24/04 até 30/04	14.371	32,22%	151	1,05%	34	0,24%
01/05 até 07/05	14.682	33,91%	160	1,09%	34	0,23%
μ	15.062	33,77%	149	0,99%	33	0,22%
σ	454	1,02%	5	0,05%	1	0,01%

Tabela B.1: Quantidade de aplicativos encontrados no FCD *DS01*. O número total e a porcentagem de todos aplicativos, aplicativos *mais utilizados*, e aplicativos *populares* para cada janela. Abaixo a média e o devio-padrão de cada elemento.

em mais dispositivos únicos. Assim, é possível afirmar que a escolha dos aplicativos *mais utilizados* pelo número de dispositivos únicos é possível.

A definição dos aplicativos *mais utilizados* reduz o escopo do número de aplicativo do conjunto de dados. A partir deste definição foi realizada uma nova tentativa da tarefa de Agrupamento, uma vez que o número de atributos é menor mas mantém uma cobertura significativa do tempo total de uso de todos os aplicativos. Contudo, tais resultados, continuaram apresentando a existência de no máximo 3 perfis e uso, como mostra a Figura B.5. A Figura B.5 apresenta resultados para as medidas *Elbow method* e *SWC* obtidas a partir da execução do algoritmo *K-Means*. Esta tarefa foi aplicada ao conjunto de dados da primeira janela do FCD *DS01* após o uso de transformação logarítmica, onde o parâmetro k foi variado de 2 até 40. É possível observar que não existe um cotovelo para a medida *Elbow method*, bem como uma grande valor para a medida de *SWC*, da mesma forma que ocorreu coma execução no conjunto de dados com uso de categorias de aplicativos.

Tempo de uso dos aplicativos

Além disso, foi realizado uma análise do tempo de uso de cada aplicativo visto o grande número de aplicativos utilizados nos dispositivos. Para esta verificação utilizou-se o FCD *DS02*. Neste FCD, em uma janela, cada dispositivo apresenta, em média, uso de 20 aplicativos distintos, o

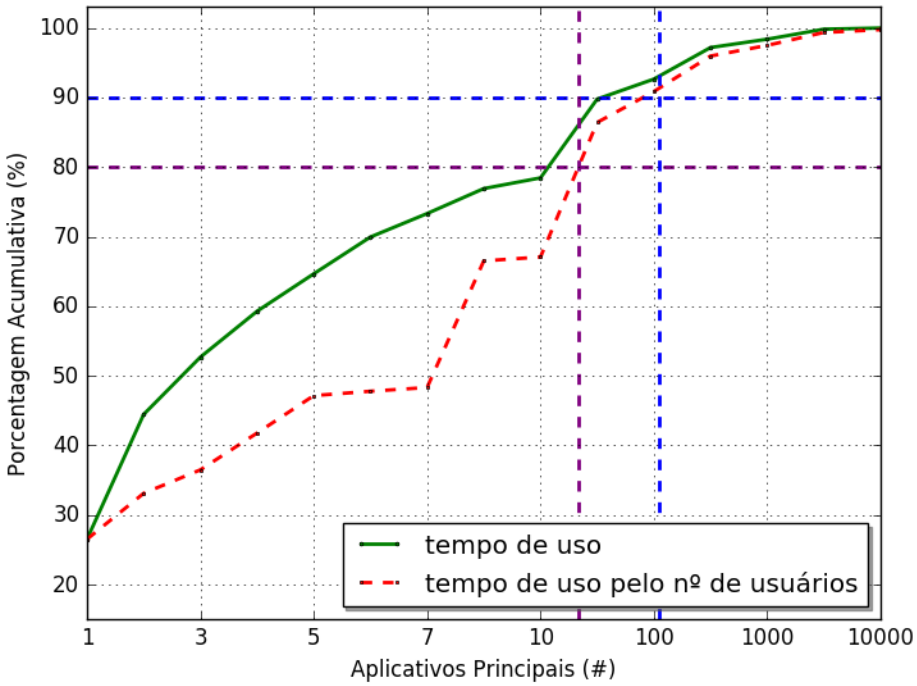


Figura B.4: Porcentagem acumulativa de tempo de uso total dos aplicativos *mais utilizados* na primeira janela do FCD *DS01*.

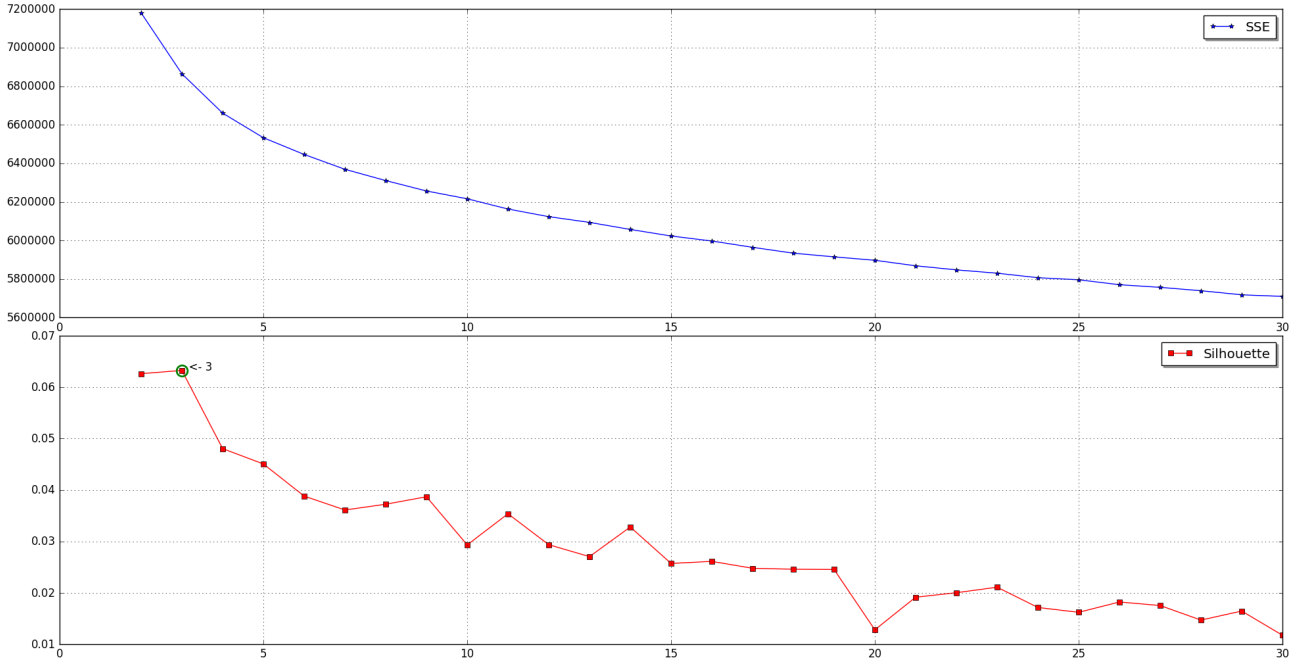


Figura B.5: Resultado das medidas de avaliação com a execução da técnica de transformação *logarítmica* e do algoritmo *K-Means* para a primeira janela do FCD *DS01*.

que representa um número um pouco elevado em comparação ao que é reportado em Annie (2017). Tais utilizações variam de 1,03 segundos até 133 horas por semana. Ferreira et al. (2014) afirmam que a utilização mínima aceitável de um aplicativo é cerca de 15 segundos, chamando este tipo de utilização de micro utilização. Assim, os aplicativos que são utilizados em um dispositivo por

menos tempo que um determinado limiar aceitável podem ser removidos, pois representam um uso irrelevante para tal dispositivo. Nesse sentido, é investigado como determinar o número real de aplicativos utilizados para cada dispositivo, tendo como base um limiar mínimo de utilização em comparação ao tempo total de uso de aplicativos para tal usuário. A utilização de um aplicativo por menos que 10 segundos pode simplesmente representar um erro cometido ao tentar abrir outro aplicativo. Além disso, alguns aplicativos são usados diariamente enquanto outros esporadicamente. Dessa forma, determinar um limiar de tempo de uso mínimo por usuário pode ajudar a determinar quais são os aplicativos que são realmente utilizados.

Inicialmente, para cada dispositivo obteve-se o tempo total de uso de aplicativos, onde foi realizada a soma do tempo de uso de todos os aplicativos utilizados em tal dispositivo. Quando um aplicativo de um determinado usuário apresentar um tempo total de uso menor que 1% do tempo total de uso de tal dispositivo ou menos de 70 segundos, este é desconsiderado para tal dispositivo. Dessa forma, somente aqueles aplicativos que representam ao menos 1% do *ttu* de um usuário que são também utilizados por mais de 70 segundos, são considerados os aplicativos realmente utilizados por tal usuário. Com a aplicação deste filtro o número médio de aplicativos por dispositivo, em uma janela, caiu para 14. A Tabela B.2 apresenta um resumo do FCD *DS02*, onde são apresentados o número de aplicativos *mais utilizados* após a definição do limiar mínimo de tempo de uso para cada dispositivo, bem como a porcentagem destes aplicativos em relação ao total de aplicativos na sua respectiva janela. Além disso a Tabela B.2 apresenta o total de aplicativos encontrados para cada janela, bem como a porcentagem destes aplicativos em relação a todos os aplicativos detectados ao longo do FCD *DS02*.

Em resumo, este processo adicional mostrou melhorar o conjunto de dados e seus resultados foram aplicados em outras abordagens realizadas pelas demais fases do *framework* proposto durante os estudos de caso (ver Apêndice C).

Aplicativos Populares

Todas os aplicativos *mais utilizadas* remanescentes foram ainda mais explorados. Alguns destes aplicativos foram utilizadas por 40% ou mais dispositivos únicos enquanto outros foram utilizadas por 90% ou mais dispositivos únicos. Neste sentido, alguns aplicativos foram considerados *populares*. Mais especificamente, foram reconhecidos como aplicativos *populares*, aqueles aplicativos com 10% ou mais dispositivos únicos. Tais aplicativos são utilizados em muitos dispositivos e precisam ser tratados de maneira diferente por serem muito frequentes e apresentarem muitas formas de uso. O número de aplicativos populares, bem como a porcentagem destes aplicativos em relação ao total de aplicativos em sua respectiva janela também são apresentados pela Tabela B.1. Em média, o número de aplicativos *populares* foi de 33. Desta forma, em relação a Figura B.4, é possível observar que somente os aplicativos *populares* são responsáveis por cerca de 80% do tempo total de uso na primeira janela do FCD *DS01* (linhas lilás). Com base na avaliação do tempo de uso dos aplicativos *populares*, foi sentida a necessidade da divisão de cada um destes atributos em alguns

Janela	Aplicativos					
	Total	**%**	***Mais utilizados***	**%**	***Populares***	**%**
11/12 até 17/12	18.248	30,35%	62	0.34%	14	0,08%
18/12 até 24/12	18.117	30,14%	61	0.34%	14	0,08%
25/12 até 31/12	18.150	30,19%	62	0.34%	15	0,08%
01/01 até 07/01	18.521	30,81%	62	0.33%	14	0,08%
08/01 até 14/01	18.502	30,78%	61	0.33%	14	0,08%
15/01 até 21/01	18.571	30,89%	61	0.33%	14	0,08%
22/01 até 28/01	18.580	30,91%	60	0.32%	14	0,08%
29/01 até 04/02	18.449	30,69%	60	0.33%	14	0,08%
05/02 até 11/02	18.034	30,00%	58	0.32%	14	0,08%
12/02 até 18/02	18.131	30,16%	57	0.31%	14	0,08%
μ	18.330	30,49%	60	0,33%	14	0,08%
σ	203	0,34%	2	0,01%	0	0,00%

Tabela B.2: Quantidade de aplicativos encontrados no FCD *DS02*. O número total e a porcentagem de todos aplicativos, aplicativos *mais utilizados*, e aplicativos *populares* para cada janela. Abaixo a média e o desvio-padrão de cada elemento.

novos atributos. Três técnicas de discretização foram abordadas em ambos FCDs para estudos de caso. Tais técnicas são por *frequência igual* no FCD *DS01* e por *K-Means* e *IP* no FCD *DS02*.

Discretização por Frequência Igual

Para cada aplicativo popular, foi executado a divisão dos N dispositivos existentes, por 10% do número total de dispositivos que fazem uso de tal aplicativo. Como passo seguinte, todos os dispositivos foram divididos, depois de serem ordenados de forma ascendente pelo tempo total de uso do aplicativo, em k intervalos conforme o resultado de tal divisão. Neste sentido, é considerado que o número de intervalos k seja entre 2 e 10. Ao final da discretização, cada aplicativo *popular* foi composto por no máximo 10% dos usuários que utilizaram este aplicativo. Além disso, a discretização melhorou a distribuição de utilização de cada aplicativo. Exemplos de aplicativos *populares* discretizados em k intervalos na primeira janela do FCD *DS01* são apresentados na Tabela B.3.

Na Tabela B.3 são apresentados alguns aplicativos *populares*, o número de dispositivos que utilizam tal aplicativo, a porcentagem destes dispositivos em relação ao total de dispositivos observados e o respectivo número de intervalos em que o aplicativo foi dividido. Por exemplo, o aplicativo *Whatsapp*, foi dividido em 10 intervalos, (*whatsapp*_1_10, *whatsapp*_2_10, ..., *whatsapp*_10_10) onde 1_10 representa dispositivos que apresentam tempo total de uso menor e 10_10 representa

Aplicativos	Dispositivos	%	Intervalos
whatsapp	24.302	98%	10
android.chrome	21.917	88%	9
android.mms	18.938	76%	8
facebook.orca	15.690	63%	7
android.calculator2	12.657	51%	6
google.android.music	11.330	46%	5
android.documentsui	7.772	31%	4
google.android.gm	7.279	29%	3
google.android.apps.docs	4.794	19%	2

Tabela B.3: Algumas discretizações por *frequência igual* com base no tempo de uso de aplicativos *populares* na primeira janela do FCD *DS01*. Quantidade e porcentagem de dispositivos que usam tais aplicativos e o número de intervalo para cada discretização.

dispositivos com com tempo total de uso maior. Tais números demonstram, o intervalo em que o tempo total de uso do aplicativo se insere (primeiro valor, 1, 2, ..., 10), e o total de intervalos que o aplicativo foi discretizado, (segundo valor 10). Os demais usuários, que não estão inseridos em um dos intervalos, apresentaram um tempo total de uso para esse aplicativo igual a zero.

Discretização por Agrupamento

Outra forma de tentar discretizar atributos contínuos é pela execução de algoritmos de Agrupamento. Nesse sentido, foi realizada a discretização dos aplicativos *populares* encontrados no FCD *DS02* antes da definição do limiar mínimo de tempo de uso de cada aplicativo por dispositivo

Neste estudo de caso foi executado o algoritmo *K-Means* e a medida de avaliação *SWC*, variando o parâmetro k de 2 até $\sqrt{N}$. A Tabela B.4 apresenta os resultados dos melhores números de agrupamentos para alguns dos aplicativos *populares* na primeira janela do FCD *DS02*.

Na tabela B.4 são apresentados alguns aplicativos *populares*, o total de dispositivos que fazem uso de tais aplicativos, a porcentagem destes dispositivos em relação a população total do FCD e os melhores resultados obtidos pela medida *SWC*. A coluna *SWC - 1* apresenta o melhor número de grupos encontrado, enquanto a coluna *SWC - 2* apresenta o segundo melhor número de grupos encontrado por tal medida. É possível perceber que o número de grupos sugerido se mantém em 2 ou 3 para praticamente todos aplicativos. Esse mesmo comportamento foi detectado nas demais janelas de tal FCD para todos os aplicativos *populares*. Além disso, o valor obtido pela medida *SWC*, para todos os valores de k se manteve próximo de 0.58, variando sempre na segunda casa decimal.

Aplicativos	Dispositivos	%	*SWC* - 1	*SW* - 2
whatsapp	33.285	98%	2	3
android.chrome	32.831	95%	2	3
android.mms	30.606	87%	2	4
facebook.orca	25.548	74%	2	3
android.calculator2	21.209	61%	2	3
google.android.gm	16.127	47%	2	3
google.android.music	14.978	43%	2	3
android.documentsui	13.683	40%	2	3
google.android.apps.docs	12.055	34%	2	3

Tabela B.4: Algumas discretizações por *K-Means* com base no tempo de uso de aplicativos *populares* na primeira janela do FCD *DS02*. Quantidade e porcentagem de dispositivos que usam tais aplicativos e o número de intervalos para cada discretização.

Após a obtenção dos resultados acima investigou-se a amplitude interquatil (Dawson, 2011) dos valores de tempo de uso de cada aplicativo *popular*. Esta avaliação é calculada com base no cálculo de quartis. Assim, é necessário determinar o intervalo interquatil (IIQ), o qual representa a diferença entre o quartil superior ($Q3$) e o quartil inferior ($Q1$). Dessa forma, dois limites são definidos, limite superior (LS) e limite inferior (LI). Então, o valor de LI é $LI = Q1 - 1,5 \times IIQ$, enquanto o valor de LS é $LS = Q3 + 1,5 \times IIQ$. Assim, os valores que estão abaixo do LI ou acima do LS são considerados *outliers* e são separados do conjunto de valores analisado. Após a remoção dos valores *outliers* foi novamente realizado a execução do algoritmo *K-Means*, conforme procedimento anterior, para todos os aplicativos *populares*. Mesmo com a modificação nos dados que compõem tais aplicativos *populares* os resultados obtidos pela medida de *SWC* continuaram entre 2 e 3. Como os resultados por meio de Agrupamento não foram satisfatórios para a discretização dos aplicativos *populares* resolveu-se investigar outra técnica discretização, por *IP*.

Discretização por IP

A discretização por *IP* a partir da regra 3-4-5 busca dividir os valores de um conjunto de dados em intervalos naturais. Normalmente o conjunto é dividido em 3, 4 ou 5 intervalos, com algumas exceções. Este tipo de discretização foi aplicada nos aplicativos *populares* encontrados no FCD *DS02* após a definição do limiar mínimo de tempo de uso de cada aplicativo por dispositivo. A Tabela B.5 apresenta alguns aplicativos *populares* da primeira janela do FCD abordado e o número de intervalos em que tais aplicativos foram divididos.

Nesta investigação os aplicativos *google.android.gm*, *google.android.apps.docs* e *android.documentsui* não são mais considerados aplicativos populares dado a definição do limiar mínimo de utilização apli-

Aplicativos	Dispositivos	%	Intervalos
whatsapp	33.539	97%	5
android.chrome	22.087	64%	4
android.calculator2	21.209	61%	5
facebook.orca	10.824	31%	5
google.android.music	4.585	13%	4
android.mms	4.437	13%	6

Tabela B.5: Algumas discretizações por *IP* com base no tempo de uso de aplicativos populares na primeira janela do FCD *DS02*. Quantidade e porcentagem de dispositivos que utilizam tais aplicativos e o número de intervalos para cada discretização.

cado. Também é possível perceber que alguns aplicativos populares foram divididos em 6 intervalos. Esse intervalo extra é consequência da necessidade de cobrir o valor máximo de tempo de uso do aplicativo *popular*. Contudo a maioria dos aplicativos *populares* são divididos em 3, 4 ou 5 intervalos. Por fim, os resultados obtidos por meio da discretização por *IP* foram investigados em estudos de caso de outra fases do *framework* proposto (Ver Apêndice C)

APÊNDICE C – ESTUDO DE CASO DA FASE DE ASSOCIAÇÃO

Neste Apêndice são apresentados os estudos de caso relacionados a fase de Associação do *framework f-DOPE*. Para tais estudos, foi aplicado o algoritmo *Apriori* introduzido por Agrawal and Srikant em Agrawal et al. (1994). O primeiro estudo, apresenta o emprego de tal algoritmo após a discretização dos aplicativos populares por *frequência igual*, enquanto que o segundo estudo demonstra a execução do mesmo algoritmo após a discretização de tais aplicativos por *IP*. Para ambos os estudo, este algoritmo recebe como entrada a estrutura de dados previamente gerada na fase de Absorção (ver Apêndice B) e a definição de valores mínimos de *suporte* e *all-confidence*, para a geração dos conjuntos de itens candidatos, e com valores mínimos de *confiança* e *lift*, para a geração das regras de associação. Assim, a estrutura de dados que sumariza o FCD é transformada em um conjuntos de transações. Em resumo, este conjunto de transações indica, por dispositivo, quais atributos (aplicativos) são e quais não são utilizados em seus respectivos intervalos.

Regra de associações após discretização por frequência igual

A Tabela C.1 apresenta alguns exemplos do conjunto de transações gerado na primeira janela do FCD *DS01* após a discretização por *frequência igual* (ver Apêndice B), onde *n* representa o número de dispositivos observados em tal janela.

TID	**Itens**
1	snapchat=3_3, facebook.orca=5_7, whatsapp=9_10,instagram=2_5, ...
2	snapchat=4_3, facebook.orca=4_7, whatsapp=9_10, instagram=1_5, ...
3	snapchat=1_3, facebook.orca=2_7, whatsapp=6_10, instagram=3_5, ...
n	...

Tabela C.1: Exemplos de transações do conjunto de transações de uso de aplicativos da primeira janela do FCD *DS01*.

Inicialmente foi escolhido somente gerar itens candidatos de tamanho dois. Neste sentido, o valor de *suporte mínimo* escolhido neste estudo de caso foi de $0,01$. Além do mais, como o objetivo é encontrar padrões de uso de aplicativos correlacionados, somente itens candidatos com valores de uso devem ser gerados. A Tabela C.2 apresenta alguns conjuntos de itens candidatos de tamanho 2, juntamente com seus respectivos valores de *suporte* e *all-confidence* gerados na primeira janela de tempo com a aplicação da discretização por *frequência igual* no FCD *DS01*.

O número total de conjuntos de itens candidatos produzidos, para cada janela, é apresentado na Tabela C.3. Além disso, a Tabela C.3 exibe o número total destes conjuntos de itens candidatos que possuem valor de *all-confidence* maior que o valor de *all-confidence mínimo* calculado. O cálculo do valor de *all-confidence mínimo* foi realizado pela média dos valores de *all-confidence* de

Itens candidatos	suporte	all-confidence
instagram=5_5, snapchat=3_3	0,100	0,409
facebook.katana=9_9, facebook.orca=7_7	0,090	0,356
⋱	⋱	⋱
chrome=9_9, youtube=8_8	0,0598	0,273

Tabela C.2: Alguns itens candidatos gerados pela primeira etapa da mineração de regras de associação na primeira janela do FCD *DS01* após discretização por *frequência igual*.

Janela	Conjuntos de itens candidatos		
	Total	> all-confidence mínimo	%
28/02 até 05/03	4.237	1.696	40,03%
06/03 até 12/03	4.221	1.686	39,94%
13/03 até 19/03	4.019	1.613	40,13%
20/03 até 26/03	3.918	1.578	40,28%
27/03 até 02/04	3.984	1.573	39,48%
03/04 até 09/04	3.931	1.587	40,37%
10/04 até 16/04	3.884	1.565	40,29%
17/04 até 23/04	3.884	1.590	40,94%
24/04 até 30/04	3.841	1.538	40,04%
01/05 até 07/05	4.015	1.625	40,47%
μ	3.993	1.605	40,20%
σ	130	49	0,6%

Tabela C.3: Resumo da geração dos conjuntos de itens candidatos. O número total de conjuntos de itens inicialmente gerados, o número total e a porcentagem de conjuntos de itens selecionados em cada janela do FCD *DS01*. Abaixo a média e o desvio-padrão de cada elemento.

todos os conjuntos de itens candidatos, uma vez que não existe na literatura uma fórmula concreta para calcular este valor. Visando obter todas as possíveis regras de associação o valore de *suporte mínimo* utilizado continua o mesmo, $0,01$. Neste caso, também deve ser escolhido um valor de *confiança mínima*. Assim, é importante lembrar que a *confiança* mede a probabilidade condicional de Y dado X, tendendo a dar ênfase a regras não correlacionadas (Tan et al., 2006; Han et al., 2011). Neste sentido, o valor de *confiança mínima* escolhido neste estudo foi 0.10. Na Tabela C.4 são apresentadas algumas das regras obtidas neste processo, juntamente com seus respectivos valores de *suporte*, *confiança* e *lift*. É importante recordar que o *lift* diz respeito a probabilidade de

ocorrência de X e Y independentemente um do outro, onde regras com *lift* > 1, são regras onde X e Y são correlacionados positivamente (Tan et al., 2006; Han et al., 2011). Neste sentido, todas as regras obtidas, que não possuam pelo menos *lift* > 1 devem ser desconsideradas. Na Tabela C.5, são apresentados o número total de regras obtidas, o número total de regras após a remoção das *regras redundantes* e o número total de regras que possuem valor de *lift* maior que o valor de *lift mínimo* calculado.

X	Y	suporte	confiança	lift
snapchat=3_3	instagram=5_5	0,100	0,613	2,512
facebook.katana=9_9	facebook.orca=7_7	0,090	0,557	2,199
⋱	⋱	⋱	⋱	⋱
chrome=9_9	youtube=8_8	0,059	0,330	1,544

Tabela C.4: Algumas regras de associação geradas pelo segundo passo da mineração de regras de associação na primeira janela do FCD *DS01*.

Janela	Regras				
	Total	Sem redundantes	%	> min lift	%
28/02 até 05/03	3.315	1.696	51,16%	507	15,29%
06/03 até 12/03	3.267	1.686	51,61%	486	14,88%
13/03 até 19/03	3.098	1.613	52,07%	485	15,66%
20/03 até 26/03	3.055	1.578	51,65%	456	14,93%
27/03 até 02/04	3.036	1.573	51,81%	473	15,58%
03/04 até 09/04	3.036	1.587	52,27%	489	16,11%
10/04 até 16/04	3.011	1.565	51,98%	463	15,38%
17/04 até 23/04	3.065	1.590	51,88%	456	14,88%
24/04 até 30/04	2.975	1.538	51,70%	447	15,03%
01/05 até 07/05	3.170	1.625	51,26%	467	14,73%
μ	3.103	1.605	51,74%	473	15,24%
σ	107	49	0,33%	18	0,42%

Tabela C.5: Resumo do tarefa de mineração de regras de associação. O número total de regras geradas, o número total e a porcentagem de regras após a remoção de regras redundantes e o número total e porcentagem de regras selecionadas para cada janela do FCD *DS01*. Abaixo a média e o desvio-padrão de cada elemento.

O cálculo do valor de *lift mínimo* foi realizado pela média dos valores de *lift* das regras remanescentes após o descarte das regras redundantes. É esperado que o valor de *lift mínimo* calculado seja > 1, caso contrário ele deve ser, obrigatoriamente $= 1$. Em geral, o número de regras obtida foi satisfatório. Contudo, foi investigado a aplicação destes conjuntos de regras nos demais estudos de casos das fases posteriores do *framework*. Esta investigação foi realizada visando avaliar a qualidade de tais regras na obtenção de perfis de uso (Ver Apêndice D).

Regras de associação após discretização por *IP*

Após a definição de um limiar mínimo de utilização de aplicativos, que foi apresentado como estudo de caso no Apêndice B a geração de regras de associação foi um pouco afetada. A geração de itens de tamanho dois fez com que o número de regras obtidas para o FCD *DS02* diminuísse bastante (ex: menos de 300 regras) em relação ao FCD *DS01*. Dado os valores de *suporte mínimo* $0,01$ e *confiança mínima* $0,10$, seria inviável reduzir tais valores, os quais são razoavelmente baixo, o que ajudaria a aumentar o número de regras. Desta maneira, foi aumentado o tamanho máximo das regras a serem geradas. Assim, foi definido 4 como o tamanho máximo para a obtenção de regras a partir do FCD *DS02* com os aplicativos *populares* discretizado por *IP*. O processo para a obtenção de regras de associação seguiu praticamente da mesma maneira. A única variação foi a seleção de regras com *lift* > 1, para comparação. Com a nova definição para o tamanho das regras foram obtidas em média 1.785 regras, como mostra a Tabela C.6.

Com a filtragem pelo *lift* > 1 obteve-se em média 1.424 regras, enquanto com a filtragem pelo *lift* maior que a média de *lift* calculada obteve-se em média 674 regras. Em comparação com o processo executado anteriormente, os resultados apresentados por esta Secção mostram uma quantidade maior de regras obtidas tanto antes quanto depois da filtragem pelo *lift*. As regras obtidas por ambas as formas de filtragem possuem, em muitos casos, regras muito similares dado o aumento no tamanho máximo de tais regras. Estas regras similares apresentam os mesmo conjuntos de itens, em uma ordenação distinta, que apresentam um mesmo valor de *suporte* e com valores de *confiança* e de *lift* distintos, como mostra a Tabela C.7. Dada a repetição de tais regras, buscou-se investigar os conjuntos de itens candidatos que geraram as regras finais obtidas. Assim, os conjuntos de itens candidatos que geraram as regras obtidas pela filtragem com *lift* > 1 são investigados. A captura dos conjuntos de itens candidatos que geram tais regras é utilizada como base para o cálculo de distância entre usuários que é apresentado no estudo de caso no Apêndice D.

Em resumo, com este estudo de caso, foi possível verificar que a redundância de regras ocorre principalmente pelo aumento no tamanho máximo de tais regras como, por exemplo, as apresentadas na Tabela C.7, as quais são geradas a partir de um único conjunto de itens candidatos. Além disso, a ampliação no tamanho das regras beneficiou a identificação de perfis de uso abordado por esta pesquisa.

Janela	Regras				
	Total	**lift > 1**	**%**	**lift > média**	**%**
11/12 até 17/12	1.801	1.434	79,62%	686	21,00%
18/12 até 24/12	1.908	1.512	79,25%	710	22,92%
25/12 até 31/12	1.906	1.532	80,38%	688	22,52%
01/01 até 07/01	1.907	1.535	80,49%	728	23,98%
08/01 até 14/01	1.706	1.367	80,13%	639	21,05%
15/01 até 21/01	1.744	1.386	79,47%	672	22,32%
22/01 até 28/01	1.675	1.330	79,40%	648	21,14%
29/01 até 04/02	1.698	1.357	79,92%	639	21,48%
05/02 até 11/02	1.752	1.395	79,62%	659	20,79%
12/02 até 18/02	1.757	1.396	79,45%	668	21,53%
μ	1785	1424	79.77%	674	21.87%
σ	86	72	0.41%	28	0.98%

Tabela C.6: Resumo do tarefa de mineração de regras de associação. O número total de regras geradas, o número total e a porcentagem de regras após a filtragem por *lift* > 1 e o número total e porcentagem de regras selecionadas após a filtragem por *lift* > que a média de *lift* calculada, para cada janela do FCD *DS02*. Abaixo a média e o desvio-padrão de cada elemento.

LHS	RHS	suporte	confiança	lift
android.music=1_4, snapchat=1_5	instagram=1_4	0,0211	0,532	1,800
instagram=1_4, snapchat=1_5	android.music=1_4	0,0211	0,400	1,650
android.music=1_4, instagram=1_4	snapchat=1_5	0,0211	0,243	2,246

Tabela C.7: Algumas regras similares encontradas após o aumento no tamanho máximo de itens na primeira janela do FCD *DS02*.

APÊNDICE D – ESTUDO DE CASO DA FASE DE CARACTERIZAÇÃO

Neste Apêndice são apresentados os estudos de caso relacionados com a fase de Caracterização do *framework f-DOPE*. Duas abordagens foram realizadas como estudo de caso para esta fase. Na primeira buscou-se formar perfis de regras obtidas na fase de associação com o FCD *DS01*. Assim, a similaridade entre tais regras foi baseada nos dispositivos que dão suporte para tais regras. Na segunda investigou-se a obtenção de perfis de dispositivos. Dessa forma, a similaridade entre dispositivos é baseada nos conjunto de itens finais que geram as regras obtidas após o processo da fase de associação. Ambos conjuntos, de regras e de itens, são oriundos dos estudos de caso do Apêndice C.

Agrupamento de Regras de Associação com Mapeamento de Dispositivos

Para conseguir encontrar a verdadeira relação entre duas regras e definir uma medida de similaridade alguns autores baseiam-se nos dados originais, utilizando medidas que baseiam-se no conjunto de itens que dão suporte a tais regras. Neste estudo de caso, o conjunto de itens são os dispositivos que dão suporte para tais regras. Assim, uma matriz de distâncias foi calculada e estruturada com base em todas as regras restantes oriundas da execução da fase de associação após a discretização por *frequência igual*. A criação de tal matriz foi realizada com base no Índice de *Jaccard* apresentado pelo Capítulo 2, Seção 2.2.1, Equação 2.22, porém com uma modificação (ver Equação D.1. A única mudança é a substituição da subtração (1−) existente no índice original, pela negação de uma função logarítmica ($-log$), o que proporciona um ganho no intervalo de valores possíveis para as distâncias calculadas. A distância é calculada a cada par de regras (ex: $(X \Rightarrow Y)$) e $(X \Rightarrow Y)'$) com base em seus conjuntos de itens, os quais são compostos por dispositivos que dão suporte a tais regras. Assim, o conjunto $I_{(X \Rightarrow Y)}$ pode ou não conter os mesmo dispositivos do conjunto $I_{(X \Rightarrow Y)'}$, pois os dispositivos que utilizam os aplicativos contidos em cada uma das regras podem ser diferente. Em resumo o resultado do cálculo de distância, para duas regras que possuam o mesmo conjunto de dispositivos de suporte, será = 0. Por outro lado, se duas regras possuírem um conjunto diferente de dispositivos de suporte, o resultado vai tender ao infinito.

$$Dist\_J((X \Rightarrow Y), (X \Rightarrow Y)') = -log\left(\frac{\left|I_{(X \Rightarrow Y)} \cap I_{(X \Rightarrow Y)'}\right|}{\left|I_{(X \Rightarrow Y)}\right| + \left|I_{(X \Rightarrow Y)'}\right| - \left|I_{(X \Rightarrow Y)} \cap I_{(X \Rightarrow Y)'}\right|}\right) \quad \text{(D.1)}$$

Após o cálculo e a estruturação da matriz de distância das regras foi realizada a tarefa de Agrupamento. Para este fim, a matriz de distância foi utilizada como entrada para alguns algoritmos testados, os quais são *WARD*, *UPGMA*, *Complete Linkage* e *Single Linkage*. Estes algoritmos são possíveis de serem utilizados quando se possui uma matriz de distâncias pré-computada. Além disso, medidas de avaliação que foram utilizadas são *SWC* e *Elbow method* (Vendramin et al., 2010).

Para este estudo, o algoritmo *Complete Linkage* apresentou os melhores resultados de acordo com as avaliações obtidas pelas medidas de avaliação. Além disso, as medidas *SWC* e *Elbow method* apresentaram resultados sugerindo números similares de grupos em todas as janelas do FCD *DS01*.

Janela	Grupos	Outliers	
		Total	%
28/02 até 05/03	8	403	1,63%
06/03 até 12/03	5	611	2,50%
13/03 até 19/03	5	573	2,36%
20/03 até 26/03	5	622	2,58%
27/03 até 02/04	9	557	2,33%
03/04 até 09/04	4	569	2,39%
10/04 até 16/04	4	599	2,52%
17/04 até 23/04	4	610	2,59%
24/04 até 30/04	5	646	2,76%
01/05 até 07/05	6	633	2,72%
μ	6	582	2,44%
σ	2	66	0,30%

Tabela D.1: Resumo da tarefa de Agrupamento, o número de grupos representando diferentes perfis, o número total e porcentagem de dispositivos *outliers* para cada janela do FCD *DS01*. Abaixo a média e o desvio-padrão de cada elemento.

Após a definição dos grupos cada dispositivo analisado precisou ser mapeado para um dos grupos identificados. Neste sentido, cada dispositivo foi adicionado ao grupo que ele fornece maior porcentagem de suporte em relação a todas as regras agrupadas em tal grupo. Esta definição ocorreu pelo fato de que cada grupo obteve uma quantidade diferente de regras agrupadas. Assim, se tal mapeamento fosse realizado pelo total de regras para as quais um dispositivo apresentasse suporte, este mapeamento seria desvirtuado. É importante notar que alguns dispositivos podem dar suporte somente a regras que foram descartadas durante a fase de Associação (ver Apêndice C). Como alguns conjuntos de itens bem como regras são descartadas ao longo de tal fase ,os dispositivos que dão suporte a tai regras são considerados *outliers*. Na Tabela D.1 também é apresentado o número total e a porcentagem de dispositivos *outliers* para cada janela.

A caracterização dos perfis neste estudo de caso é realizada pela identificação dos aplicativos mais frequentes em cada perfil. A identificação destes principais aplicativos (aplicativos mais frequentes) visou caracterizar um perfil analisando suas principais características. Como cada perfil é formado pelo agrupamento de padrões similares de uso de aplicativos, as principais características

de um grupo será diferente dos demais. Para ajudar na visualização deste tipo de caracterização, foi utilizado o recurso de *word clouds* (nuvens de palavras) (Figura D.1). Está técnica visa ajudar a visualizar e entender as palavras mais frequentemente utilizadas. Neste estudo, as palavras de tal nuvem são os nomes dos aplicativos mais utilizados pelos dispositivos em um perfil obtido. Neste sentido, quanto mais utilizado for um determinado aplicativo mais visível este será na nuvem, representando as características de um determinado perfil.

Figura D.1: Representação visual dos aplicativos mais frequentes, por meio de nuvens de palavras de um dos agrupamentos gerados na primeira janela do FCD *DS01*.

A Figura D.1 mostra uma nuvem de palavras obtida para um dos perfis encontrados na primeira janela do FCD *DS01*. Esta nuvem apresenta o perfil de dispositivos mapeados que utilizam, em grande quantidade, os aplicativos *Instagram* e *Snapchat*. Neste caso, os intervalos em que estes aplicativos foram discretizados são os maiores possíveis (5_5 e 3_3) neste perfil.

Agrupamento de Dispositivos por meio de Conjuntos de Itens

Neste estudo de caso, uma nova forma de calcular uma matriz de distâncias é investigada. Tal abordagem se baseia nos conjuntos de itens que geraram as regras finais obtidas no estudo de caso com discretização por *IP* (ver Apêndice C). Para este estudo é utilizada a mesma função apresentado pelo Capítulo 2, Seção 2.2.1, Equação 2.22, com a mesma modificação apresentada na Equação 2.22 no estudo de caso do Apêndice C. Contudo, desta vez visou-se encontrar a verdadeira relação entre dois dispositivos i e $i^{'}$ ao invés de duas regras. Desta forma, a função se baseia nos conjuntos de itens para quais os dispositivos apresentam suporte. Assim, a distância é calculada a cada par de dispositivos com base em seus conjuntos de itens (ex: P_i e $P_{i'}$), os quais são conjuntos de itens para os quais tais dispositivos são suporte. Da mesma maneira, o conjunto P_i pode ou não conter os mesmo conjuntos de itens de $P_{i'}$. Enquanto que, alguns dispositivos não apresentam suporte para os conjuntos de itens finais. Esta falta de suporte mostra, por exemplo, que tais dispositivos utilizam poucos aplicativos ou usam aplicativos que não são importantes para a geração de padrões de uso. Estes dispositivos representam menos de $0,01\%$ dos dispositivos analisados. Assim, tais dispositivos são considerados *outliers*.

Após o cálculo e a estruturação da matriz de distância entre dispositivos, foi realizada a tarefa de Agrupamento. Para tal tarefa os algoritmos *WARD*, *UPGMA*, *Complete Linkage* e *Single*

Linkage e as medidas de avaliação *SWC* e *Elbow method* (Vendramin et al., 2010) foram novamente utilizadas. Com base nas avaliações obtidas pelas medidas de avaliação o algoritmo *Complete Linkage* apresentou os piores resultado, enquanto o algoritmo *Ward* apresentou os melhores resultados. No entanto os algoritmos *UPGMA* e *Single Linkage*, dado os valores apresentados pelas medidas de avaliação, tiveram um bom desempenho. Contudo, após a avaliação da distribuição dos grupos com tais algoritmos foi possível perceber, que apesar dos resultados das medidas de avaliação, os dispositivos foram distribuídos de uma maneira indesejada. A Tabela D.2 apresenta a distribuição dos dispositivos para os algoritmos *Single Linkage*, *UPGMA* e *WARD* em seus respectivos números de perfis.

Algoritmos	Grupos	G1	G2	G3	G4	G5	G6	G7	G8
Single	6	34.273	1	1	1	1	1	-	-
UPGMA	8	2	2	4	13	5	2	34.249	1
WARD	5	3.052	6.481	6.423	3.466	14.856	-	-	-

Tabela D.2: Distribuição dos dispositivos nos perfis obtidos por meio de diferentes algoritmos na primeira janela do FCD *DS02*.

É possível observar que para os algoritmos *Single Linkage* e *UPGMA* somente um dos perfis (respectivamente G1 e G7) agrupam praticamente todos os dispositivos, enquanto os demais perfis possuem menos de 15 dispositivos. As medidas de avaliação até então utilizadas são baseadas nas distâncias entre os objetos. Contudo, pode ser necessário uma avaliação da distribuição destes usuários em cada perfis, uma vez que distribuições como as apresentadas pelos algoritmos *Single Linkage* e *UPGMA* podem não representar perfis de uso adequados. Além disso, outras medidas podem ser aplicadas.

Por fim, os resultados apresentados por todo o processo com a aplicação do algoritmo *WARD* apresentaram melhorias em relação ao processo apresentado ao final do estudo de caso do Apêndice C. Em resumo a técnica de discretização por *IP* e o uso de conjuntos de itens, ao invés das regras, melhora os resultados da tarefa de Agrupamento. Além disso, o cálculo da matriz de distância por dispositivos permite diferenciar os resultados da tarefa de Agrupamento.

APÊNDICE E – ESTUDO DE CASO DA FASE DE MONITORAMENTO

Neste Apêndice é apresentado o estudo de caso relacionado a fase de Monitoramento do *framework f-DOPE*. Para esta fase foi realizado um estudo de caso com o FCD *DS01* dando continuidade aos resultados do estudo de caso onde perfis de usuários foram formados pelo agrupamento de regras de associações em combinação com o mapeamento destes usuários para tais grupos (ver Apêndice D). Contudo, as abordagens aqui testadas podem ser aplicadas aos resultados obtidos pelo estudo de caso que buscou perfis de uso pelo agrupamento de dispositivos (ver Apêndice D).

Visando distinguir o comportamento dos dispositivos, foi buscado entender como estes dispositivos mudam de perfis, baseando-se nas mudanças de conceitos que podem ocorrer ao longo do tempo. É importante notar que um grupo, em janelas distintas, pode representar um único perfil, mesmo quando seu rótulo for diferente. Neste sentido, em uma única janela cada rótulo representa um perfil diferente, enquanto rótulos diferentes podem representar o mesmo perfil em diferentes janelas. Isto é possível pelo fato de que o algoritmo de agrupamento é o responsável por rotular cada um dos grupos formados. Além disso, quando um dispositivo não envia eventos de uso de aplicativos, este não é agrupado (–). A Tabela E.1 apresenta alguns dispositivos e a sequência de grupos em que tais dispositivos foram mapeados[1].

IMEI	28/02	06/03	13/03	20/03	27/03	03/04	10/04	17/04	24/04	01/05
iea15...	5	1	3	0	4	2	1	2	0	4
ie7bd...	0	3	1	2	4	0	0	0	3	2
⋱	⋱	⋱	⋱	⋱	⋱	⋱	⋱	⋱	⋱	
ie892...	3	2	0	-	-	-	-	-	-	-

Tabela E.1: Alguns dispositivos e o rótulo do grupo para o qual cada um foi mapeado no decorrer das janelas do FCD *DS01*.

Após a obtenção dos perfis é possível investigar as mudanças de conceitos de uma janela para outra. Nesse sentido, foi utilizado o conceito de representação de perfis por *enumeração* apresentado no Capítulo 3, Seção 3.3. Para tal abordagem, um perfil é monitorado por meio da verificação da quantidade de objetos existente em perfis de diferentes janelas (Spiliopoulou et al., 2006; Oliveira e Gama, 2010c). Neste caso, foram observados os dispositivos pertencentes a cada perfil buscando investigar onde tais dispositivos foram agrupados em janelas posteriores. Para este estudo de caso, utilizou-se a verificação da continuidade ou não dos perfis, sem averiguar, divisões e uniões. Assim, para um perfil em uma janela posterior ser o mesmo de uma janela anterior, optou-se por considerar inicialmente que este grupo deve conter mais da metade ($> 50\%$) dos objetos que estavam agrupados no mesmo perfil anterior. Após este teste foi possível identificar a existência de

[1]São indicados, como cabeçalho, os datas de início de cada janela do FCD *DS01*

10 perfis distintos ao longo de todas as janelas do FCD *DS01*. Tais perfis são apresentados pela Tabela E.2.

Principais aplicativos	**nº de Janelas**
Update, Searchbox, Whatsapp_*(pouco_uso)*[2]	2
Snapchat, Instagram	10
Facebook.orca, Facebook.katana, Whatsapp_*(muito_uso)*[3]	7
Vending, UI, Settings	10
Calculator, Gmail, Drive	10
Google Play, Vending, Services	2
Mms, Contacts, Dialer	1
Transalate, Youtube, Twitter	1
Gallery, Music, Camera	1
Contacts, Photos, Mms	1

Tabela E.2: Principais aplicativos para cada um dos dez perfis identificados e o número de janelas em que estes perfis foram identificados ao longo do FCD *DS01*.

Na Tabela E.2, também são apresentados os aplicativos *mais utilizados* por cada perfil. Foram desprezados os intervalos em que estes aplicativos *populares* foram divididos pois o número de intervalos varia, em alguns casos, de acordo com a janela analisada. Esta variação existe mas é pequena e somente o aplicativo *Whatsapp* apresenta diferentes tipos de intervalos, com menor tempo de uso (ex: 1_10 e 2_10) e maior tempo de uso (ex: 9_10 e 10_10). Além disso, a Tabela E.2 mostra também o número de janelas em que estes perfis foram identificados. Por fim, além da verificação por enumeração, a frequência de uso dos aplicativos e as nuvens de palavras mostraram existência de alguns perfis em todas as janelas enquanto outros surgem e desaparecem ao longo do tempo.

[2]Representa intervalos da discretização com pouco tempo total de uso (ex: 1_10 e 2_10)
[3]Representa intervalos da discretização com grande tempo total de uso (ex: 9_10 e 10_10)

APÊNDICE F – ESTUDO DE CASO DA FASE DE SEGMENTAÇÃO

Neste Apêndice é apresentado o estudo de caso relacionado a fase de Segmentação do *framework f-DOPE*. Para este estudo é necessário levar em consideração os resultados do estudo de caso da fase de Monitoramento (ver Apêndice E). Os resultados de tal estudo mostraram que existe uma diferença de uso de aplicativos entre usuários e também entre diferentes períodos de tempo. Neste sentido, o monitoramento da evolução dos perfis e também do comportamento de usuários permite apontar para diferentes tipos de comportamentos, com ou sem ocorrência de mudança conceitos ao longo do tempo. No caso de mudança no comportamento de um usuário, tal mudança pode sugerir, por exemplo, um usuário candidato a *churn* ou uma insatisfação. Tendo em vista os diferentes perfis, mudança de conceito e as mudanças de perfis dos usuários são investigados os *ciclos comportamentais*. Neste sentido, foram produzidos diferentes *ciclos comportamentais* por meio da investigação das mudanças de comportamentos dos usuários de acordo com seu perfis e as mudanças de conceitos detectadas. Alguns ciclos foram compostos por muitos usuários e outros por poucos usuários. Ciclos com poucos usuários (ex: 2) apresentam comportamentos de usuários que mudam seu comportamento constantemente. Assim foi escolhido selecionar ciclos representados por um número mínimo de usuários, pois ciclos com poucos usuários indicam comportamentos de ruídos em uma grande população. Mais precisamente, ciclos representados por mais de três usuários foram selecionados. Ciclos com até três usuários não são usuais indicando comportamentos incertos e impossibilitando a obtenção de conhecimento. Assim, ao total, foram encontrados 706 *ciclos comportamentais*, os quais são representados por mais de três usuários. A Tabela F.1 mostra alguns dos ciclos monitorados durante este processo.

28/02	06/03	13/03	20/03	27/03	03/04	10/04	17/04	24/04	01/05	Dispositivos
-	L	L	L	L	L	L	L	L	L	758
-	L	L	L	M	L	L	L	L	L	423
-	L	L	L	M	L	L	L	M	M	252

Tabela F.1: Três dos *ciclos comportamentais* encontrados no FCD *DS01*.

O *ciclo comportamental* mais frequente monitorado, foi o ciclo onde, ao passar das janelas somente ações leais (L) foram identificadas. Tal ciclo é representado por 748 usuários que sempre são leais conforme a evolução de tempo. Este ciclo não sugere mudanças no comportamento dos usuários e apresenta uma grande quantidade de usuários, não representando um comportamento infrequente ou de risco.

Por outro lado, o segundo ciclo mais frequente, representado por 423 usuários, é um ciclo onde usuários são leais por algumas janelas e mudam de comportamento (M) exatamente na metade do tempo analisado. Após esta mudança de comportamento, estes usuários mantém ações leais ao novo perfil, continuando assim até o final da última janela. Este ciclo sugere uma única mudança no comportamento dos usuário. Tal mudança pode indicar alguma dificuldade de tais usuários. Assim,

o conhecimento obtido com esta mudança, pode ser ou não um indicativo de um comportamento a ser investigado.

Outro comportamento detectado é um comportamento com 252 usuários. Este comportamento apresenta 3 mudanças, uma delas no meio do tempo e outras duas no final da avaliação. A mudança ocorrida na metade do período ocorre justamente na mesma janela que o comportamento anterior, contudo a ocorrência de mais duas mudanças indica uma alta probabilidade para que tal comportamento seja investigado. A ocorrência de três mudanças no comportamento em um conjunto relativamente pequeno de usuários pode indicar para uma situação que deve ser investigada. Estes mudanças podem representar, uma insatisfação ou problema, o que pode levar usuários a abandonarem a marca.

Neste pesquisa, investiga-se as mudanças de comportamentos de usuários as quais podem ajudar na identificação de comportamentos de risco. Estas mudanças devem ser aprendidas e utilizadas para segmentar comportamento infrequentes visando minimizar impactos gerados por tais situações de risco.

Com o objetivo de observar o comportamento dos usuários do FCD *DS01* foi desenvolvido um sistema de monitoramento de comportamento de usuários para o contexto abordado (Figura F.1). Tal sistema foi desenvolvido com o uso da biblioteca *3D.js* que utiliza a linguagem de programação *javascript*. Neste sistema, os 10 diferentes perfis caracterizados, bem como o rótulo de *Outlier* são visualizados ao redor da área esférica da imagem. Além destes, o rótulo Ausente (*Missing*) é visualizado no centro da área esférica da imagem. Os comportamentos, que venham a ser segmentados por apresentar *ciclos comportamentais* a serem investigados, deverão ser ressaltados neste mesmo sistema em uma versão a ser desenvolvida. Estes comportamentos podem determinar perfis, mudanças de comportamentos e *ciclos comportamentais*, que favorecem o abandono de usuários, devendo ser investigados.

Na Figura F.1 uma amostra de 3.000 usuários é analisada e o FCD *DS01* está na quinta janela (*week*). Nesta janela, 8 perfis de usuários (Ver Tabela D.1) são formados. Além disso, o agrupamento central representa usuários ausentes em tal janela, enquanto o grupo de cor rosa próximo a nuvem de palavras representa usuários *outliers*. Tais agrupamentos são representados por diferentes cores e representam diferentes perfis, enquanto cada *pixel* representa um usuário monitorado. Estes usuários, com o passar das semanas, se movimentam em diferentes perfis. Ao colocar o *mouse* em cima de um perfil uma nuvem de palavras que representa tal perfil surge ao lado esquerdo. Esta nuvem de palavras mostra os principais aplicativos utilizados e também a porcentagem de usuários do perfil que utilizam tais aplicativos. O gráfico que se encontra abaixo da janela indica, no eixo x, a janela analisada e, no eixo y, o percentual de usuários em cada agrupamento. Os agrupamento são as linhas, as quais possuem as mesmas cores presentes no círculo principal.

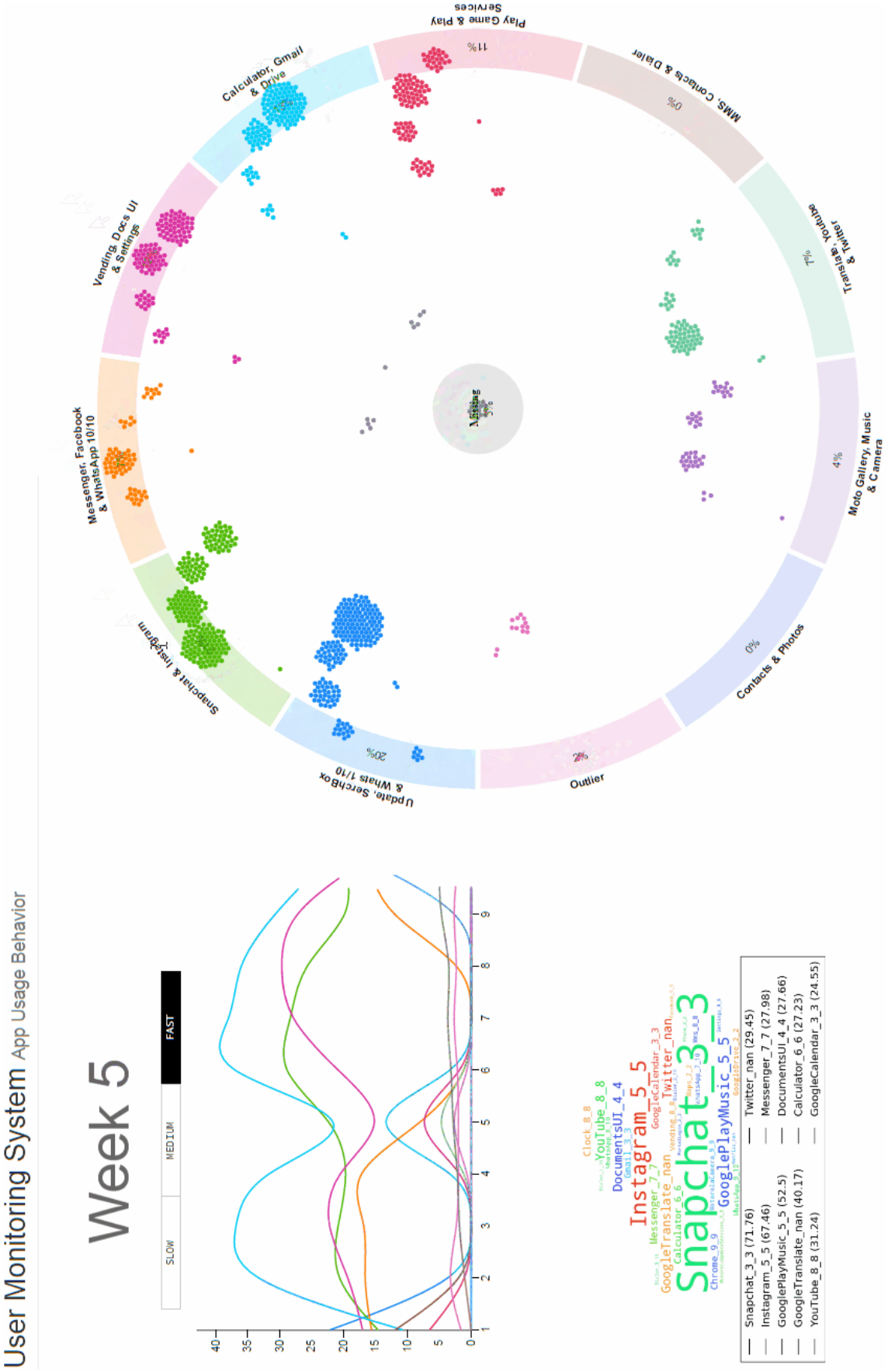


Figura F.1: Captura de tela do sistema desenvolvido para monitorar os comportamentos dos usuários ao longo do tempo.

APÊNDICE G – RESULTADOS DAS PREDIÇÕES COM *F-DOPE* DADAS TODAS COMBINAÇÕES DOS LIMIARES.

Limiares		*Distribuição*		*Sensibilidade*		*FPR*		*Precisão*		*F-Measure*		*Acurácia*	*Sensibilidade*	*FPR*	*Precisão*	*F-Measure*
τ_{match}	τ_{split}	$\neg L$	L	$\neg L$	L	$\neg L$	L	$\neg L$	L	$\neg L$	L		*Ponderada*	*Ponderada*	*Ponderada*	*Ponderada*
0,50	0,10	20,67%	79,33%	0,191	0,994	0,006	0,809	0,896	0,825	0,315	0,902	0,828	0,828	0,643	0,840	0,781
0,50	0,15	28,73%	71,27%	0,148	0,988	0,012	0,852	0,837	0,742	0,251	0,848	0,747	0,747	0,611	0,770	0,677
0,50	0,20	52,45%	47,55%	0,608	0,544	0,456	0,392	0,595	0,557	0,601	0,551	0,578	0,578	0,426	0,577	0,577
0,50	0,25	65,22%	34,78%	0,811	0,453	0,547	0,189	0,736	0,562	0,772	0,502	0,687	0,687	0,422	0,675	0,678
0,50	0,30	89,11%	10,89%	0,999	0,001	0,999	0,001	0,891	0,950	0,942	0,002	0,890	0,890	0,890	0,804	0,840
0,50	0,35	89,11%	10,89%	0,999	0,001	0,999	0,001	0,891	0,118	0,942	0,002	0,890	0,890	0,890	0,807	0,840
0,50	0,40	89,11%	10,89%	1,000	0,000	1,000	0,000	0,891	0,100	0,942	0,001	0,891	0,891	0,891	0,805	0,840
0,55	0,10	11,68%	88,32%	0,333	0,996	0,004	0,667	0,915	0,919	0,489	0,956	0,919	0,919	0,589	0,918	0,901
0,55	0,15	39,58%	60,42%	0,194	0,930	0,070	0,806	0,645	0,638	0,298	0,757	0,639	0,639	0,515	0,641	0,575
0,55	0,20	63,35%	36,65%	0,872	0,280	0,720	0,128	0,677	0,559	0,762	0,373	0,655	0,655	0,503	0,633	0,620
0,55	0,25	76,11%	23,89%	0,811	0,453	0,547	0,189	0,736	0,562	0,772	0,502	0,687	0,687	0,442	0,675	0,687
0,55	0,30	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,55	0,35	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,55	0,40	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,60	0,10	11,68%	88,32%	0,333	0,996	0,004	0,667	0,920	0,919	0,489	0,956	0,919	0,919	0,590	0,919	0,901
0,60	0,15	39,58%	60,42%	0,253	0,900	0,100	0,747	0,624	0,648	0,360	0,753	0,644	0,644	0,491	0,638	0,598
0,60	0,20	63,35%	36,65%	0,878	0,274	0,726	0,122	0,676	0,565	0,764	0,369	0,657	0,657	0,505	0,636	0,619
0,60	0,25	76,11%	23,89%	0,985	0,036	0,964	0,015	0,765	0,419	0,861	0,066	0,758	0,758	0,738	0,682	0,671
0,60	0,30	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,60	0,35	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,60	0,40	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,65	0,10	11,68%	88,32%	0,333	0,996	0,004	0,667	0,920	0,919	0,489	0,956	0,919	0,919	0,590	0,919	0,901
0,65	0,15	39,58%	60,42%	0,187	0,942	0,058	0,813	0,678	0,639	0,239	0,761	0,643	0,643	0,514	0,654	0,576
0,65	0,20	87,23%	12,77%	0,997	0,003	0,997	0,003	0,872	0,127	0,931	0,005	0,870	0,870	0,870	0,777	0,813
0,65	0,25	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,65	0,30	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,65	0,35	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,65	0,40	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,70	0,10	11,68%	88,32%	0,331	0,996	0,004	0,669	0,917	0,918	0,487	0,956	0,918	0,918	0,591	0,918	0,901
0,70	0,15	39,58%	60,42%	0,181	0,944	0,056	0,819	0,677	0,637	0,285	0,761	0,642	0,642	0,517	0,653	0,573
0,70	0,20	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,70	0,25	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,70	0,30	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,70	0,35	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,70	0,40	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,75	0,10	11,68%	88,32%	0,333	0,995	0,005	0,667	0,903	0,919	0,487	0,955	0,918	0,918	0,590	0,917	0,901
0,75	0,15	39,58%	60,42%	0,163	0,950	0,050	0,837	0,682	0,643	0,263	0,761	0,639	0,639	0,525	0,653	0,654
0,75	0,20	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,75	0,25	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,75	0,30	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,75	0,35	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,75	0,40	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,80	0,10	11,68%	88,32%	0,333	0,995	0,005	0,667	0,903	0,919	0,486	0,955	0,918	0,918	0,590	0,917	0,901
0,80	0,15	44,63%	55,37%	0,376	0,779	0,221	0,624	0,578	0,607	0,376	0,779	0,599	0,599	0,445	0,594	0,581
0,80	0,20	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,80	0,25	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,80	0,30	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,80	0,35	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,80	0,40	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,85	0,10	11,68%	88,32%	0,331	0,996	0,004	0,669	0,913	0,918	0,486	0,956	0,918	0,918	0,591	0,918	0,901
0,85	0,15	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,85	0,20	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,85	0,25	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,85	0,30	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,85	0,35	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,85	0,40	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,90	0,10	11,68%	88,32%	0,332	0,995	0,005	0,668	0,898	0,918	0,485	0,955	0,918	0,918	0,591	0,916	0,900
0,90	0,15	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,90	0,20	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,90	0,25	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,90	0,30	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,90	0,35	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,90	0,40	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												

Tabela G.1: Todos resultados das medidas de avaliação após execução do algoritmo de classificação nos *ciclos comportamentais* obtidos pelo *f-DOPE* dado as variações de τ_{match} e τ_{split}.

APÊNDICE H – RESULTADOS DAS PREDIÇÕES COM *X-MEANS* DADAS TODAS COMBINAÇÕES DOS LIMIARES.

Limiares		*Distribuição*		*Sensibilidade*		*FPR*		*Precisão*		*F-Measure*		*Acurácia*	*Sensibilidade*	*FPR*	*Precisão*	*F-Measure*
τ_{match}	τ_{split}	$\neg L$	L	$\neg L$	L	$\neg L$	L	$\neg L$	L	$\neg L$	L		*Ponderada*	*Ponderada*	*Ponderada*	*Ponderada*
0,50	0,10	29,29%	70,71%	0,190	0,957	0,043	0,810	0,644	0,740	0,293	0,835	0,732	0,732	0,586	0,712	0,676
0,50	0,15	29,29%	70,71%	0,183	0,957	0,043	0,817	0,639	0,739	0,285	0,834	0,730	0,730	0,590	0,710	0,673
0,50	0,20	29,29%	70,71%	0,165	0,968	0,032	0,835	0,680	0,737	0,266	0,837	0,733	0,733	0,600	0,720	0,669
0,50	0,25	32,89%	67,11%	0,251	0,911	0,089	0,749	0,581	0,713	0,351	0,800	0,694	0,694	0,532	0,669	0,652
0,50	0,30	42,85%	57,15%	0,595	0,760	0,240	0,405	0,650	0,715	0,621	0,737	0,689	0,689	0,334	0,687	0,687
0,50	0,35	42,85%	57,15%	0,595	0,760	0,240	0,405	0,650	0,715	0,621	0,737	0,598	0,598	0,334	0,687	0,687
0,50	0,40	42,85%	57,15%	0,580	0,737	0,263	0,420	0,623	0,701	0,601	0,718	0,670	0,670	0,353	0,667	0,668
0,55	0,10	29,29%	70,71%	0,190	0,957	0,043	0,810	0,648	0,741	0,294	0,835	0,733	0,733	0,585	0,714	0,677
0,55	0,15	29,29%	70,71%	0,182	0,957	0,043	0,818	0,636	0,738	0,283	0,834	0,730	0,730	0,591	0,708	0,672
0,55	0,20	29,29%	70,71%	0,160	0,970	0,030	0,840	0,688	0,736	0,260	0,837	0,733	0,733	0,603	0,722	0,668
0,55	0,25	32,89%	67,11%	0,230	0,921	0,079	0,770	0,587	0,709	0,330	0,801	0,693	0,693	0,543	0,669	0,646
0,55	0,30	42,85%	57,15%	0,585	0,770	0,230	0,415	0,656	0,712	0,618	0,740	0,690	0,690	0,336	0,688	0,688
0,55	0,35	42,85%	57,15%	0,558	0,790	0,210	0,442	0,666	0,704	0,607	0,745	0,690	0,690	0,343	0,688	0,686
0,55	0,40	42,85%	57,15%	0,545	0,799	0,201	0,455	0,671	0,701	0,601	0,747	0,690	0,690	0,346	0,688	0,684
0,60	0,10	29,29%	70,71%	0,194	0,956	0,044	0,806	0,645	0,741	0,299	0,835	0,733	0,733	0,583	0,713	0,678
0,60	0,15	29,29%	70,71%	0,177	0,956	0,044	0,823	0,625	0,737	0,276	0,832	0,728	0,728	0,595	0,704	0,670
0,60	0,20	29,29%	70,71%	0,168	0,961	0,039	0,832	0,641	0,736	0,266	0,834	0,729	0,729	0,600	0,708	0,667
0,60	0,25	32,89%	67,11%	0,246	0,917	0,083	0,754	0,593	0,713	0,348	0,802	0,696	0,696	0,533	0,673	0,653
0,60	0,30	42,85%	57,15%	0,568	0,792	0,208	0,432	0,672	0,710	0,615	0,749	0,696	0,696	0,336	0,693	0,691
0,60	0,35	42,85%	57,15%	0,556	0,811	0,189	0,444	0,688	0,709	0,615	0,757	0,702	0,702	0,335	0,700	0,696
0,60	0,40	42,85%	57,15%	0,538	0,822	0,178	0,462	0,694	0,703	0,606	0,758	0,700	0,700	0,340	0,699	0,693
0,65	0,10	23,53%	76,47%	0,148	0,986	0,014	0,852	0,762	0,786	0,248	0,875	0,785	0,785	0,652	0,780	0,725
0,65	0,15	23,53%	76,47%	0,144	0,987	0,013	0,856	0,777	0,786	0,244	0,875	0,785	0,785	0,654	0,784	0,724
0,65	0,20	44,23%	55,77%	0,636	0,741	0,259	0,364	0,660	0,719	0,648	0,730	0,694	0,694	0,318	0,693	0,694
0,65	0,25	57,79%	42,21%	0,848	0,657	0,343	0,152	0,772	0,76	0,808	0,705	0,768	0,768	0,262	0,767	0,765
0,65	0,30	62,07%	37,93%	0,880	0,669	0,301	0,120	0,827	0,780	0,853	0,737	0,811	0,811	0,232	0,809	0,809
0,65	0,35	62,07%	37,93%	0,883	0,702	0,298	0,117	0,829	0,786	0,855	0,741	0,814	0,814	0,229	0,813	0,812
0,65	0,40	62,07%	37,93%	0,882	0,710	0,290	0,118	0,833	0,786	0,857	0,746	0,817	0,817	0,225	0,815	0,815
0,70	0,10	22,77%	77,23%	0,156	0,986	0,014	0,844	0,763	0,798	0,259	0,882	0,797	0,797	0,655	0,790	0,740
0,70	0,15	22,77%	77,23%	0,144	0,991	0,009	0,856	0,820	0,797	0,245	0,833	0,798	0,798	0,663	0,802	0,738
0,70	0,20	49,08%	50,92%	0,743	0,698	0,302	0,257	0,704	0,738	0,723	0,718	0,720	0,720	0,729	0,721	0,720
0,70	0,25	62,64%	37,36%	0,831	0,648	0,352	0,169	0,798	0,696	0,814	0,671	0,763	0,763	0,284	0,760	0,761
0,70	0,30	66,92%	33,08%	0,877	0,609	0,391	0,123	0,819	0,711	0,847	0,656	0,789	0,789	0,302	0,783	0,784
0,70	0,35	66,92%	33,08%	0,890	0,586	0,414	0,110	0,813	0,724	0,850	0,648	0,789	0,789	0,314	0,784	0,783
0,70	0,40	66,92%	33,08%	0,890	0,586	0,414	0,110	0,813	0,725	0,750	0,648	0,789	0,789	0,314	0,784	0,783
0,75	0,10	22,77%	77,23%	0,153	0,986	0,014	0,847	0,759	0,798	0,255	0,882	0,796	0,796	0,657	0,789	0,739
0,75	0,15	22,77%	77,23%	0,150	0,990	0,010	0,850	0,821	0,798	0,254	0,884	0,799	0,799	0,659	0,803	0,740
0,75	0,20	49,08%	50,92%	0,706	0,720	0,280	0,294	0,709	0,718	0,707	0,719	0,713	0,713	0,287	0,713	0,713
0,75	0,25	62,64%	37,36%	0,844	0,613	0,387	0,156	0,785	0,700	0,813	0,654	0,757	0,757	0,301	0,753	0,754
0,75	0,30	66,92%	33,08%	0,903	0,545	0,455	0,097	0,800	0,734	0,848	0,626	0,784	0,784	0,337	0,779	0,775
0,75	0,35	66,92%	33,08%	0,921	0,514	0,486	0,079	0,793	0,763	0,852	0,614	0,786	0,786	0,352	0,783	0,773
0,75	0,40	66,92%	33,08%	0,921	0,514	0,486	0,079	0,793	0,763	0,852	0,614	0,786	0,786	0,351	0,783	0,774
0,80	0,10	74,61%	25,39%	0,937	0,311	0,689	0,063	0,800	0,628	0,863	0,416	0,778	0,778	0,530	0,756	0,750
0,80	0,15	76,16%	23,84%	0,934	0,399	0,601	0,066	0,832	0,653	0,880	0,895	0,806	0,806	0,474	0,790	0,788
0,80	0,20	82,16%	17,84%	0,958	0,331	0,669	0,042	0,868	0,632	0,911	0,434	0,846	0,846	0,557	0,826	0,826
0,80	0,25	95,72%	4,28%	0,989	0,386	0,614	0,011	0,973	0,611	0,981	0,474	0,963	0,963	0,588	0,958	0,959
0,80	0,30	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,80	0,35	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,80	0,40	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,85	0,10	88,16%	11,84%	0,982	0,153	0,847	0,018	0,896	0,529	0,937	0,237	0,884	0,884	0,749	0,853	0,854
0,85	0,15	89,72%	10,28%	0,978	0,310	0,690	0,022	0,925	0,615	0,951	0,413	0,909	0,909	0,621	0,893	0,895
0,85	0,20	95,72%	4,28%	0,992	0,299	0,701	0,008	0,969	0,620	0,980	0,404	0,962	0,962	0,671	0,954	0,956
0,85	0,25	95,72%	4,28%	0,991	0,330	0,670	0,009	0,971	0,629	0,981	0,433	0,963	0,963	0,642	0,956	0,957
0,85	0,30	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,85	0,35	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,85	0,40	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,90	0,10	95,72%	4,28%	0,994	0,240	0,760	0,006	0,967	0,655	0,980	0,351	0,962	0,962	0,728	0,954	0,954
0,90	0,15	95,72%	4,28%	0,992	0,296	0,704	0,008	0,969	0,619	0,980	0,400	0,962	0,962	0,674	0,954	0,956
0,90	0,20	95,72%	4,28%	0,992	0,296	0,704	0,008	0,969	0,619	0,980	0,400	0,962	0,962	0,674	0,954	0,956
0,90	0,25	95,72%	4,28%	0,992	0,296	0,704	0,008	0,969	0,619	0,980	0,400	0,962	0,962	0,674	0,954	0,956
0,90	0,30	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,90	0,35	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												
0,90	0,40	100,00%	0,00%	Todos comportamentos são $\neg L$ na última janela												

Tabela H.1: Resultados das medidas de avaliação após execução do algoritmo de classificação nos *ciclos comportamentais* obtidos pelo *X-Means* dado as variações de τ_{match} e τ_{split}.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br