PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE ENGENHARIA
MESTRADO EM ENGENHARIA ELÉTRICA

IDENTIFICAÇÃO DE CAUSAS DE DESLIGAMENTOS NÃO PROGRAMADOS EM REDES DE DISTRIBUIÇÃO

Alex Bernsts Tronchoni

Porto Alegre
Março de 2008

ALEX BERNSTS TRONCHONI

IDENTIFICAÇÃO DE CAUSAS DE DESLIGAMENTOS NÃO PROGRAMADOS EM REDES DE DISTRIBUIÇÃO

Dissertação apresentada como requisito para obtenção do grau de Mestre pelo Programa de Pós–graduação da Faculdade de Engenharia Elétrica da Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Flávio Antonio Becon Lemos, Dr.

Co-Orientador: Prof. Daniel Ferreira Coutinho, Dr.

Porto Alegre

Março de 2008

"IDENTIFICAÇÃO DE CAUSA DE DESLIGAMENTOS NÃO PROGRAMADOS EM REDES DE DISTRIBUIÇÃO"

ALEX BERNSTS TRONCHONI

Esta dissertação foi julgada para a obtenção do título de MESTRE EM ENGENHARIA e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Pontifícia Universidade Católica do Rio Grande do Sul.

Flávio Antonio Becon Lemos, Dr.
Orientador

Rubem Dutra Ribeiro Fagundes, Dr.
Coordenador
Programa de Pós-Graduação em Engenharia Elétrica

Banca Examinadora:

Flávio Antonio Becon Lemos, Dr.
Presidente – PUCRS

Daniel Ferreira Coutinho, Dr.
PUCRS

Luís Alberto Pereira, Dr. Ing.
PUCRS

Vinícius Licks, PhD.
PUCRS

Jacqueline Gisele Rolim, Drª.
UFSC

Agradecimentos

O autor agradece a RGE Rio Grande Energia S. A. e as Centrais Elétricas de Santa Catarina – CELESC pelo suporte financeiro e ao desenvolvimento desta dissertação. Além disso, agradeço o orientador e professor, Dr. Flávio Antonio Becon Lemos, que me deu a oportunidade de fazer um mestrado e crescer profissionalmente.

Resumo da Dissertação apresentada a PUCRS como parte dos requisitos necessários para obtenção do grau de Mestre em Engenharia Elétrica.

Identificação de Causas de Desligamentos Não Programados em Redes de Distribuição

Alex Bernsts Tronchoni

Março de 2008.

Orientador: Flávio Antonio Becon Lemos, Dr.

Área de Concentração: Planejamento e Gestão de Sistemas de Energia.

Palavras-chave: Desligamentos Não-Programáveis, Confiabilidade de Sistemas de Distribuição, Mineração de Dados, Redes Bayesianas, Redes Neurais Artificiais.

Os desligamentos não programados são um dos fatores que mais contribuem para a interrupção do fornecimento de energia e, portanto, na qualidade do serviço prestado. Uma correta identificação das causas que originaram os desligamentos torna-se cada vez mais indispensável para distribuir de forma mais eficaz os investimentos e recursos para a redução de problemas no sistema elétrico, trazendo como conseqüência direta destes investimentos a melhoria dos índices de confiabilidade. Dessa forma, torna-se necessário o desenvolvimento de ferramentas para gerenciamento, análise e diagnóstico de causas de eventos não programados que ocorrem nos sistema de distribuição das empresas. Nesta dissertação são apresentados dois métodos para identificação da causa de desligamentos não programados na rede de distribuição: um modelo probabilístico utilizando Redes Bayesianas e um modelo usando Redes Neurais Artificiais. Inicialmente é apresentada uma conceituação sobre aspectos teóricos fundamentais ao entendimento de Redes Bayesianas e Redes Neurais Artificiais, seguida de uma revisão sobre definições básicas acerca de confiabilidade e causas de desligamentos em sistemas de distribuição. Após, são descritas as etapas realizadas para treinamento e validação dos dois sistemas de identificação da causa de desligamentos não programados. A base de conhecimento utilizada para o aprendizado foi extraída de um banco de dados de eventos fornecido por uma concessionária de energia, cujo processo de extração de conhecimento consistiu em uma série de etapas, incluindo uma de mineração de dados. Esse processo tornou a base de dados mais confiável e adequada resultando em 8888 amostras para a construção, geração dos conjuntos de treinamento e validação dos modelos de Rede Bayesiana e de Rede Neural utilizados. Ambas heurísticas foram validadas através do método da prova bipartida (*split-half method*). O processo de aprendizagem da Rede Bayesiana foi realizado através do algoritmo de maximização da expectância (*Expectation Maximization*), enquanto que para a Rede Neural o algoritmo de treinamento escolhido foi o *Resilient back propagation*, devido as suas características de desempenho e velocidade de convergência.

Abstract of Dissertation presented to PUCRS as a partial fulfillment of the requirements for the Degree of Master in Electrical Engineering.

Forced Outage Cause Identification in Distribution Networks

Alex Bernsts Tronchoni

March 2008.

Advisor: Flávio Antonio Becon Lemos, Dr.

Area of Concentration: Electrical Energy Planning and Management.

Keywords: Forced Outage, Distribution System Reliability, Data Mining, Bayesian Networks, Artificial Neural Networks.

Forced outages are one of the most relevant elements of influence in the energy supply interruption and, thus, in the service quality. A correct identification of the causes that led to an outage become essential, once it provides a better way to allocate resources and investments to reduce problems in the electrical system, and, as a consequence, the improvement of reliability indices. To achieve this goal it is necessary to develop tools for the management, analysis and diagnostic of forced outage causes in the electric distribution system. This dissertation presents two methodologies to identify forced outage causes: a probabilistic model using Bayesian Networks, and an Artificial Neural Networks model. Initially, theoretical concepts and definitions required to understand Bayesian Networks and Artificial Neural Networks are presented, followed by a review on basic definitions of distribution system reliability and forced outage causes in the distribution system. After, are described training and validation steps of both forced outage cause identification methods. The knowledge base used for the network learning process was extracted from an event database provided by an electric utility. The knowledge discovery process comprised several stages, including one of data mining. This process turns the database into a more reliable and appropriate format, resulting in 8888 samples for construction, generation of the training and validation dataset of the proposed Bayesian Network and Neural Network models. Both heuristics were validated through the split-half method. The learning process of the Bayesian Network was done using the Expectation Maximization Algorithm, while for Neural Network was used Resilient back propagation learning algorithm, chosen specially because of its fast convergence and good performance.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AMR – Automatic Meter Reading (leitura automática de medidores)

ANEEL – Agência Nacional de Energia Elétrica

ASAI - Average Service Availability Index (índice de disponibilidade média do serviço)

BNT – Bayes Net Toolbox for Matlab (biblioteca de redes bayesianas para *MATLAB*)

CAIDI - Customer Average Interruption Duration Index (índice de duração médio de interrupção por consumidor)

CAIFI - Customer Average Interruption Frequency Index (índice de freqüência média de interrupção por consumidor)

CEEE – Companhia Estadual de Energia Elétrica

COD - Centro de Operação da Distribuição

CPT – Conditional Probabilities Table (tabela de probabilidades condicionais)

DEC - Duração Equivalente de Interrupção por Unidade Consumidora

DIC - Duração de Interrupção Individual por Unidade Consumidora

DMIC - Duração Máxima de Interrupção Contínua por Unidade Consumidora

EM – Expectation Maximization (E - expectância, M - maximização)

ERIS - Equipment Reliability Information System (sistema de informação para confiabilidade de equipamentos)

FEC - Freqüência Equivalente de Interrupção por Unidade Consumidora

FIC - Freqüência de Interrupção Individual por Unidade Consumidora

GAD – Grafo Acíclico Direcionado

KDD – Knowledge Discovery in Databases (extração de conhecimento em banco de dados)

MATLAB – Matrix Laboratory (software de ferramentas matemáticas e matriciais)

MLE – Maximum Likelihood Estimator (estimador de máxima verossimilhança)

MLP – Multi-Layer Perceptron (rede neural perceptron de múltiplas camadas)

MP – Matriz de Pertinência

PDA – Personal Digital Assistant (assistente pessoal digital)

RB – Redes Bayesianas

RGE – Rio Grande Energia

RNA – Redes Neurais Artificiais

SAIDI - System Average Interruption Duration Index (índice de duração média das interrupções do sistema)

SAIFI - System Average Interruption Frequency Index (índice da freqüência média das interrupções do sistema)

SCADA - Supervisory Control and Data Acquisition (sistema de controle de supervisão e aquisição de dados)

SOM - Self Organizing Maps (Mapas Auto-Organizáveis)

UTM – Universal Transversal Mercator (sistema universal transverso de Mercator)

Sumário

1. Introdução

1.1. O Tema e sua Importância

A reestruturação do setor elétrico brasileiro conduziu a uma mudança nas relações entre as empresas, especialmente para as concessionárias de distribuição de energia, forçando a busca por melhorias que atendessem as novas exigências de qualidade impostas pelo órgão regulamentador, a Agência Nacional de Energia Elétrica (ANEEL). Este órgão surgiu para, entre outras coisas, zelar pela qualidade do serviço prestado ao consumidor [1].

Com o objetivo de adequar-se a esses padrões de qualidade, as empresas de energia elétrica precisam identificar corretamente os fatores que influenciam nos índices de confiabilidade do seu sistema, e desta forma, planejar os investimentos necessários para um contínuo melhoramento do sistema, reduzindo a freqüência e a duração de desligamentos não programados.

O grande número de eventos não programados que uma distribuidora de energia apresenta anualmente influi diretamente no desempenho do seu sistema de distribuição e nos índices de confiabilidade. Isso motivou o trabalho realizado em [2], onde foi elaborado um sistema computacional que proporciona uma metodologia organizada e automática de coleta de dados, garantindo um número mínimo necessário de informações sobre os desligamentos atendidos pela empresa. Essa garantia de coleta de informações acaba promovendo um gradual aumento das possibilidades de uso da base de dados históricos sobre desligamentos, uma vez que dados de maior qualidade constituem uma ferramenta muito útil para guiar processos de inspeção, manutenção e expansão de rede. Em outras palavras, para um bom desempenho de sistemas especialistas que auxiliem no suporte desses processos, é necessária uma base de conhecimento confiável para realizar o aprendizado do sistema.

Atualmente, muitas empresas estão utilizando técnicas de mineração de dados com o objetivo de melhorar a qualidade de suas bases de dados, e conseqüentemente tornar a informação armazenada mais confiável.

1.2. Objetivos

O trabalho de Pretto [2] propôs a coleta de dados no local do evento utilizando computação móvel, técnicas como árvores de decisão e matriz de pertinência para a identificação de causas, deixando em aberto a utilização de outros métodos e suas relações com a qualidade da informação de uma base de dados pré-existente, uma vez que na referência [2] a base de informações foi criada com uma formatação adequada e com as informações necessárias para ser acessada pelo sistema proposto. De modo geral, nas concessionárias de energia é comum a existência de banco de dados com o histórico de desligamentos incompleto, não só pela falta de metodologia na coleta das informações a respeito dos eventos, mas também porque muitas vezes algumas variáveis não podem ser observadas diretamente no local onde aconteceu o evento.

Esta dissertação parte das pesquisas desenvolvidas por Pretto [2] e propõe o desenvolvimento de uma nova metodologia de análise sobre um conjunto de dados sobre desligamentos não programados de energia, podendo ser utilizada para criação de ferramentas para apoio na análise e estudo das principais causas de desligamentos não programados em uma rede de distribuição de energia.

As principais contribuições desta dissertação podem ser sintetizadas como:

- Aprofundamento do estudo sobre identificação de causas de desligamentos não programados em sistemas de distribuição;
- Desenvolvimento de um processo de extração de conhecimento em banco de dados (*KDD – Knowledge Discovery in Databases*) para a construção de uma base de dados de eventos a partir de um conjunto incompleto de informações;
- Utilização de Redes Bayesianas e Redes Neurais Artificiais para identificação da causa de desligamentos não programados na rede de distribuição.

A escolha de uma metodologia que utiliza a técnica de Redes Bayesianas foi decidida após a pesquisa e análise do problema e das fontes de informações disponíveis para criar um sistema de apoio à decisão, o qual fosse capaz de identificar a causa, ou causas, de desligamentos não programados, uma vez que esta técnica permite utilizar algoritmos para o aprendizado a partir de dados incompletos ou incertos, e ser sintonizado com o conhecimento de um especialista.

As Redes Neurais Artificiais têm a capacidade de aprender com base em exemplos e a partir disso realizar classificações e generalizações das classes de dados presentes no domínio de treinamento do sistema. Assim como nas Redes Bayesianas, esse método também utiliza algoritmos que permitem o treinamento do sistema a partir de dados incompletos ou incertos. Essas características das RNA foram determinantes para utilização da técnica na metodologia proposta de identificação de causas.

Dessa forma, os métodos propostos nesta dissertação permitem avançar nas pesquisas sobre o diagnóstico de causas através de múltiplas fontes de informação, tratando com a incerteza de forma mais eficiente. Esta proposta tem características herdadas de metodologias já utilizadas na empresa, sendo que o desenvolvimento do sistema tentou manter o máximo possível das características dos processos.

1.3. Caracterização do Problema

A identificação correta das causas de um desligamento não programado está diretamente relacionada à quantidade e à qualidade da informação adquirida no local do evento, das condições de entorno que cercam o evento e da identificação correta das atividades e dos fatos que poderiam indicar uma causa de falta de energia não programada [3]. Muitas vezes a atribuição de uma causa de desligamento está associada a fatos locais aparentes, mas que na realidade podem ser a conseqüência ou causa secundária, e terem sido desencadeados por uma falta primária não detectável na inspeção local durante o restabelecimento de energia. Outras vezes, por não haver uma causa aparente, costuma ser indicado como falta transitória o motivo do desligamento.

Um exemplo é o desligamento por curto-circuito. Ao restabelecer o fornecimento de energia após detectar o curto-circuito que originou a atuação da proteção, o eletricista pode anotar na sua planilha como causa a atuação da proteção

e/ou o curto-circuito. Na realidade, o curto-circuito e a atuação da proteção são conseqüências de algum evento que ocorreu na rede. A correta identificação da causa raiz ou primária é tão ou mais importante, pois muitas vezes permite identificar um conjunto de causas que estão originando a degradação da rede e por conseqüência os indicadores de continuidade. Um exemplo disto é a identificação do padrão de construção e manutenção da rede, como por exemplo, postes desalinhados, postes podres, estais frouxos, condutores com problemas de tracionamento, etc, que pode de fato estar contribuindo para os desligamentos, bem como a vegetação presente no trecho de rede envolvido.

Dessa forma, é importante criar mecanismos para que a indicação visual, direta e conclusiva da causa de desligamento apontada pelo eletricista seja apoiada por um conjunto de dados coletados no local do evento e no seu entorno que comprove através de inferências lógicas e estatísticas a causa apontada, e remova qualquer subjetividade ou informação incompleta, do tipo defeito transitório ou causa desconhecida, como a origem de um evento de desligamento não programado.

As referências [2]-[6] apresentam um sistema para a coleta e análise de dados de eventos não programados utilizando computação móvel, para posterior análise utilizando técnicas de inteligência computacional como lógica *fuzzy* e árvores de decisão [2].

Entretanto, ainda não é prática corrente nas empresas dispor de sistemas automatizados de coleta, armazenagem e avaliação em campo dos dados do evento, o que permitiria adicionar quantidade e qualidade na informação para posterior análise. Dessa forma, a aplicação de técnicas de identificação e análise de eventos baseadas em dados incompletos e incertos, que possam de alguma forma ser ajustadas por especialistas devem ser utilizadas, uma vez que em modelos que representam problemas do mundo real, muitas vezes é necessário levar em consideração a incerteza ou aleatoriedade das variáveis envolvidas. Uma técnica muito utilizada e adequada para este tipo de problema é a de Redes Bayesianas [7]-[12], uma vez que são apropriadas para lidar com manipulação de dados sob incerteza e passíveis de ajustes por especialistas, portanto, adequadas para tratar o tipo de problema estudado nesta dissertação.

Para suprir essa demanda foi criado um sistema utilizando Redes Bayesianas que, com base num histórico de eventos de desligamentos, que pode ser treinado para

diagnosticar as possíveis causas de um corte no fornecimento de energia. Neste trabalho, uma das abordagens o problema de identificação da causa de desligamentos não-programados será utilizando um modelo probabilístico representado através de uma Rede Bayesiana. Adicionalmente, em função do tipo de problema, foi realizada uma pesquisa sobre RNA. Assim, uma RNA do tipo MLP foi implementada com o objetivo de validar o modelo feito através de Redes Bayesianas e ao mesmo tempo para permitir uma comparação entre as duas heurísticas.

Inicialmente, pretendia-se treinar a Rede Bayesiana e a RNA com a base de dados gerada pelo sistema de coleta e identificação de causas desenvolvido em [2]. Entretanto a quantidade de dados que foram coletados no projeto piloto não possui um conjunto suficiente de amostras. Para contornar esse problema foi realizado um processo de extração de conhecimento em banco de dados (*KDD*), em um histórico de desligamentos disponível no banco de dados. Assim, foi possível extrair conhecimento para o treinamento da rede a partir de uma base de dados inconsistente e sem padronização. O que inicialmente parecia um problema acabou tornando-se um desafio e adicionou uma nova pesquisa na dissertação. Essa fase inicial de mineração de dados permitiu o aproveitamento dos dados que originalmente seriam descartados, já que anteriormente ao trabalho de Pretto em [2] não havia na empresa um sistema de coleta e tratamento de dados apropriado para o treinamento de sistemas especialistas.

Atualmente, em grande parte das empresas, o processo de coleta de dados ainda é feito com total autonomia do eletricista, respeitando somente uma lista de possíveis causas, de onde o mesmo deve escolher a que melhor se encaixa no evento que está atendendo. O eletricista passa para o operador do Centro de Operação da Distribuição (COD) o diagnóstico realizado e esse cadastra a causa utilizando a relação de códigos da lista. Apesar desses esforços para a melhoria do desempenho da rede dessas empresas, não existe uma metodologia que retire do eletricista a tarefa de diagnosticar a causa do desligamento. O trabalho realizado nessa dissertação se propõe a isso, criar um sistema que através das informações coletadas sobre o evento desligamento seja possível identificar a causa do mesmo com o auxílio de um sistema especialista.

1.4. Revisão Bibliográfica

Uma revisão bibliográfica criteriosa sobre os temas abordados, e a busca pelo estado da arte da pesquisa, são fundamentais para estabelecer os limites da pesquisa e apresentar as contribuições da dissertação. Nesta seção é apresentada uma revisão bibliográfica sobre os principais tópicos envolvidos na dissertação: aspectos de confiabilidade de sistemas de distribuição de energia elétrica com ênfase na identificação de causas de desligamentos não programados, utilização de técnicas de inteligência computacional com destaque para as Redes Bayesianas e Redes Neurais Artificiais abordando seus aspectos teóricos e práticos na determinação de causas de desligamentos, assim como aplicação de mineração de dados para criação de uma base de dados consistente a partir de uma base de informações incompletas.

Em [2] o autor expõe uma abordagem para a coleta e tratamento de informações sobre desligamentos não programados na rede de energia elétrica. Foi desenvolvido um sistema que, além de melhorar a qualidade dos dados obtidos em campo pelos eletricistas, também indica a provável causa de um desligamento não programado. Propôs-se que a coleta dos dados fosse feita através de um questionário implementado em um *Personal Digital Assistant* (*PDA*). Nesse dispositivo, o eletricista deve escolher a resposta mais apropriada para traduzir condições do elemento do sistema de distribuição que falhou e as variáveis temporais relativas às condições climáticas da hora da ocorrência do evento. A identificação da causa do desligamento é feita através de um algoritmo implementado em um aplicativo que analisa as informações colhidas em campo com o computador portátil. No trabalho foram apresentadas duas abordagens para a apropriação de conhecimento de especialistas e identificação de causas, sendo a primeira através de uma tabela relacional entre as causas do desligamento e os indícios coletados em campo, chamada de Matriz de Pertinência (MP), e a segunda baseada na teoria de árvores de decisão.

Um sistema de coleta de dados de interrupções não programadas (*forced-outages*) em equipamentos de distribuição é descrito em [13]. Esse sistema é o terceiro estágio do desenvolvimento de um sistema de informação para confiabilidade de equipamentos (*ERIS*) adotado pela *Canadian Electric Utilities*. No processo de desenvolvimento do sistema, uma série de premissas tiveram que ser adotadas a fim de fornecer as informações adequadas para alimentar o banco de dados do sistema. Entre

elas podem-se destacar: qual equipamento deve ser identificado, que dados devem ser coletados e de que forma armazená-los. O primeiro estágio lida com os equipamentos de geração de energia e foi implantado em 1977. O segundo estágio lida com os equipamentos de transmissão e foi posto em funcionamento em 1978. No terceiro estágio, foram classificados os componentes e as principais causas de falha na rede de distribuição. Isso permitiu a construção de forma sistemática de uma base de dados de desligamentos para ser utilizada no sistema *ERIS*, que torna o gerenciamento de recursos e equipamentos mais eficiente, na geração, transmissão e distribuição de sistemas de energia.

Na referência [14], destaca-se a importância da determinação da confiabilidade de sistemas e equipamentos (*system and equipment performance assessment*). Neste artigo, o autor descreve os índices de avaliação adotados por uma concessionária de energia do Canadá fazendo uma breve descrição de cada um deles: *SAIFI , CAIFI , SAIDI , CAIDI* e *ASAI*. Além disso, o autor classifica os componentes em sete classes: linha de distribuição, cabo de distribuição, transformador de distribuição, transformador de potência, chaves, regulador e capacitor. Ainda, sugerem-se seis tipos de causas primárias de defeito: equipamento defeituoso, clima adverso, ambiente adverso, elemento humano, interferência externa e causa desconhecida. Neste trabalho, o autor conclui que a qualidade dos dados é de fundamental importância na busca de bons resultados na estimação de confiabilidade e desempenho já que estes são obtidos geralmente através de dados estatísticos. Em outras palavras, o grau de precisão da identificação de causas de defeitos esta ligado diretamente à qualidade das informações disponíveis para análise.

A referência [15] apresenta conceitos sobre confiabilidade em sistemas de distribuição, apresentando os aspectos de duração e freqüência de interrupções, com a preocupação na origem das informações dos desligamentos, proporcionando um melhor entendimento do impacto das causas nesses índices. Nesse trabalho, destaca-se a existência de um percentual elevado da causa “desconhecida” o que em muitas situações dificulta a identificação dos efeitos, pois neste caso os dados estão sujeitos a um elevado grau de incerteza.

Em [16] o autor questiona o que é confiabilidade na distribuição de energia, apresentando uma preocupação centrada no controle dos índices de confiabilidade através de uma configuração adequada dos circuitos alimentadores, os quais

proporcionam perfeitas condições de coordenação e seletividade da sua proteção, bem como aspectos de automação das redes de distribuição.

A referência [17] discute a importância da prevenção de falhas no fornecimento de energia e como afetam a confiabilidade, segurança e qualidade dos sistemas de distribuição. A análise realizada no histórico de eventos de desligamento da empresa mostra que animais são a segunda maior causa de desligamentos na região de cobertura da concessionária. Um estudo ainda mostra a correlação existente entre as falhas causadas por animais e fatores externos como, ID do circuito, condições climáticas, estação do ano, horário e número de fases afetadas, permitindo uma melhor compreensão das causas e conseqüências das falhas causadas por animais. A colocação de protetores cobrindo os transformadores e isoladores diminuiu o número de falhas evitando que ocorra curto-circuito quando caso um animal se encoste em uma fase e no transformador. Essas ações de prevenção reduziram de forma significante as falhas causadas por animais e, como conseqüência, os índices de confiabilidade *SAIDI* e *SAIFI*.

Métodos de Inteligência Artificial são muito utilizados em aplicações de sistemas de potência. Em 1986, Fukui e Kawakami da Hitachi Ltd., do Japão desenvolveram um sistema especialista utilizando informação de relés de proteção e disjuntores para estimar a seção em que ocorreu a falha no sistema de transmissão [18].

Em [19], um sistema de classificação de falhas em sistemas de distribuição foi desenvolvido para demonstrar a aplicação de métodos de classificação como Redes Neurais Artificiais e Regressão Logística aplicado ao histórico de eventos de desligamento da empresa *Duke Energy*. As causas consideradas para ilustrar a técnicas de classificação utilizadas são contato com animais e árvores. O artigo ainda utiliza quatro medidas de desempenho para comparar as duas heurísticas: taxa de classificação correta, taxa de positivo verdadeiro, taxa de negativo verdadeiro, e média geométrica. As vantagens e desvantagens entre as duas técnicas também são abordadas, assim como citado em [20].

Na referência [21], um sistema especialista foi proposto para diagnosticar falhas em tempo real. O sistema utiliza a informação baseada no estado de relés e disjuntores da rede elétrica para identificar o mais provável elemento com falha, servindo de suporte na tomada de decisões no centro de operações. É utilizada uma árvore de decisão para identificar os elementos com maior probabilidade de estar em falha.

Em [22] é apresentada uma metodologia que utiliza informações de três fontes diferentes, sendo elas: *Call Center*, dados do *SCADA* e dados do *Automatic Meter Reading* (*AMR)* para identificar falhas no sistema de distribuição de energia elétrica através de lógica nebulosa e sistema baseado no conhecimento. Neste trabalho, mostra-se que a combinação de sistemas inteligentes e métodos numéricos pode melhorar a qualidade da análise de sistemas de distribuição de energia.

Na referência [23] é proposto um sistema similar ao da referência [22], onde é usado um sistema baseado no conhecimento para a identificação de locais de falhas nas linhas de transmissão de energia através das três fontes de informação utilizadas em [22] (*Call Center*, dados do *SCADA* e dados do *AMR*). Neste trabalho, métodos e técnicas de inteligência artificial proporcionaram uma forma de capturar o conhecimento de especialistas e engenheiros para o treinamento do sistema. Em particular, o sistema especialista G2 foi aplicado como uma ferramenta para métodos de operação avançados, acesso a banco de dados e interface gráfica.

Buscando mostrar a influência das condições climáticas nos índices de confiabilidade de uma concessionária de energia, foi realizado em [24] um estudo objetivando dar suporte à alocação de equipes de atendimentos de emergência em condições climáticas adversas. Para isso, duas técnicas de agrupamento foram comparadas, Redes Neurais tipo *SOM (Self Organizing Maps)* e o algoritmo *K-means*. Concluiu-se que ambas técnicas são compatíveis com a realidade técnica/climática e que se pode validar o modelo desenvolvido para que seja utilizado em outros períodos do ano.

A referência [25] propôs, a partir de um banco de dados com informações metereológicas e de um banco de dados com índices de desempenho de empresas do setor elétrico, uma metodologia para identificar padrões climáticos que influenciem na qualidade do fornecimento de energia. Foram aplicados métodos de estatística multivariada, tais como análise de componentes principais e análise de *cluster*, e Redes Neurais tipo *SOM*. As duas técnicas mostraram uma concordância de 93.73 % entre os padrões descobertos aplicando as duas metodologias ao longo do ano em estudo. Este trabalho permitiu que as empresas definissem com mais segurança a quantidade de equipes de apoio para ficar em alerta em períodos de condições climáticas adversas.

Uma Rede Bayesiana é utilizada em [7] para localização de falhas no sistema de distribuição. A rede foi montada com base nas informações de especialistas e dados

históricos de uma concessionária de energia de Taiwan. Os resultados obtidos com a Rede Bayesiana mostraram o grande potencial que essa abordagem tem para localização de falhas em alimentadores.

No artigo [8] é feita a modelagem de uma Rede Bayesiana desenvolvida a partir do conhecimento de *clusters* (ou agrupamentos) climáticos e sua relação com o número de paradas não programadas, utilizando-se uma base de dados considerando as observações diárias num período de 12 meses. Primeiramente, foram identificados os padrões climáticos mensais através de técnicas estatísticas e redes neurais. Posteriormente, investigou-se a relação dos *clusters* climáticos com o número de paradas não programadas. Finalmente, foi desenvolvida uma Rede Bayesiana que se constitui num cenário que propaga a informação climática diária de forma a oferecer uma previsão do número de paradas emergenciais. Dessa forma, a Rede Bayesiana pode ser útil como apoio à gestão de equipes de atendimentos a falhas no sistema elétrico.

Atualmente, não existem modelos matemáticos adequados para análise da relação de causa e efeito em falhas no sistema de distribuição. Isso se dá principalmente pela complexidade e não-linearidade desses sistemas. Por esse motivo, o uso de RNA é indicado devido à natureza do problema. Em [26], uma RNA é implementada para o diagnóstico de falhas causadas por animais. A metodologia envolve organização e análise dos dados coletados sobre os eventos de desligamento, assim como a escolha das variáveis apropriadas para a entrada da RNA. Muitas vezes, falhas causadas por diferentes fatores podem levar às mesmas conseqüências, por exemplo, tanto um curto-circuito quanto um acidente de trânsito podem levar à abertura de um relé. Então, se a causa real for detectada de antemão, é possível alocar a equipe e os equipamentos adequados para restauração do sistema o mais rápido possível. A RNA mostrou-se capaz de predizer corretamente falhas causadas por animais em aproximadamente 98% dos casos de teste. Esses resultados mostram o potencial de RNA para identificação de falhas em sistemas de potência.

Assim como em muitas outras áreas, em sistemas de potência existe um grande crescimento na quantidade de dados armazenada pelos sistemas de informação. Segundo [27], as principais fontes de desses dados são: (i) dados de campo, que podem se coletados através de diversos dispositivos ao longo do sistema, (ii) sistema *SCADA*, (iii) dados obtidos através de ambientes de simulação de planejamento ou operação. Esse enorme conjunto de dados torna difícil para um profissional a tarefa de entendê-lo

como um todo. Nesse contexto é que a utilização de algoritmos capazes de sintetizar estruturas a partir de dados se torna uma necessidade. Implementar e aplicar esses algoritmos a problemas reais é o que se propõe o campo de mineração de dados. Esse artigo descreve diversas ferramentas utilizadas na mineração de dados, exemplificando através de um software desenvolvido para estudos de avaliação da segurança dinâmica.

Em [28], é descrito uma série de aplicações de mineração de dados para sistemas de potência. O autor diz que técnicas de inteligência artificial, principalmente sob a forma de sistemas especialistas, são utilizadas como ferramentas de apoio para os operadores. No entanto, ele lembra que essas técnicas possuem algumas limitações. O desempenho do sistema especialista irá depender da quantidade e da qualidade da informação utilizada com base de conhecimento. É difícil construir sistemas especialistas que possam capturar totalmente o conhecimento de especialistas humanos. A mineração de dados é utilizada para contornar esse problema. Conhecida também como extração de conhecimento em banco de dados, esse processo permite organizar a informação de uma forma mais adequada para a utilização em algoritmos de aprendizado de sistemas especialistas.

Regras de indução são utilizadas em [29], a empresa *General Motors* utiliza um banco de dados de relatórios de problemas ocorridos com seus automóveis para construir sistemas especialistas de diagnóstico. Similarmente, mineração de dados pode ser utilizada para casos de desligamentos não programados na rede de distribuição de energia, gerando dados para construção de um sistema especialista para identificação de causas de desligamento.

No trabalho realizado em [30] equipes de manutenção coletaram cada falha ocorrida no sistema numa tabela que incluía hora, data, mês, ano, endereço, equipamento em que ocorreu a falha, causa ou acidente, e etc. O banco de dados acumulou uma grande base de informação durante muitos anos. Nesse foram utilizados dados brutos e mineração de dados para derivar padrões e regras para as o diagnóstico de falhas ocorridas nos equipamentos da rede de distribuição e localização de falhas.

O artigo [31] apresenta um estudo sobre o estado da arte em mineração de dados para sistemas de potência. Devido a não linearidade existente nos sistemas de potência, o autor ressalta a importância da utilização de técnicas de extração de conhecimento de banco de dados para suporte da operação e no planejamento. A mineração de dados é uma etapa de todo o processo de extração de conhecimento. Alguns métodos utilizados

em mineração de dados, como árvores de decisão, RNA e sistemas *fuzzy*, são citados e sucintamente descritos. Ainda são apresentadas 42 referências sobre trabalhos que dão uma visão geral sobre o estado da arte em aplicações de mineração de dados para sistemas de potência.

Em alguns casos existe a necessidade de realizar um trabalho de pré-processamento da informação utilizada no treinamento de sistemas baseados em conhecimento ou para mineração de dados. Isso se aplica quando o conjunto de treinamento está desbalanceado (*imbalanced data*). Para contornar essa situação sistemas baseados em lógica *fuzzy*, *I-algorithm e E-algorithm* [32] e no algoritmo AIRS (*Artificial Immune Recognition System*) [33] foram criados para aliviar os efeitos de dados desbalanceados na performance da identificação de falhas no sistema de distribuição.

1.5. Estrutura da Dissertação

Este trabalho está organizado conforme segue. O capítulo apresentado mostra o contexto do estudo, através de seus objetivos, caracterização do problema e uma revisão bibliográfica sobre confiabilidade de sistemas de energia e diagnóstico de falhas em redes de distribuição. O capítulo 2 se propõe a realizar a fundamentação teórica, dando ênfase a conceitos básicos de probabilidade, redes bayesianas, redes neurais artificiais, métodos alternativos, extração de conhecimento e confiabilidade de sistemas de energia. No capítulo 3 se encontra a metodologia, com detalhamento de todas etapas realizadas, definição das variáveis, tratamento dos dados, implementação da rede neural, implementação da rede bayesiana e apresentação dos resultados. Por fim, o Capítulo 4 apresenta as conclusões e indicações para trabalhos futuros.

2. Fundamentação Teórica

Neste Capítulo são apresentados conceitos e definições que serão necessários para a construção da metodologia proposta. A primeira seção apresenta conceitos básicos sobre teoria de probabilidade. Na segunda seção as Redes Bayesianas (RB) são apresentadas com os respectivos algoritmos de inferência e aprendizado utilizados. Em seguida, é feita uma introdução sobre RNA, assim como a descrição do algoritmo de treinamento utilizado. Na quarta seção, alguns métodos alternativos às RB e RNA são apresentados de forma sucinta. Na seção seguinte, são descritos os passos do processo de extração de conhecimento em banco dados, conhecido pela sigla *KDD*, com destaque para o método utilizado no processo mineração de dados. Por último, apresenta-se uma revisão e alguns conceitos básicos de confiabilidade de sistema distribuição de energia elétrica necessários ao estudo de identificação e análise de interrupções de fornecimento de energia.

2.1. Conceitos Básicos de Probabilidade

Uma Rede Bayesiana pode ser representada por uma tabela de conjunção de probabilidades. As análises feitas com base em uma Rede Bayesiana estão sempre relacionadas à probabilidade de algo ocorrer condicionado a um ou mais eventos. Assim, conhecimentos de teoria de probabilidades tornam-se essenciais para a compreensão das Redes Bayesianas. Por essa razão, a seguir é apresentado um conjunto de conceitos mínimos necessários ao entendimento deste assunto.

2.1.1. Modelos Matemáticos

Na natureza existem fenômenos que podem ter seu comportamento representado através de modelos matemáticos chamados determinísticos. Um modelo determinístico pode ser referido como um modelo que estipula que as condições sob as quais um

experimento é executado determinem os resultados do experimento [34]. Pode-se citar, por exemplo, as leis do movimento de Newton onde as equações são modelos matemáticos que descrevem de uma forma bastante precisa o comportamento do movimento dos corpos. Na prática, isso seria como medir o tempo que uma bola leva para atingir o solo quando ela é solta de uma determinada altura repetidas vezes. Num modelo determinístico esse tempo é sempre, ou quase sempre, o mesmo, uma vez que pequenos distúrbios como o vento, por exemplo, podem alterar essas medidas.

A teoria de probabilidades está diretamente relacionada ao estudo de fenômenos aleatórios e à incerteza. Segundo a referência [35], "*a teoria das probabilidades se fundamenta nas situações da vida real quando uma pessoa realiza uma experiência cujo resultado não pode ser o esperado. Tal experiência denomina-se experimento aleatório*".

Este tipo de fenômeno caracteriza-se pelo fato de que o comportamento futuro do modelo é imprevisível. A representação deste tipo de sistema utilizando modelos determinísticos não é geralmente possível. No entanto, pode-se criar um outro tipo de modelo nesta situação. Por exemplo, utiliza-se um experimento que realizado repetidas vezes nas mesmas condições e circunstâncias produz resultados diferentes. Este é tipicamente o caso do lançamento de dois dados não viciados. Eles são denominados modelos não-determinísticos ou modelos probabilísticos. No exemplo dos dados, não se pode determinar com certeza quais serão os números sorteados, mas pode-se nesta situação saber a probabilidade de saírem os números escolhidos. Estes modelos consistem em uma listagem com todos resultados possíveis e suas respectivas probabilidades. A teoria das probabilidades nos permite então predizer ou deduzir padrões de futuros resultados [36].

2.1.2. Conjuntos, Espaço Amostral e Evento

Um conjunto é uma coleção de objetos. Usualmente, conjuntos são representados por letras maiúsculas A, B, C,...,Z [34]. Os objetos de um conjunto são chamados elementos ou membros. Um conjunto qualquer A pode ser descrito através de palavras, como $A=\{1, 2, 3 , 4\}$ ou ainda $A=\{x \mid 1 \leq x \leq 4 \}$.

O espaço amostral S define-se como sendo o conjunto de todos os resultados possíveis de um experimento aleatório E [36]. Por exemplo [34]:

E_1: Jogar um dado e observar o número mostrado na face de cima.

S: {1,2,3,4,5,6}

E_2: Jogar uma moeda quatro vezes para cima e observar o número de caras obtido.

Um evento A é simplesmente um conjunto de resultados possíveis, que pode ser tanto um subconjunto do espaço amostral S, o próprio espaço amostral S, bem como o conjunto vazio $\varnothing$.

2.1.3. Freqüência Relativa

Em estatística, freqüência é o número de vezes que um evento ocorre em um experimento. Admita-se que n_A é o número de vezes que o evento A ocorre nas n repetições do experimento E.

$$f_A = \frac{n_A}{n} \qquad 2.1$$

a equação 2.2 é denominada freqüência relativa do evento A nas n repetições de E. A freqüência relativa de um evento A tenderá a variar cada vez menos à medida que o número de repetições for aumentada [34].

Sejam A e B eventos de S, suas freqüências relativas apresentam as seguintes propriedades:

a) $0 \leq f_A \leq 1$,

b) $f_A = 1$, se $n_A = n$,

c) $f_A = 0$, se $n_A = 0$,

d) $f_{A \cup B} = f_A + f_B$, se $A \cap B = \varnothing$.

2.1.4. Axiomas Básicos da Probabilidade

A probabilidade $P(A)$ de um evento A é um número no intervalo unitário [0,1] e obedece aos seguintes axiomas básicos:

a) $P(A) = 1$, se e somente se A é verdadeiro;

b) Se A e B são eventos disjuntos de S, então $P(A \cup B) = P(A) + P(B)$;

c) Se $A_1, A_2, ..., A_n$ é uma família de eventos de S, dois a dois disjuntos, então $P(A_1 \cup A_2 \cup ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$.

2.1.5. Redes Causais e d-separação

Uma rede causal consiste num conjunto de variáveis e um conjunto de arcos direcionados entre as variáveis. Matematicamente, a estrutura é chamada grafo direcionado. Para expressar as relações entre as variáveis utiliza-se a seguinte notação: se existe um arco conectando A a B, então se diz que B é filho de A e A é pai de B [37].

As variáveis representam os nós. Uma variável pode ter qualquer número de estados. As redes causais podem ser utilizadas a fim de verificar como a certeza de uma variável pode influenciar na certeza de outras variáveis.

Conexão Serial

Na rede causal de conexão serial mostrada na Figura 2.1, A exerce influência na certeza de B, que por sua vez influencia na certeza de C. Da mesma forma, uma evidência em C irá influenciar na certeza de A através de B. No entanto, dado que o estado de B é conhecido, então o caminho através de B está bloqueado, e A e C se tornam independentes. Diz-se então que A e C estão d-separados dado B, e quando o estado de uma variável é conhecido, diz-se que ela está instanciada.

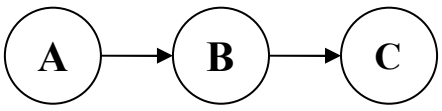


Figura 2.1 – Conexão serial.

Conexão Divergente

Na Figura 2.2, a influência dos nós filhos só poderá ser transmitida entre eles se estado de B não for conhecido. Se B for instanciado, então o caminho de comunicação entre os nós A e C estará bloqueado. Diz-se que A e C estão d-separados dado B.

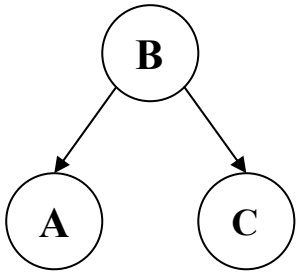


Figura 2.2 – Conexão Divergente.

Conexão Convergente

No caso de uma rede convergente, mostrada na Figura 2.3, se nada é conhecido a respeito de *A*, com exceção do que se pode inferir a partir do conhecimento de seus pais, então seus pais são independentes, ou seja, evidência em um deles não causará nenhuma influência sobre o outro. A informação só será transmitida entre os nós pais *B* e *C* se o estado de *A* ou de um de seus filhos for conhecido. Então, para este caso, *A* deve estar instanciado para que *B* e *C* influenciem um ao outro através de *A*.

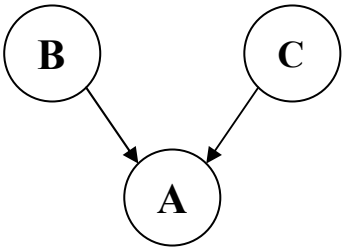


Figura 2.3 – Conexão Convergente.

Segundo Jensen [37], duas variáveis *A* e *B* numa rede causal são d-separadas se, para todos os caminhos entre *A* e *B*, existe uma variável intermediária *V* que para o caso de:

a) a conexão é serial ou divergente, e a variável *V* está instanciada;

ou

b) a conexão é convergente, e ou *V* ou qualquer um de seus filhos está instanciado.

2.1.6. Probabilidade Condicional

Sejam dois eventos *A* e *B* contidos em um espaço amostral *S*, associado a um experimento *E*. Supondo que seja conhecido o subconjunto *B* do espaço amostral *S*, com $P(B) \neq 0$. O conhecimento da ocorrência do evento *B* pode mudar a probabilidade de ocorrência do evento *A* [36].

De acordo com [38], a probabilidade condicional do evento *A* dado que o evento *B* ocorreu primeiro é denotada por $P(A \mid B)$ e definida como,

$$P(A \mid B) = \frac{P(A,B)}{P(B)} \qquad P(B) > 0 \tag{2.3}$$

onde $P(A,B)$ é a probabilidade conjunta de *A* e *B*. Da mesma forma,

$$P(B \mid A) = \frac{P(A,B)}{P(A)} \qquad P(A) > 0 \tag{2.4}$$

é a probabilidade condicional de um evento *B* dado o evento *A*. Das equações 2.3 e 2.4, tem-se que,

$$P(A,B) = P(A \mid B)P(B) = P(B \mid A)P(A) \tag{2.5}$$

a equação 2.5 é utilizada para o cálculo da probabilidade conjunta de eventos.

Da equação 2.5 pode-se chegar a seguinte regra, chamada Regra de Bayes:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2.6}$$

O principal uso desse teorema em aplicações de probabilidade é reverter o condicionamento dos eventos, isto é, mostrar como a probabilidade *A/B* está relacionada com a de *B/A*.

2.1.7. Cálculo de Probabilidades para Variáveis

Seja A uma variável com estados $a_1,...,a_n$, então $P(A)$ representa a distribuição de probabilidade sobre os estados:

$$P(A) = (a_1,...,a_n); \qquad a_i \geq 0; \qquad \sum_{i=1}^{n} a_i = 1$$

onde a_i é a probabilidade de A estar no estado a_i, e é denotada por $P(A = a_i)$.

Segundo [37], se a variável B tem os estados $b_1,...,b_m$, então $P(A \mid B)$ representa uma tabela $n \times m$ contendo os números $P(a_i \mid b_j)$, como mostrado na Tabela 2.1.

Tabela 2.1 – Exemplo de $P(A \mid B)$.

	b_1	b_2	b_3
a_1	0.4	0.3	0.6
a_2	0.6	0.7	0.4

A conjunção de probabilidades $P(A,B)$ para as variáveis A e B, também é uma tabela $n.m$. Ela consiste da probabilidade de cada configuração de (a_i,b_j), como mostrado na Tabela 2.2. Note que a soma das colunas é 1.

Tabela 2.2 – Exemplo de $P(A,B)$.

	b_1	b_2	b_3
a_1	0.16	0.12	0.12
a_2	0.24	0.28	0.08

A partir da Tabela 2.2, a probabilidade $P(A)$ pode ser calculada utilizando:

$$P(a_i) = \sum_{j=1}^{m} P(a_i, b_i) \qquad 2.7$$

Este cálculo é chamado marginalização de *B* em $P(A,B)$, diz-se que a variável *B* é marginalizada para fora de $P(A,B)$, resultando em *P(A)*, tal que

$$P(A) = \sum_{B} P(A,B)$$

Então, utilizando a Tabela 2.2, o cálculo para marginalização de *B* ficaria:

$$P(A = a_1) = P(a_1, b_1) + P(a_1, b_2) + P(a_1, b_3) = 0.16 + 0.12 + 0.12 = 0.4$$
$$P(A = a_2) = P(a_2, b_1) + P(a_2, b_2) + P(a_2, b_3) = 0.24 + 0.28 + 0.08 = 0.6$$

Assim, marginalizando *B* na Tabela 2.2, tem-se *P(A)=(0.4,0.6)*.

É possível reverter a dependência dos eventos e assim obter $P(B \mid A)$. Para isso aplica-se a regra de Bayes usando a equação 2.6. *P(A)* e *P(A|B)* são conhecidos, então é necessário marginalizar a variável *A* de *P(A,B)* para obter *P(B)* e tornar possível o cálculo de $P(B \mid A)$.

Marginalização de *A* em $P(A,B)$:

$$P(B) = \sum_{A} P(A,B)$$

Utilizando a Tabela 2.2, o cálculo para marginalização de *A* fica:

$$P(B = b_1) = P(a_1, b_1) + P(a_2, b_1) = 0.16 + 0.24 = 0.4$$
$$P(B = b_2) = P(a_1, b_2) + P(a_2, b_2) = 0.12 + 0.28 = 0.4$$
$$P(B = b_3) = P(a_1, b_3) + P(a_2, b_3) = 0.12 + 0.08 = 0.2$$

Assim, marginalizando *A* na Tabela 2.2, tem-se *P(B)=(0.4,0.4,0.2)*

Cálculo de *P(B|A)*:

$$P(B=b_1 \mid A=a_1)=\frac{P(A=a_1 \mid B=b_1)}{P(A=a_1)}=\frac{0.4\cdot 0.4}{0.4}=0.4$$

$$P(B=b_2 \mid A=a_1)=\frac{P(A=a_1 \mid B=b_2)}{P(A=a_1)}=\frac{0.3\cdot 0.4}{0.4}=0.3$$

$$P(B=b_3 \mid A=a_1)=\frac{P(A=a_1 \mid B=b_3)}{P(A=a_1)}=\frac{0.6\cdot 0.2}{0.4}=0.3$$

$$P(B=b_1 \mid A=a_2)=\frac{P(A=a_2 \mid B=b_1)}{P(A=a_2)}=\frac{0.6\cdot 0.4}{0.6}=0.4$$

$$P(B=b_2 \mid A=a_2)=\frac{P(A=a_2 \mid B=b_2)}{P(A=a_2)}=\frac{0.7\cdot 0.4}{0.6}=0.47$$

$$P(B=b_3 \mid A=a_2)=\frac{P(A=a_2 \mid B=b_3)}{P(A=a_2)}=\frac{0.4\cdot 0.2}{0.6}=0.13$$

A Tabela 2.3 mostra os resultados obtidos da aplicação da Regra de Bayes.

Tabela 2.3 – $P(B \mid A)$

	a_1	a_2
b_1	0.4	0.4
b_2	0.3	0.47
b_3	0.3	0.13

2.1.8. Eventos Independentes

Segundo [38], dois eventos A e B são ditos independentes se e somente se

$$P(A,B)=P(A).P(B) \qquad 2.8$$

Segue que, se A e B são independentes, então pelas equações 2.3 e 2.4,

$$P(A \mid B)=P(A) \quad \text{e} \quad P(B \mid A)=P(B) \qquad 2.9$$

2.1.9. Independência condicional

Como mostrado no item 2.1.8, duas variáveis são independentes se $P(A \mid B) = P(A)$ e $P(B \mid A) = P(B)$. Para o caso de três variáveis aleatórias discretas *A, B,* e *C* se diz que *A* e *C* são condicionalmente independentes dado *B* se:

$$P(A \mid B) = P(A \mid B, C) \qquad 2.10$$

Sendo conhecido o estado de *B*, então o conhecimento de *C* não irá causar influência na probabilidade da ocorrência de *A*. A independência condicional aparece nos casos de redes causais que apresentam conexão do tipo serial e divergente. Ver Figura 2.1 e Figura 2.2.

2.2. Redes Bayesianas

Nesta seção apresenta-se uma definição e conceitos relacionados a Redes Bayesianas. O algoritmo utilizado para realizar inferências na rede e o algoritmo de treinamento são explicados, sendo apresentado um exemplo para melhor entendimento.

2.2.1. Definição de Redes Bayesianas

As Redes Bayesianas se originam da aplicação do Teorema de Bayes [39] a problemas de natureza prática em diversos campos do conhecimento. Apesar das teorias de probabilidades e Bayesiana existirem há bastante tempo, foi somente no final dos anos 80 que as redes probabilísticas emergiram como uma nova abordagem para representação gráfica de dependências probabilísticas, da qual as Redes Bayesianas são exemplos desse tipo de representação. Isso ocorreu porque foi nos últimos anos que algoritmos e ferramentas de *software* foram desenvolvidas permitindo a propagação de evidências nas redes para um número razoável de variáveis.

As Redes Bayesianas permitem a manipulação de dados sujeitos à incerteza e podem ser vistas como um modelo gráfico de representação compacta da tabela de

conjunção de probabilidades sobre seu universo de variáveis [37]. Utilizando essa abordagem, um problema pode ser representado como um conjunto de variáveis e suas relações probabilísticas.

Segundo [40], uma Rede Bayesiana é um grafo acíclico direcionado (GAD) onde cada nó possui informação quantitativa de probabilidade. As regras Bayesianas apresentam as seguintes características:

1. Um conjunto de variáveis aleatórias representa os nós da rede. As variáveis podem ser discretas ou contínuas;
2. Um conjunto de arcos direcionados conecta os pares de nós. Se existe um arco conectando o nó *A* ao nó *B*, diz-se que o no *A* é pai de *B*, e por conseqüência *B* é filho de *A*;
3. Cada nó A_i tem uma distribuição de probabilidades condicional $P(A_i|pai(A_i))$, que quantifica o efeito que os pais têm sobre aquele nó;
4. O grafo é direcionado e acíclico.

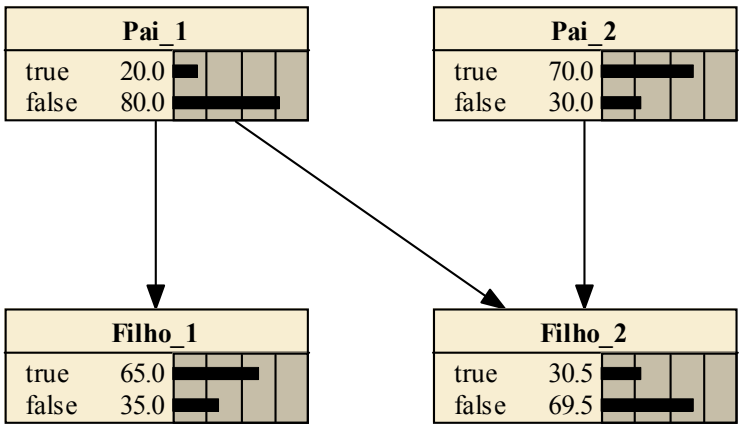


Figura 2.4 – Exemplo de Rede Bayesiana.

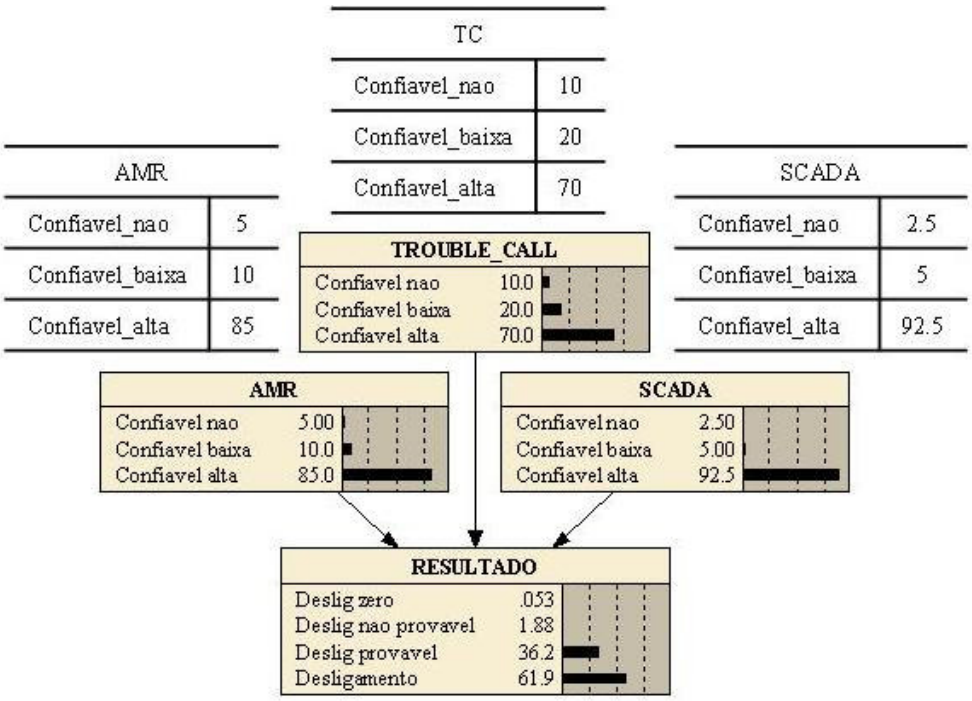
A topologia da rede, que consiste nos arcos e nós, irá estabelecer as relações de independência condicional que a rede possui. O teorema de Bayes, mostrado em 2.6, simplifica o cálculo quando existe uma independência condicional entre as variáveis. Isso também pode reduzir o número de probabilidades condicionais [40] que devem ser representadas nas Tabelas de Probabilidades Condicionais (CPT[1]) . A *CPT* é uma tabela n-dimensional com o valor de cada célula dando a probabilidade de ocorrência daquele

[1] *Conditional Probability Table*

estado específico, e pode ser utilizada para calcular qualquer probabilidade a respeito do domínio em estudo [40].

Uma vez definida a topologia, é necessário especificar a distribuição de probabilidade condicional de cada variável, dado os seus pais. Isto pode ser feito por um especialista diretamente nas tabelas de probabilidades condicionais, ou utilizando algum método de aprendizagem.

A Figura 2.5 ilustra um exemplo de Rede Bayesiana inspirada no filtro *Fuzzy* implementado em [22]. Os nós representam as variáveis envolvidas, e os estados desses nós representam o grau de certeza da variável. A cada fonte de informação sobre desligamentos, *AMR*, *SCADA* e *Trouble Call* (chamadas de clientes), são atribuídos graus de confiança no que diz respeito à veracidade dos dados obtidos. Cruzando essas fontes de informação, é possível avaliar de forma mais precisa a qualidade e veracidade das leituras obtidas via *AMR* e *SCADA*, e através das chamadas telefônicas de clientes.



AMR	TC	SCADA	Zero	Nao_provavel	Provavel	Desligamento
Confiavel_nao	Confiavel_nao	Confiavel_nao	100	0	0	0
Confiavel_nao	Confiavel_nao	Confiavel_baixa	0	90	10	0
Confiavel_nao	Confiavel_nao	Confiavel_alta	0	10	90	0
Confiavel_nao	Confiavel_baixa	Confiavel_nao	0	90	10	0
Confiavel_nao	Confiavel_baixa	Confiavel_baixa	0	95	5	0
Confiavel_nao	Confiavel_baixa	Confiavel_alta	0	5	95	0
....	....	....	....	....	....	....
Confiavel_alta	Confiavel_alta	Confiavel_alta	0	0	0	100

Figura 2.5 – Exemplo de RB. Somente alguns casos são mostrados.

2.2.2. Regra da Cadeia para Redes Bayesianas

Seja $U = \{A_1,...,A_n\}$ um universo de variáveis. Utilizando a tabela de probabilidades conjuntas $P(U) = (A_1,...,A_n)$ é possível calcular $P(A_i)$. No entanto, $P(U)$ cresce exponencialmente com o número de variáveis do domínio, tornando a tabela muito difícil de ser tratada. Representando $P(U)$ através de uma Rede Bayesiana pode-se obter uma forma mais compacta para a distribuição de probabilidades conjuntas, permitindo que $P(U)$ seja calculado se necessário [37].

Seja uma Rede Bayesiana sobre o domínio $U = \{A_1,...,A_n\}$, então, a distribuição de probabilidades conjunta $P(U)$ é o produto de todas probabilidades especificadas na rede [37].

$$P(U) = \prod_i P(A_i \mid pa(A_i)) \qquad 2.11$$

onde $pa(A_i)$ é o conjunto de pais de A_i.

A seguir será mostrado um exemplo de representação da distribuição de probabilidades conjunta utilizando a regra da cadeia.

Supondo a Rede Bayesiana da Figura 2.6 consistindo de cinco variáveis A, B, C, D, E, onde todas são dependentes. A regra da cadeia permite calcular a distribuição de probabilidades conjunta de $P(A,B,C,D,E)$ como:

$$P(A,B,C,D,E) = P(E \mid A,B,C,D)P(D \mid A,B,C)P(C \mid A,B)P(B \mid A)P(A)$$

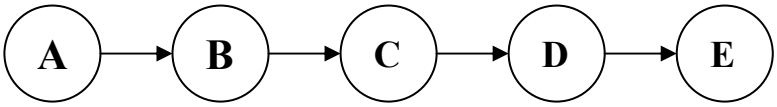


Figura 2.6 – Exemplo de RB onde as variáveis são dependentes.

Agora, supondo que as dependências estão explicitamente modeladas, como mostrado na Figura 2.7, então a representação da distribuição de probabilidades conjunta fica bastante simplificada:

$$P(A,B,C,D,E) = P(A)P(B)P(C \mid A,B)P(D \mid C)P(E \mid C)$$

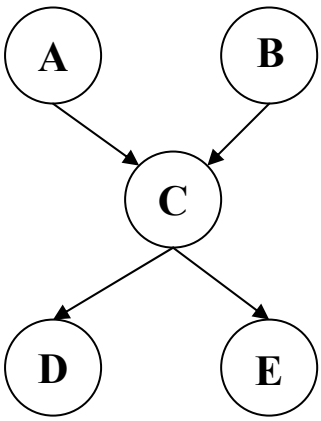


Figura 2.7 – Exemplo de RB onde as estruturas de independência condicional estão explícitas.

Através desse exemplo verifica-se a importância da utilização do conceito de independência condicional em Redes Bayesianas, permitindo que o modelo de representação de distribuição de probabilidades conjuntas seja feito de uma forma mais compacta, já que a existência de nós condicionalmente independentes reduz drasticamente o esforço computacional para o cálculo das probabilidades desejadas.

2.2.3. Inferência em Redes Bayesianas

O processo de inferência em Redes Bayesianas consiste em obter estimativas de probabilidades de eventos relacionados aos dados, dado o conhecimento de uma nova informação ou evidência. A inferência pode ser realizada através de métodos exatos, no entanto, alguns problemas podem se tornar intratáveis utilizando esse tipo de inferência [40]. Para essas situações onde não é possível utilizar métodos exatos, podem-se usar métodos aproximados e também métodos simbólicos. Neste trabalho foi adotado um método exato para realizar as inferências, que serão feitas através do método de eliminação de variáveis [41].

Seja um conjunto de variáveis $U = \{A_1,...,A_n\}$ de distribuição de probabilidades conjunta $P(U)$ para A, expressa pela equação 2.11. Dada uma certa evidência E, que é um subconjunto de variáveis conhecidas $E \subset U$, o processo de propagação irá considerar essas evidências no cálculo das novas probabilidades dos nós da rede. Esse processo consiste no cálculo *a posteriori* $P(A_i \mid e)$ para cada variável $A_i \notin E$, dada a evidência $E = e$.

Definição 2.1 (Evidência):
Um subconjunto de variáveis $E \subset U$ com valores conhecidos, E = e, em uma dada situação, é conhecido como conjunto de evidência, ou simplesmente evidência.

Se não há evidência, as funções condicionadas $P(A_i \mid e)$ são as funções de probabilidades marginais $P(A_i)$, para cada $A_i \in A$. O cálculo de probabilidades marginais de variáveis foi abordada na subseção 2.1.7. De forma genérica, quando não existem evidências, o cálculo da probabilidade de qualquer variável do conjunto pode ser obtida através da equação,

$$P(A_i) = \alpha \sum_{A \setminus A_i} P(A_1, ..., A_n) \qquad 2.12$$

onde $A \setminus A_i$ são todas as combinações dos estados das variáveis em A sem considerar a variável A_i e α é uma constante de normalização para assegurar que a soma da distribuição de probabilidades seja 1.

Esse método é bastante ineficiente, pois envolve um número elevado de combinações para o cálculo de uma determinada probabilidade. Mesmo para um conjunto pequeno de variáveis envolvidas, o problema pode se tornar intratável, tendo em vista que, para o caso de variáveis binárias, a equação 2.12 requer a soma de 2^{n-1} probabilidades distintas.

O problema do método exposto anteriormente é que não se considera a estrutura de independência condicional contida na função *P(U)*. Através de uma Rede Bayesiana pode-se representar a distribuição de probabilidades *P(U)*, e assim simplificar o processo de eliminação de variáveis levando-se em conta a independência condicional existente entre as variáveis da função de distribuição de probabilidades *P(U)*. A regra da cadeia, equação 2.11, dá uma representação compacta da distribuição de probabilidades conjuntas.

Considerando o seguinte exemplo, exposto em [40]:

"Você possui um novo alarme contra ladrões em casa. Este alarme é muito confiável na detecção de ladrões, entretanto, ele também pode disparar caso ocorra um terremoto. Você tem dois vizinhos, João e Maria, os quais prometeram telefonar-lhe no trabalho caso o alarme dispare. João sempre liga quando ouve o alarme, entretanto, algumas vezes confunde o alarme com o telefone e também liga nestes casos. Maria, por outro lado, gosta de ouvir música alta e às vezes não escuta o alarme". A Rede Bayesiana que representa esse problema pode ser vista na Figura 2.8.

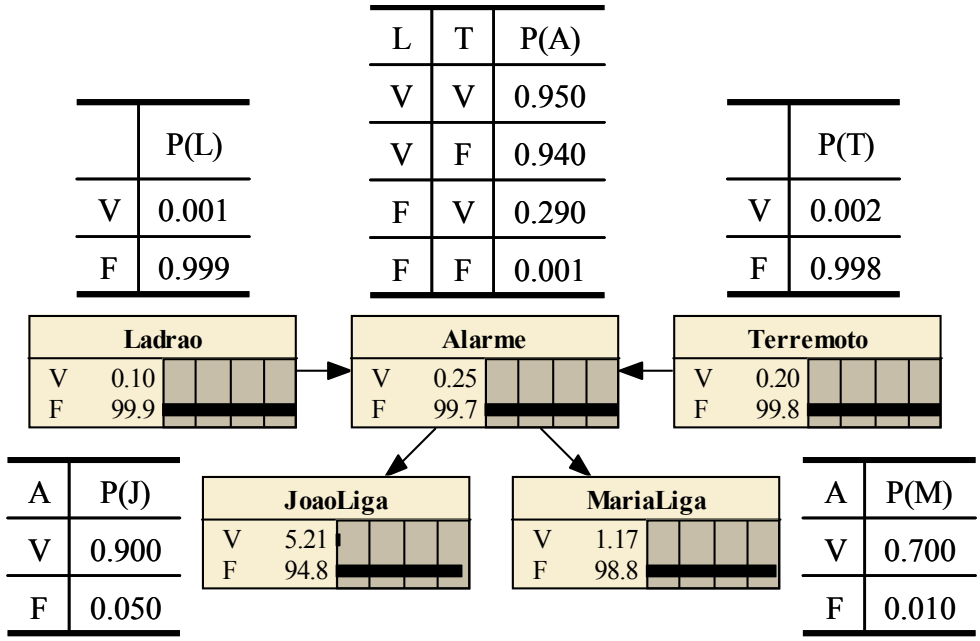


Figura 2.8 – Rede Bayesiana do Alarme e suas *CPT*.

Considere a consulta,

$P(Alarme \mid Terremoto = verdadeiro, MariaLiga = verdadeiro)$, ou seja, a probabilidade do alarme tocar, dado que ocorreu um terremoto e Maria telefonou. As variáveis incógnitas para essa consulta são *Ladrão* e *Joãoliga*. Da equação 2.12, utilizando as iniciais das variáveis para simplificar as expressões, tem-se:

Para $A=V$

$$P(A \mid t, m) = \alpha \times P(A, t, m) = \alpha \times \sum_{l} \sum_{j} P(A, t, m, l, j)$$

Utilizando a regra da cadeia, equação 2.11, e valendo-se da estrutura de independência contida na distribuição de probabilidades conjuntas representada pela Rede Bayesiana, pode-se reduzir o número de termos da expressão, dada por

$$P(A \mid t, m) = \alpha \times \sum_{l} \sum_{j} P(l) \times P(t) \times P(a \mid l, t) \times P(j \mid a) \times P(m \mid a)$$

Reagrupando os termos calcula-se cada um dos somatórios de forma independente. Os termos $P(t)$ e $P(m \mid a)$ são constantes e podem ser movidos para fora do somatório:

$$P(A \mid t, m) = \alpha \times P(t) \times P(m \mid a) \times \sum_{l} P(l) \times P(a \mid l, t) \times \sum_{j} P(j \mid a)$$

O termo $\sum_j P(j \mid a)$ é igual a 1, de acordo com a propriedade do potencial unitário [37], simplificando bastante o cálculo. A expressão resulta em

$$P(A \mid t,m) = \alpha \times P(t) \times P(m \mid a) \times \sum_l P(l) \times P(a \mid l,t)$$

$$P(A \mid t,m) = \alpha \times P(t=v) \times P(m=v \mid a=v) \times [P(l=v) \times P(a=v \mid l=v, t=v) + P(l=f) \times P(a=v \mid l=f, t=v)]$$

Das *CPT*, obtém-se os valores das probabilidades, dado por

$$P(A \mid t,m) = \alpha \times 0.002 \times 0.7 \times [0.001 \times 0.95 + 0.999 \times 0.29] = \alpha \times 0.000406924$$

Fazendo o mesmo procedimento para *A=F*, obtém-se $\alpha \times 0.0000141868$.

Normalizando, chega-se a:

$$P(A \mid t,m) = \alpha \times \langle 0.000406924, 0.0000141868 \rangle = \langle 0.96631, 0.03689 \rangle$$

Então a probabilidade do alarme tocar, dado que ocorreu um terremoto e que Maria telefonou é aproximadamente de 96.6%. Foi utilizado o software *Netica* [42] para construir as Redes Bayesianas e validar a inferência realizada. A Figura 2.9 mostra os dois eventos, *MariaLig=V* e *Terremoto=V* instanciados, e a probabilidade do Alarme tocar dadas essas evidências.

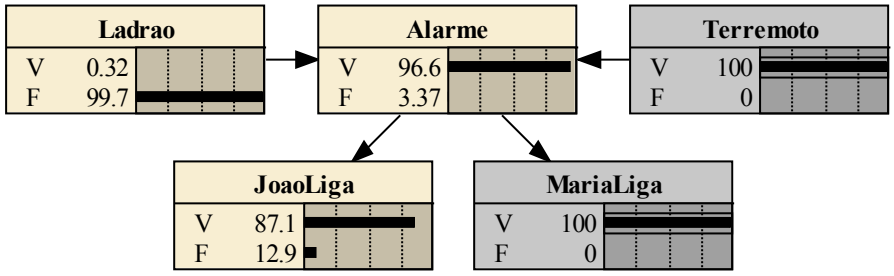


Figura 2.9 – Rede Bayesiana com os eventos instanciados.

2.2.4. Aprendizagem em Redes Bayesianas

Para o caso em que a rede tem estrutura conhecida, o problema da aprendizagem em Redes Bayesianas irá se resumir ao aprendizado dos parâmetros da rede, ou em outras palavras, a definição da relação das probabilidades condicionais e incondicionais entre as variáveis da rede.

Esse processo de aprendizagem pode ser realizado por um especialista, introduzindo as relações de probabilidade condicionais diretamente nas *CPT*. No entanto, para modelos que tenham um número grande de variáveis, esse processo pode se tornar bastante complicado, devido ao número de combinações possíveis entre os estados das variáveis. A aprendizagem pode também ser feita através de indução, a partir de uma amostra de dados. O método pode variar de acordo com o tipo de amostra de dados disponível. O caso mais simples é o para situações em que existe um conjunto de dados completo.

Nesse trabalho, a base de dados que é utilizada para o treinamento da Rede Bayesiana apresenta falta de dados e informações incompletas, o que torna o problema do aprendizado da rede um pouco mais complicado. Para contornar este tipo de problema foi utilizado o algoritmo *Expectation Maximization (EM),* o qual é adequado para este tipo de situação, sendo descrito a seguir.

Algoritmo *Expectation Maximization* (*EM*)

O algoritmo *EM* aplica-se para solução de problemas onde alguns estados das variáveis não puderam ser observados, então se podem utilizar os casos para os quais foram observados os estados para estimar o estado das variáveis que não puderam ser observados. O algoritmo *EM* pode ser utilizado também para variáveis cujos valores nunca foram observados, sempre e quando seja conhecida a forma geral da distribuição de probabilidade das variáveis. O algoritmo começa com uma hipótese inicial arbitrária. A seguir calcula-se repetidamente os valores esperados para as variáveis não observadas (assumindo que a hipótese atual está correta), depois recalcula-se a estimativa de máxima verossimilhança (assumindo que as variáveis não observadas têm seus valores calculados no primeiro passo). Esse procedimento converge para uma estimativa de máxima verossimilhança local com os valores estimados para as variáveis não observadas.

Antes de demonstrar o funcionamento do algoritmo *EM* deve-se definir função de verossimilhança e estimador de máxima verossimilhança conforme [43], para o caso de n variáveis aleatórias que dependem do parâmetro θ, que pode até ser um vetor de parâmetros.

Definição 2.2 - Função de Verossimilhança e Estimador de Máxima Verossimilhança:

Suponha que X é uma variável aleatória com distribuição de probabilidades $f(x;\theta)$, *onde* θ *é um parâmetro desconhecido. Sejam* $x_1,...,x_m$ *os valores observados em uma amostra aleatória de tamanho n. Então a* ***função de verossimilhança*** *da amostra é:*

$$L(\theta) = f(x_1;\theta) \cdot f(x_2;\theta) \cdot ... \cdot f(x_n;\theta) \cdot \qquad 2.13$$

Note que a função de máxima verossimilhança é agora função somente dos parâmetros desconhecidos θ. *O* ***estimador de máxima verossimilhança*** *de* θ *é o valor de* θ *que maximiza a função de verossimilhança* $L(\theta)$.

Para o caso de variáveis discretas, o estimador de máxima verossimilhança é um estimador que maximiza a probabilidade de ocorrência dos valores da amostra.

Dada a função de verossimilhança $L(w \mid y)$. De acordo com [44], o estimador de máxima verossimilhança pode ser obtido através da maximização do logaritmo da função de verossimilhança, $\ln L(w \mid y)$. Assumindo que o logaritmo da função de verossimilhança, $\ln L(w \mid y)$ é diferenciável, se w_{MLE} existe, então ele deve satisfazer a seguinte equação diferencial parcial conhecida como equação de verossimilhança:

$$\frac{\partial L(w \mid y)}{\partial w_i} = 0 \qquad 2.14$$

em $w_i = w_{i,MLE}$ para todo *i=1,...,k.* Isto se dá porque pela definição de máximo ou mínimo de uma função diferencial contínua, a derivada primeira é zero nesses pontos.

A equação de verossimilhança é uma condição necessária para a existência de uma estimação de máxima verossimilhança. No entanto, uma condição adicional deve ser também satisfeita para assegurar que $\ln L(w \mid y)$ é um máximo e não um mínimo, já que a derivada primeira não mostra isso. Para que seja um máximo, o logaritmo da função de verossimilhança deve ser convexo na vizinhança de w_{MLE}. Isso pode ser verificado calculando a derivada segunda do logaritmo da função de verossimilhança, mostrando que ela é sempre negativa em $w_i = w_{i,MLE}$ para *i = 1,...,k.*

$$\frac{\partial^2 L(w \mid y)}{\partial w_i^2} < 0 \qquad 2.15$$

Conforme descrito em [45], o algoritmo *EM* pode ser aplicado nos casos onde se deseja estimar algum conjunto de parâmetros θ, que descreve uma certa distribuição de probabilidades conjunta, dada somente uma parte observada dos dados produzidos por esta distribuição.

Seja, $X = \{x_1,...,x_m\}$ o conjunto de dados observados em um conjunto m de eventos ocorridos independentemente, seja $Z = \{z_1,...,z_m\}$ os dados não observados nestes mesmos eventos, e seja $Y = X \cup Z$ o total de dados. Pode-se tratar *Z* como uma variável aleatória cuja distribuição de probabilidades depende do conjunto de parâmetros desconhecidos θ e dos dados observados *X*. Analogamente, *Y* é uma variável aleatória porque é definida em termos da variável aleatória *Z*. Para descrever a forma geral do algoritmo *EM*, *h* denota a hipótese atual dos parâmetros θ, e *h'* denota a hipótese revisada que é estimada em cada iteração do algoritmo *EM*.

O algoritmo *EM* busca a hipótese *h'* de máxima verossimilhança, isto é, que maximize $E[\ln P(Y \mid h')]$. Este valor esperado é calculado sob a distribuição de probabilidade de *Y*, que é determinada pelos parâmetros desconhecidos θ.

$E[\ln P(Y \mid h')]$ tem o seguinte significado:

P(Y |h') é a verossimilhança de todos os dados *Y*, dada a hipótese *h'*. Deseja-se encontrar *h'* que maximize a função. Maximizando o logaritmo desta função *ln P(Y |h')* também se maximiza *P(Y|h')*. Introduz-se o valor esperado $E[\ln P(Y \mid h')]$ porque o total de dados *Y* é, ele próprio, uma variável aleatória.

Dado que os dados *Y* são uma combinação dos dados observados *X* e não observados *Z*, deve-se mediar sobre os possíveis valores não observados *Z*, ponderando cada um de acordo com suas probabilidades. Em outras palavras, toma-se o valor esperado $E[\ln P(Y \mid h')]$ sobre a distribuição de probabilidade da variável aleatória *Y*. A distribuição de probabilidades de *Y* é determinada pelos valores completamente conhecidos para *X*, mais a distribuição de probabilidade de *Z*.

Em geral não se sabe a distribuição de probabilidade de *Y* porque ela é determinada pelos parâmetros θ que se está tentando estimar. Entretanto, o algoritmo *EM* usa sua hipótese atual *h* no lugar do parâmetro θ atual para estimar a distribuição de probabilidades de *Y*.

Considere a definição de uma função *Q(h'|h)* que dá $E[\ln P(Y \mid h')]$ como uma função de *h'*, sob a suposição que $\theta = h$ e dada a porção observada *X* dos dados *Y*, expressa por,

$$Q(h' \mid h) = E[\ln P(Y \mid h') \mid h, X] \qquad 2.16$$

Escreve-se esta função *Q* na forma $Q(h' \mid h)$ pra indicar que ela é definida em parte pela suposição que a hipótese atual *h* é igual a θ.

O algoritmo *EM* repete os dois passos seguintes, até a convergência:

Passo E (***Expectation***): calcula Q(h'|h) usando a hipótese atual *h* e os dados observados *X* para estimar a distribuição de probabilidade sobre *Y*.

$$Q(h' \mid h) \leftarrow E[\ln P(Y \mid h') \mid h, X] \qquad 2.17$$

Passo M (***Maximization***): troca a hipótese *h* pela hipótese h' que maximiza esta função Q.

$$h = \arg\max_{h'} Q(h' \mid h) \qquad 2.18$$

2.3. Redes Neurais Artificiais

As RNA são modelos matemáticos que se assemelham às estruturas neurais biológicas e que têm capacidade computacional adquirida por meio de aprendizado e generalização. Uma vantagem das redes neurais é que isso pode ser feito com base num conjunto de exemplos. Depois de treinada, o conhecimento da rede será armazenado como um padrão de pesos distribuídos através das conexões entre seus neurônios.

2.3.1. Modelo de um Neurônio

O neurônio é a unidade de processamento de informações fundamental para o funcionamento das RNA [46]. A Figura 2.10 mostra o modelo de um neurônio do tipo perceptron. Ele possui conjunto de sinapses com pesos associados **w**, um somador para somar as entradas **X** e uma função de ativação *f(v)*.

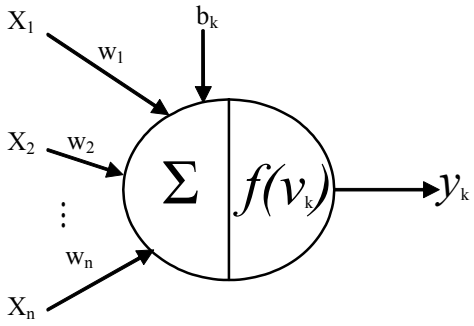


Figura 2.10 – Modelo de um neurônio Perceptron.

O *bias* b_k é utilizado para deslocar a função de ativação podendo servir para incrementar ou decrementar *f(v)*, definindo a posição da função com relação ao eixo da ordenada. Matematicamente, o neurônio pode ser representado da seguinte forma:

$$u_k = \sum_{j=1}^{n} w_{kj} x_j \qquad 2.19$$

$$v_k = u_k + b_k \qquad 2.20$$

$$y_k = f(v_k) \qquad 2.21$$

onde y_k é o sinal de saída do neurônio.

2.3.2. Funções de Ativação

A função de ativação *f(v)* define a saída do neurônio em termos do potencial de ativação *v*. Existem vários tipos de função de ativação [46], entretanto, três tipos de

função de ativação serão apresentados: a função de ativação limiar, a função de ativação linear por partes e a função de ativação sigmoidal.

Função de Ativação Limiar

A função de ativação do tipo limiar, mostrada na Figura 2.11, é descrita matematicamente pela equação 2.22:

$$f(v) = \begin{cases} 1 \ se \ v \geq 0 \\ 0 \ se \ v < 0 \end{cases} \qquad 2.22$$

Correspondentemente, a saída do neurônio k, empregando a função limiar é expressa por,

$$y_k = \begin{cases} 1 \ se \ v_k \geq 0 \\ 0 \ se \ v_k < 0 \end{cases} \qquad 2.23$$

Onde v_k é o potencial de ativação do neurônio, dado por

$$v_k = \sum_{j=1}^{m} w_{kj} x_j + b_k \qquad 2.24$$

Nesse modelo, a saída do neurônio assume valor 1 se o potencial de ativação do neurônio é não-negativo e zero caso contrário.

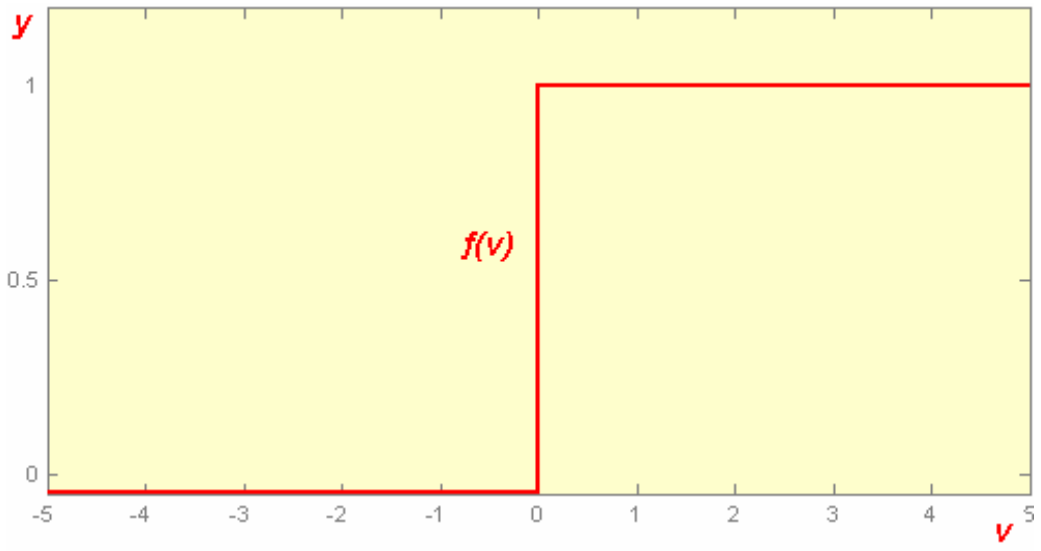


Figura 2.11 – Exemplo de função de ativação limiar

Função de Ativação Linear por Partes

A função linear pode ser restringida para produzir valores constantes em uma determinada faixa [-λ,+λ], neste caso a função passa a ser a função rampa como é mostrado na Figura 2.12 e descrito por

$$f(v) = \begin{cases} 1 & se \quad v \geq +1 \\ v & se \quad +1 > v > -1 \\ 0 & se \quad v \leq -1 \end{cases} \qquad 2.25$$

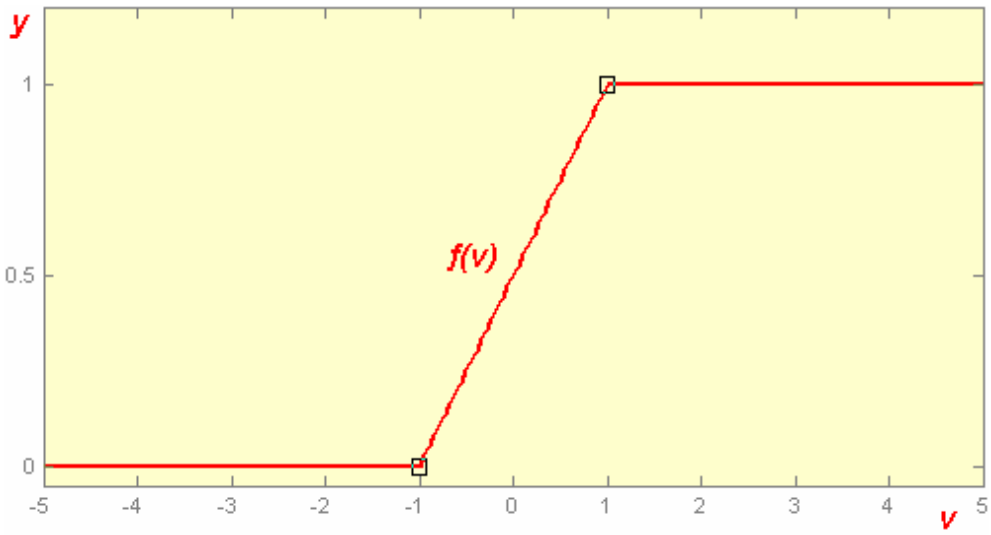


Figura 2.12 – Exemplo de função de ativação linear por partes

Função de Ativação Sigmoidal

A forma dessa função de ativação se assemelha a um "*S*", como se pode ver na Figura 2.13. Este tipo de função é o mais utilizado na construção de RNA. A função é definida como função estritamente crescente que exibe um interessante balanço entre o comportamento linear e o comportamento não-linear. Um exemplo de função sigmoidal é a função logística, definida por

$$f(v) = \frac{1}{1 + \exp(-av)} \qquad 2.26$$

onde *a* é o parâmetro de declividade da função sigmoidal.

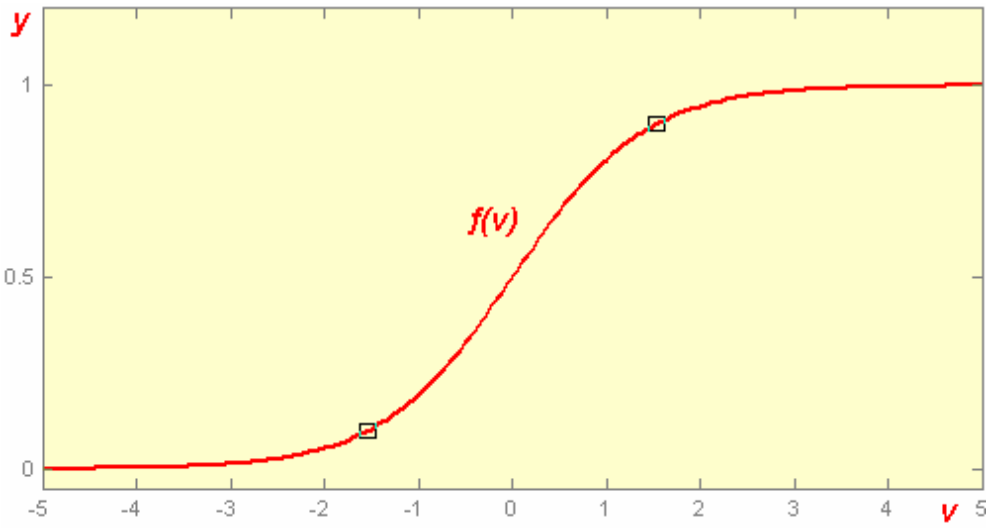


Figura 2.13 – Exemplo de função de ativação sigmoidal.

Variando o valor de *a* é possível obter funções sigmoidais com diferentes declividades. Aproximando-se o valor de *a* ao infinito, a função sigmoidal torna-se simplesmente uma função limiar. Enquanto a função limiar assume valor 0 ou 1, uma função sigmoidal assume um intervalo contínuo de valores de 0 e 1, o que torna a função sigmoidal diferenciável.

2.3.3. Arquitetura de RNA

A forma como os nós se interligam numa estrutura de rede é denominada por arquitetura ou topologia. Existem inúmeros tipos de arquiteturas de RNA, cada um com suas próprias potencialidades. Em geral, podem ser classificadas dentro de três categorias [46]:

Redes *Single-Layer Feedforward*

Esta arquitetura de RNA é a forma mais simples de rede em que os neurônios são organizados em camadas. A rede é composta por uma camada de entrada, cujos valores de saída são fixados externamente, e por uma camada de saída constituída de neurônios, que são nós computacionais, como mostra a Figura 2.14. Na realidade a entrada da rede não é contada como camada devido ao fato de nesta não se efetuarem quaisquer cálculos.

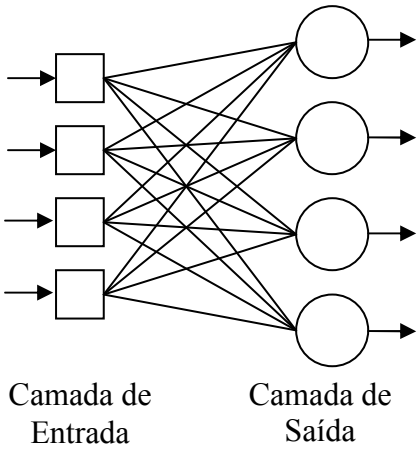


Figura 2.14 – Rede *Single-Layer Feedforward*

As topologias de camada única se limitam à aproximação de funções linearmente separáveis. A solução de problemas não linearmente separáveis envolve uso de uma ou mais camadas intermediárias.

Redes *Multilayer Feedforward*

A segunda classe de redes *feedforward* distingue-se pelo fato de possuir uma ou mais camadas intermediárias, como pode ser visto na Figura 2.15. A camada de entrada é responsável pela recepção do vetor de variáveis, com as informações a serem processadas pelo restante da rede. A segunda camada, que é denominada camada intermediária, pode ser única ou com várias subcamadas em cascata, é responsável pelo processamento da informação recebida pela camada de entrada. O processamento é realizado de forma paralela, passando pelos seus diversos neurônios. A seguir, tem-se a camada de saída, responsável pela fase final do processamento e apresentação dos resultados.

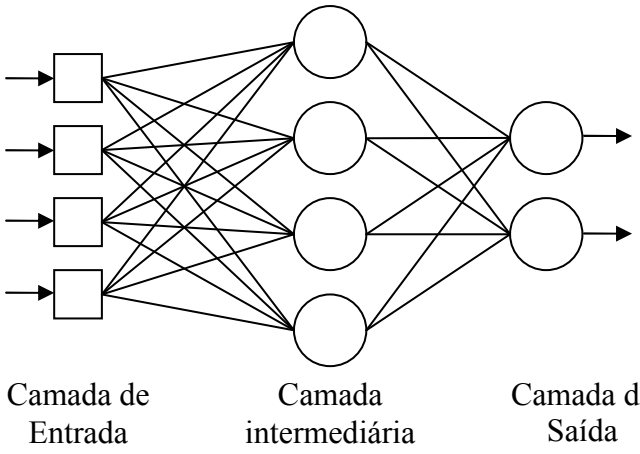


Figura 2.15 – Rede MLP típica com uma camada intermediária.

Redes Recorrentes

Uma rede neural recorrente difere de uma RNA *feedforward*, pelo fato de possuir pelo menos um laço de realimentação (*feedbackloop*). Uma rede recorrente pode consistir de uma única camada de neurônios, em que cada neurônio alimenta seu sinal de saída de volta para as entradas de todos os outros neurônios, conforme ilustra a Figura 2.16.

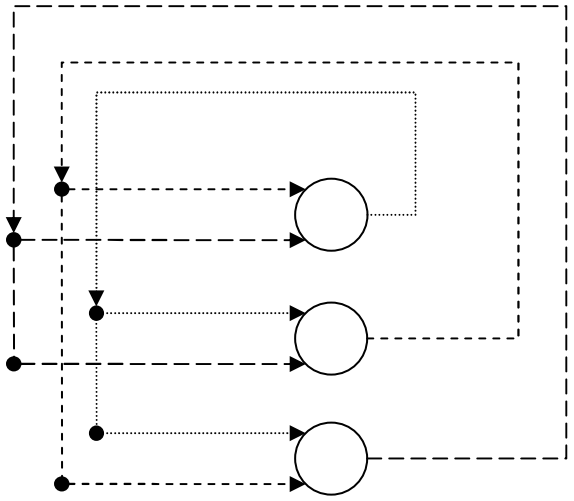


Figura 2.16 – Exemplo de Rede Recorrente.

2.3.4. Aprendizagem

Uma propriedade importante de uma rede neural artificial é o processo de aprendizagem ou de treino, onde o peso das conexões são ajustados de forma a se atingir um objetivo desejado ou estado de conhecimento da rede. Esta é a visão dos algoritmos tradicionalmente utilizados, como é o caso do algoritmo *feed forward back-propagation*, embora também seja possível modificar a topologia (ou estrutura) interna da RNA, criando uma auto-adaptação, através da eliminação ou criação de novas sinapses.

A aprendizagem de uma RNA envolve as seguintes etapas [46]-[47]:

- Estimulação da RNA pelo ambiente. Isto está relacionado às entradas da rede;
- Os pesos das conexões são ajustados em função do estímulo externo;

- A RNA responde ao estímulo do ambiente de uma nova forma, em virtude da alteração da sua estrutura interna;
- É realizada uma análise para avaliar o nível de aprendizado da RNA.

O aprendizado da rede utiliza algoritmos de aprendizado que dependerão de como a rede opera e se relaciona com o seu ambiente. Esta situação leva aos diversos tipos de algoritmos de treinamento, que se distinguem entre si pelo paradigma e regras de aprendizado que utilizam.

2.3.5. Paradigmas de Aprendizagem

Existem fundamentalmente três paradigmas de aprendizagem [46]-[47]:

Supervisionado – Trata-se, talvez, do algoritmo mais popular, pois sua essência é o do aprendizado por comparação do tipo certo ou errado avaliado por um especialista, comparado a um "professor" em algumas publicações. Neste tipo de aprendizado a rede aprende a partir de um conjunto de informações, ou padrão **P**, onde cada padrão **P**, também chamado de exemplo ou caso de treino, é composto de um vetor de entrada $\mathbf{X_p}$ e de um vetor de resposta ou saída $\mathbf{X_s}$. A comparação entre o valor esperado de resposta $\mathbf{T_p}$ com o valor de saída da rede $\mathbf{X_s}$ é realizada durante o processo de aprendizagem. Esta comparação gera um erro, o qual serve para ajustar os pesos da rede e, através de um processo iterativo, o processo é executado até que o erro esteja dentro de uma tolerância esperada, ou seja, até que se obtenha a convergência do método.

De reforço – Assume-se a presença de um instrutor, ou um especialista que tem domínio sobre o problema, embora a resposta correta não seja apresentada à rede. Dessa forma, são fornecidas apenas indicações sobre a veracidade da resposta, tal que a rede deve utilizar esta informação para melhorar a sua eficácia. Isto é realizado através do ajuste de um peso, que pode ser uma penalidade no caso de uma resposta errada ou um prêmio no caso de uma resposta correta.

Não supervisionado – No aprendizado não supervisionado não são fornecidos padrões externos sobre a resposta correta. Dessa forma, a aprendizagem é executada pela descoberta de características nos dados de entrada, adaptando-os às regularidades

estatísticas ou agrupamentos de padrões dos exemplos de treino. Um exemplo típico são as redes de Kohonen.

2.3.6. Regras de Aprendizagem

Para completar o treino tem-se, além dos paradigmas de aprendizagem, as regras de aprendizagem, as quais divide-se em cinco tipos básicos [46]:

- Hebbian – Esta regra de aprendizagem é a mais clássica e antiga, e está relacionada a um postulado de inspiração neurobiológica de aprendizagem proposto por Hebb em 1949 [47]. Para a utilização em RNA, esta regra foi dividida em duas partes:
    - Se dois neurônios em cada lado da conexão são ativados de forma simultânea, então a força, ou rigidez, dessa conexão é progressivamente aumentada;
    - Se dois neurônios em cada lado de uma conexão são ativados de forma assíncrona, então a conexão é progressivamente enfraquecida ou eliminada.

  Esta conexão é chamada de sinapse de Hebbian, tendo uma relação de dependência temporal e espacial. Esta regra é muito utilizada em aprendizagem do tipo não supervisionada. As redes auto-associativas, (como Hopfield) [47], ou hetero-associativas (memória associativa bidirecional) [47] são exemplos de redes que utilizam esta regra de aprendizagem.

- Competitiva – As saídas dos neurônios da mesma camada competem entre si para se tornarem ativas, sendo que apenas um neurônio é ativo por vez. Neste tipo de aprendizado o processo inicia com pesos pequenos e desiguais, sendo que no passo seguinte são apresentados padrões a rede, tal que, o neurônio que responder melhor do que os outros é premiado com um reforço no seu peso, por isto, está técnica também é chamada vencedor – leva – tudo. Existem situações em que pesos vizinhos podem também ser reforçados. Os mapas auto-organizáveis (redes SOM) e Kohonen utilizam este tipo de regra.

- Estocástica – Neste tipo de regra os pesos são ajustados considerando um modo probabilístico adequado [47].

- Baseada na Memória – Está baseada no armazenamento da totalidade (ou na maioria) das experiências passadas, as quais são explicitamente em uma memória de pares entrada-saída. Quando surge um novo vetor, $\mathbf{X_p}$, que não tenha sido anteriormente apresentado à rede, o algoritmo dispara uma procura de vetores na vizinhança de $\mathbf{X_p}$. Este algoritmo possui duas componentes:
    - Um critério de definição da vizinhança local de $\mathbf{X_p}$;
    - E uma regra de aprendizagem aplicada aos exemplos de treino da vizinhança local de $\mathbf{X_p}$.

  Um exemplo deste tipo de aprendizagem são as redes *Radial-Basis Function (RBF)*.

- Gradiente Descendente – No processo de aprendizagem supervisionado é utilizado a estratégia de diminuir o erro entre o valor esperado e o valor efetivo de saída, através da monitoração do erro no processo iterativo, buscando um ajuste que originará a busca da melhor resposta dentro de um erro aceitável para a convergência do processo. Basicamente, o processo visa minimizar uma função custo, ξ, definida em termos do sinal de erro ε^p. De acordo com a regra delta, o ajuste $\Delta\omega$ aplicado aos.pesos sinápticos ω a cada passo do processo pode ser definida pela equação 2.27:

$$\Delta\omega = \eta\nabla\xi \qquad 2.27$$

Onde η é a taxa de aprendizagem, representada por uma constante positiva, e $\nabla\xi$ é o gradiente da função custo. Este tipo de aprendizagem é muito utilizado e está associado a *MLP* e ao algoritmo *back-propagation* [47].

2.3.7. Algoritmo *Resilient Back-propagation*

O aprendizado de uma RNA é feito através de um algoritmo, cuja função é modificar os pesos sinápticos da rede de uma forma ordenada, para alcançar o objetivo de projeto desejado [46], como descrito nas seções 2.3.5 e 2.3.6.

Segundo Cybenko [48], uma rede com uma única camada intermediária pode aproximar qualquer função contínua. A utilização de duas camadas intermediárias permite a aproximação de qualquer função [49]. Esse tipo de rede é chamado *perceptron* multicamadas, ou *multilayer perceptron* (*MLP*), são redes que apresentam pelo menos uma camada intermediária.

Em uma RNA, o conhecimento é representado pelos seus pesos sinápticos, podendo-se determiná-los através de processos de aprendizado iterativos com base em um conjunto de dados amostrais. Diferentes arquiteturas e tipos de RNA podem ser utilizados para a construção da rede. Neste trabalho, utilizou-se uma rede *feed forward multilayer perceptron (MLP)*, o algoritmo utilizado para treinamento foi o *Resilient back propagation* [50], que é uma variação do algoritmo *back propagation* [46].

O *back propagation* é um algoritmo de aprendizado supervisionado que se baseia na regra delta proposta por Widrow e Hoff [51]. O objetivo é determinar um vetor de parâmetros **w** que minimize o erro quadrático sobre um dado conjunto de treinamento entre a saída atual e a saída desejada. O treinamento ocorre em duas fases, a fase *forward*, que é utilizada para obter uma saída para um dado padrão mostrado na entrada, e a fase *backward*, onde se utiliza a saída desejada e a saída atual para atualizar os pesos das conexões entre os nós. O método do gradiente descendente calcula a derivada parcial do erro da rede com respeito aos pesos sinápticos e através da atualização dos pesos minimiza o erro da rede em relação à função desejada.

O algoritmo *resilient back propagation* se propõe a eliminar a influência indesejável do tamanho da derivada parcial no cálculo do ajuste de pesos da rede. Essa influência prejudicial ocorre porque quando a saída de um neurônio é próxima de zero (ou um) e a saída desejada é um (ou zero), a derivada é próxima de zero, levando a um ajuste mínimo de pesos. Por essa razão, o *Rprop* utiliza somente o sinal da derivada para indicar a direção do ajuste de pesos. O quanto será modificado o peso é determinado pelo valor $\Delta w_{ij}^{(t)}$ [50], como mostra a equação 2.28:

$$\Delta w_{ij}^{(t)} = \begin{cases} -\Delta_{ij}^{(t)}, & \text{if } \dfrac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ +\Delta_{ij}^{(t)}, & \text{if } \dfrac{\partial E^{(t)}}{\partial w_{ij}} < 0 \\ 0, & \text{else} \end{cases} \qquad 2.28$$

onde $\dfrac{\partial E^{(t)}}{\partial w_{ij}}$ representa a derivada parcial do erro em relação aos pesos.

Para determinar os valores $\Delta_{ij}^{(t)}$ das novas atualizações é utilizado um processo de adaptação dependente do sinal [52][53]:

$$\Delta_{ij}^{(t)} = \begin{cases} \eta^{+} \cdot \Delta_{ij}^{(t-1)}, & \text{if } \dfrac{\partial E^{(t-1)}}{\partial w_{ij}} \cdot \dfrac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ \eta^{-} \cdot \Delta_{ij}^{(t-1)}, & \text{if } \dfrac{\partial E^{(t-1)}}{\partial w_{ij}} \cdot \dfrac{\partial E^{(t)}}{\partial w_{ij}} < 0 \\ \Delta_{ij}^{(t-1)}, & \text{else} \end{cases} \qquad 2.29$$

onde $0 < \eta- < 1 < \eta+$.

A regra de aprendizado do algoritmo *Rprop* funciona da seguinte forma, quando a derivada parcial correspondente a um peso mantém seu sinal, isso indica que a última atualização diminuiu o erro, então o valor de atualização é incrementado pelo fator η+ para acelerar a convergência. Quando o sinal da derivada parcial muda, isso indica que o último ajuste foi muito grande, e o valor de atualização é decrementado pelo fator η-, trocando a direção de atualização dos pesos.

2.3.8. Tarefas de Aprendizagem

A escolha da topologia da RNA e do método de aprendizagem tem uma influência direta pela tarefa de aprendizagem que deve ser desempenhada pela RNA, isto é, qual a aplicação, ou tipo de problema, que se espera que a RNA solucione. Esta avaliação conduz a sete categorias principais [46]:

- Memória Associativa
- Diagnóstico
- Reconhecimento de padrões
- Regressão/previsão
- Controle
- Otimização
- Filtragem/Compressão de dados

A próxima seção apresenta métodos alternativos para tratar o problema abordado por esta dissertação.

2.4. Outros Métodos

Algumas técnicas alternativas como Lógica *Fuzzy*, Matriz de Pertinência e Árvores de Decisão [5][6] foram previamente utilizadas numa tentativa de criar um sistema capaz de identificar, baseado nas evidências coletadas em campo, as causas mais prováveis que levaram a um desligamento não programado. No entanto, devido à falta de uma base apropriada de informações sobre desligamentos não programados, se tornou difícil desenvolver um sistema computacional baseado em dados históricos. Por esse motivo, nesta seção, será feita uma introdução de Lógica *Fuzzy* e Árvores de Decisão.

2.4.1. Lógica *Fuzzy*

Conjuntos *Fuzzy* foram pela primeira vez introduzidos em 1965 por Lotfi A. Zadeh [54] como uma extensão da noção clássica de conjuntos. Na teoria clássica, um elemento pode pertencer ou não a um conjunto. Isso não ocorre na teoria de conjuntos *Fuzzy* proposta por Zadeh. Em contraste, ele propôs que um conjunto seria caracterizado por uma função de pertinência, que associa os elementos a este conjunto dentro de uma faixa de pertinência entre zero e um. Então, quanto mais próximo da unidade está essa função de pertinência, maior é o grau de pertinência desse elemento ao conjunto. No

entanto, não se deve confundir esse grau de pertinência com probabilidades, existe uma diferença conceitual, a função *Fuzzy* representa o grau de pertinência a um conjunto, não a probabilidade de um evento ou condição. Numa abordagem probabilística leva-se em consideração a probabilidade de uma variável estar em determinado estado e não o quanto essa variável está nesse estado.

A Lógica *Fuzzy* é baseada na teoria dos Conjuntos *Fuzzy*. Tradicionalmente, uma proposição lógica tem dois extremos: ou "completamente verdadeiro" ou "completamente falso". Entretanto, na Lógica *Fuzzy*, uma premissa varia em grau de verdade de zero a um, o que leva a ser parcialmente verdadeira ou parcialmente falsa. Os conjuntos são rotulados qualitativamente – usando-se termos lingüísticos, tais como: alto, morno, ativo, pequeno, perto, etc. – e os elementos destes conjuntos são caracterizados variando o grau de pertinência.

Um Conjunto *Fuzzy* é definido em um universo de discurso (conjunto base) X, e caracterizado pela sua função de pertinência:

$$A: X \rightarrow [0,1]$$

onde A(X) representa o grau com que X pertence a A e expressa a extensão com que x se enquadra na classe representada por A.

Uma função de pertinência particular pode ser visualizada por meio da Equação 2.30. Como se pode constatar, esta função é triangular e as variáveis a, b e c são os parâmetros da função.

$$\mu(x) = \begin{cases} \dfrac{x-a}{b-a} \; se \; x \in [a,b) \\ \\ \dfrac{c-x}{c-b} \; se \; x \in [b,c] \\ \\ 0 \;\; caso \;\; contrário \end{cases} \qquad 2.30$$

A Figura 2.17 mostra a forma da função triangular definida acima.

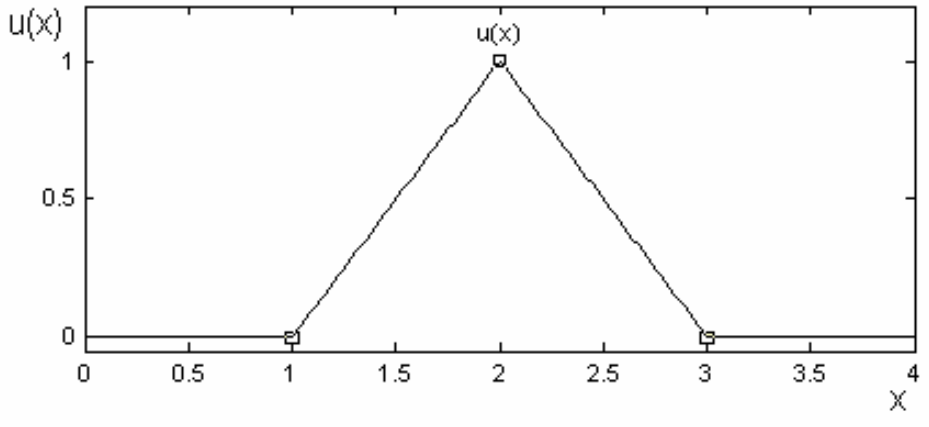


Figura 2.17 – Função de pertinência *Fuzzy*.

Conforme definida anteriormente, a teoria dos Conjuntos *Fuzzy* é uma extensão da teoria dos Conjuntos Tradicionais. Assim, as principais operações e relações entre Conjuntos *Fuzzy* são definidas como extensão das operações e relações tradicionais.

2.4.2. Árvores de Decisão

Uma árvore de decisão [55] é uma estrutura de dados definida recursivamente como:

- um nó folha que corresponde a uma classe

ou

- um nó de decisão que contém um teste sobre algum atributo. Para cada resultado do teste existe uma aresta para uma subárvore. Cada subárvore tem a mesma estrutura que a árvore.

A Figura 2.18 é um exemplo de árvore de decisão para o diagnóstico de um paciente. Na figura, cada elipse é um teste em um atributo para um dado conjunto de dados de pacientes. Cada retângulo representa uma classe, ou seja, o diagnóstico. O diagnóstico é realizado começando-se pela raiz e seguindo cada teste até que uma folha seja alcançada.

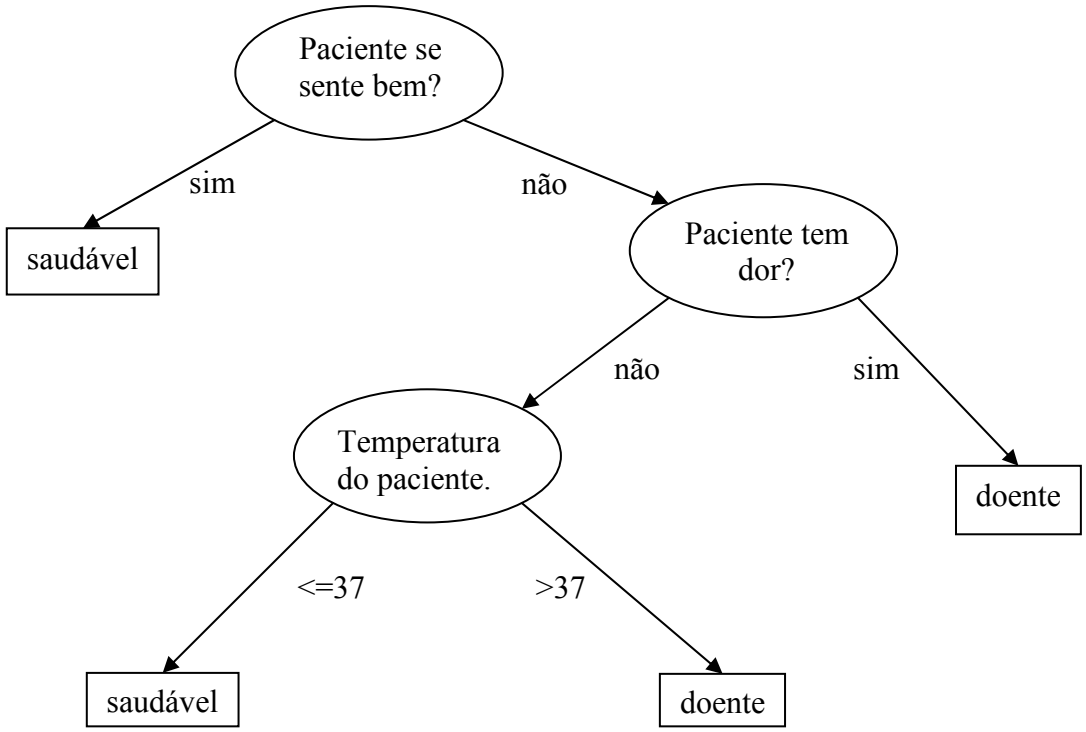


Figura 2.18 – Exemplo de árvore de decisão para diagnóstico de um paciente.

A árvore também pode ser representada por um conjunto de regras. Como as regras que representam uma árvore de decisão são disjuntas, isto é, apenas uma única regra dispara quando um novo exemplo é classificado, uma forma alternativa de representar tais regras consiste em escrever uma regra separadamente para cada nó folha, iniciando pela raiz, conseqüentemente nenhum *else* é necessário:

se paciente se sente bem = sim **então**
    classe = saudável
fim se

se paciente se sente bem = não **and** paciente tem dor = não **and** temperatura do paciente <= 37 **então**
    classe = saudável
fim se

se paciente se sente bem = não **and** paciente tem dor = não **and** temperatura do paciente > 37 **então**
    classe = doente
fim se

se paciente se sente bem = não **and** paciente tem dor = sim **então**

classe = doente

fim se

2.5. Extração de Conhecimento

A aprendizagem de Redes Bayesianas pode ser realizada a partir de uma amostra de dados do domínio de interesse, podendo-se assim retirar as distribuições de probabilidades e as relações de interdependência entre as variáveis. Por isso, para que o diagnóstico fornecido por um sistema que utilize essa abordagem seja confiável, é de extrema importância que o domínio de dados utilizado seja o mais completo possível.

A base de dados disponível para treinamento da rede apresentava campos incompletos, informações contraditórias entre outros dados que tornava a base inapropriada para representação do conhecimento sobre o domínio necessário aos propósitos deste trabalho, esta situação é descrita em detalhes na seção 3.3. Por esse motivo, se fez necessária a busca por uma técnica que pudesse adequar esse domínio para o treinamento das heurísticas propostas.

A utilização de informações armazenadas em bases de dados para compor padrões úteis de informação tem sido um constante desafio para a área de TI. Técnicas como mineração de dados, extração de conhecimento, descoberta de informação, processamento de padrões de dados são alguns termos utilizados para este tipo de tarefa. Uma forma de fazer uso adequado de dados existentes, guardando as características necessárias para o fim que se propõe, é a utilização das técnicas de Extração de Conhecimento em Banco de Dados, comumente referenciado pela sigla inglesa *KDD* [56].

2.5.1. Conceitos Básicos sobre *KDD*

A técnica de *KDD* é um processo de identificação de padrões de dados válidos, não-triviais, potencialmente úteis e compreensíveis [56]. O processo de *KDD* é apresentado na Figura 2.19, e consiste em uma seqüência iterativa dos seguintes passos [57]:

1. **Limpeza dos dados**: para remover ruído e dados irrelevantes.

2. **Integração dos dados**: onde fontes de dados múltiplos podem ser combinadas.

3. **Seleção dos dados**: onde dados relevantes para a análise são recuperados do banco de dados.

4. **Transformação dos dados**: onde os dados transformados ou consolidados no formato apropriado para mineração.

5. **Mineração de dados**: é um **processo** onde métodos inteligentes são utilizados a fim de extrair padrões de dados.

6. **Avaliação e Representação do conhecimento**: onde técnicas de visualização e representação de conhecimento são utilizadas para apresentar o conhecimento minerado para o usuário.

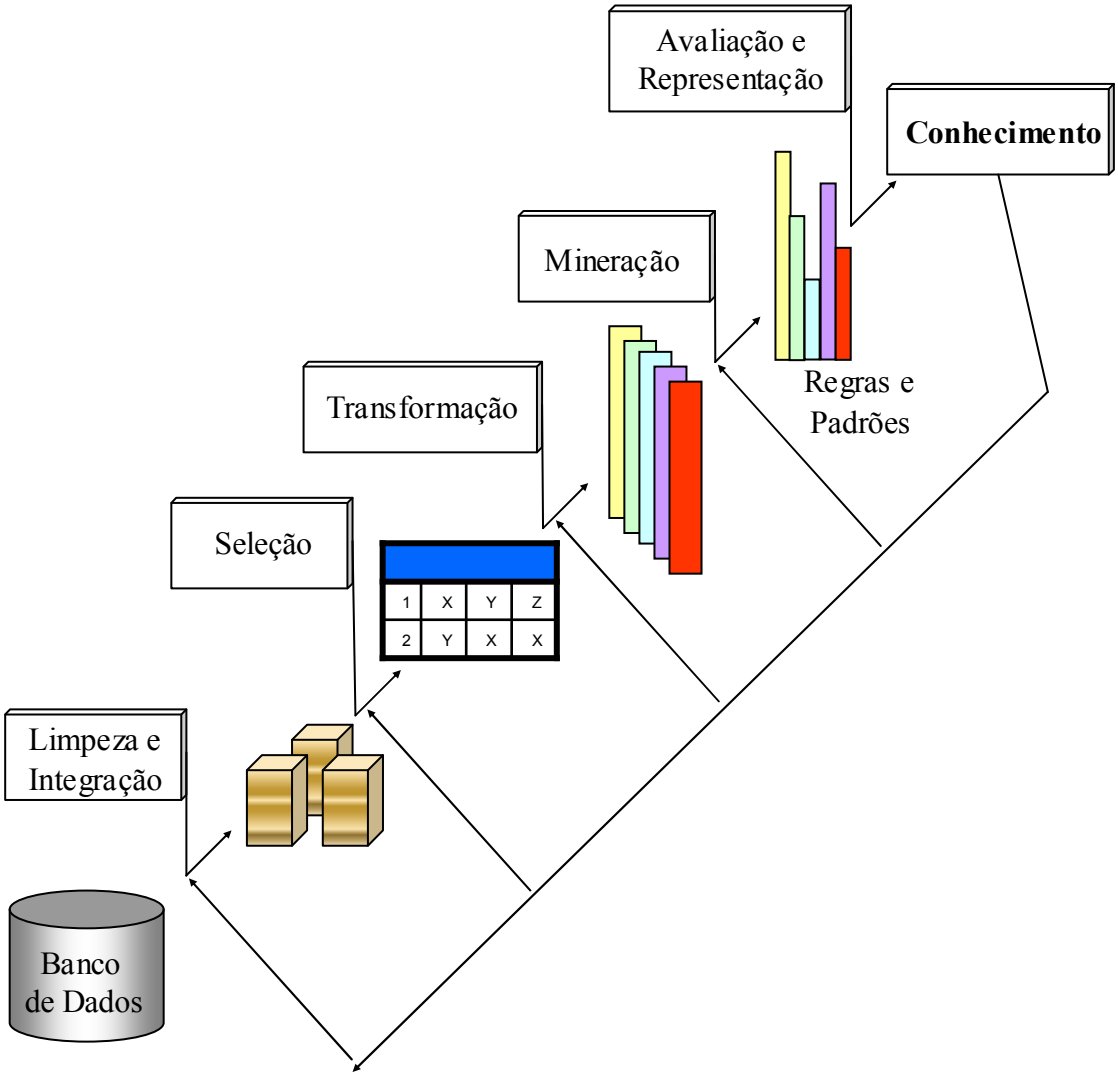


Figura 2.19 – Processo de *KDD*.

A etapa de mineração de dados irá interagir com o usuário ou com uma base de conhecimento. Os padrões extraídos são então armazenados como novo conhecimento numa base de dados construída para receber estes dados.

Alguns autores consideram *KDD* e Mineração de Dados como processos distintos [56]. Entretanto, em algumas bibliografias, o termo mineração de dados tornou-se mais popular que o *KDD* e é utilizado quando se refere ao processo de identificação de padrões a partir de grandes quantidades de dados armazenados em banco de dados ou outro tipo de banco de armazenamento [57].

Os principais componentes de um típico sistema de mineração de dados são apresentados na Figura 2.20:

1. **Banco de dados**: pode ser somente um ou um conjunto de banco de dados, planilhas, ou outro tipo de banco de informações.

2. **Servidor de banco de dados**: o servidor de banco de dados é o responsável por localizar e carregar os dados relevantes, baseado nas consultas do usuário.

3. **Base de conhecimento**: é o domínio de conhecimento utilizado para guiar a busca, ou avaliar o interesse em algum padrão resultante. Esses dados são a entrada do motor de inferência e dessa base será extraído o conhecimento de interesse para o sistema.

4. **Motor (*engine*) de mineração de dados**: é o responsável pela inferência com base nas informações obtidas na base de conhecimento.Busca padrões de interesse através de técnicas de mineração de dados tais como regras de associação, regras de classificação, padrões de seqüências, agrupamento (*clustering*), séries temporais, regras de produção, entre outras.

5. **Avaliação de padrões**: interage com o motor de mineração de dados com o objetivo de avaliar se os padrões ou regras gerados são interessantes à aplicação ou não.

6. **Interface com o usuário**: é o módulo de comunicação entre o usuário e o sistema de mineração de dados, permitindo que o usuário interaja com o sistema através de consultas e visualização da informação extraída.

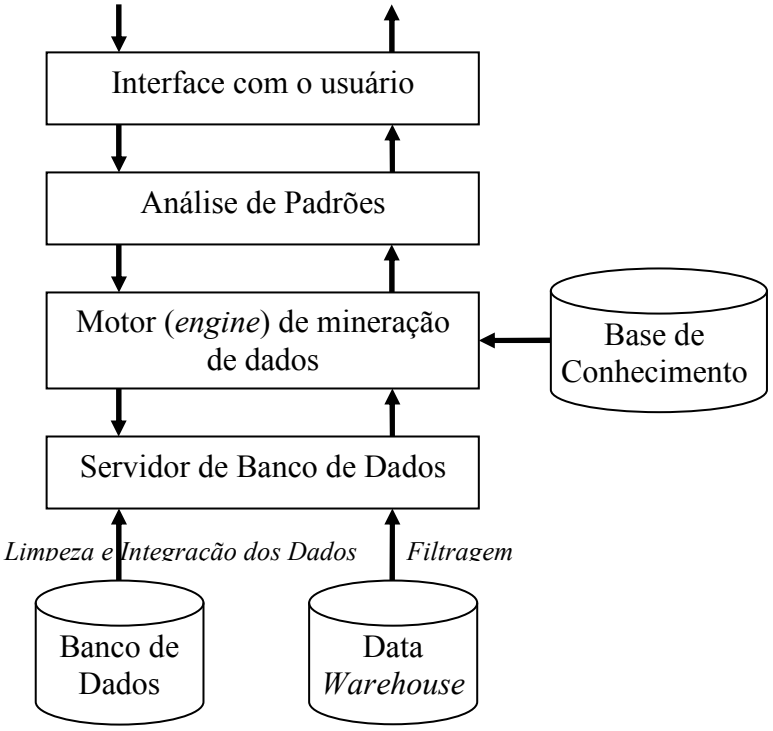


Figura 2.20 – Arquitetura típica de um sistema de mineração de dados

2.5.2. Extração de Padrões

A Mineração de Dados consiste em aplicar análise de dados e algoritmos que produzam padrões ou modelos a partir de outros dados [58]. Os primeiros sistemas baseados em conhecimento foram sistemas utilizando regras [59]. Nesses sistemas o processo de tomada de decisão humano é modelado por regras do tipo "*if P then Q*", simbolicamente $P \rightarrow Q$. Portanto, as regras podem expressar relacionamentos lógicos e equivalências de definições para simular o raciocínio humano [60].

A escolha da tarefa a ser executada no processo é feita de acordo com os objetivos desejáveis para a solução a ser encontrada. Os dois principais objetivos de mineração de dados são predição e descrição:

- Predição: envolve o uso de algumas variáveis nas bases de dados para predizer valores futuros ou desconhecidos de outras variáveis de interesse.
- Descrição: busca obter padrões que descrevam os dados e aprender uma hipótese generalizada, um modelo, a partir dos dados selecionados.

A Figura 2.21 mostra os tipos de tarefa que podem ser executadas.

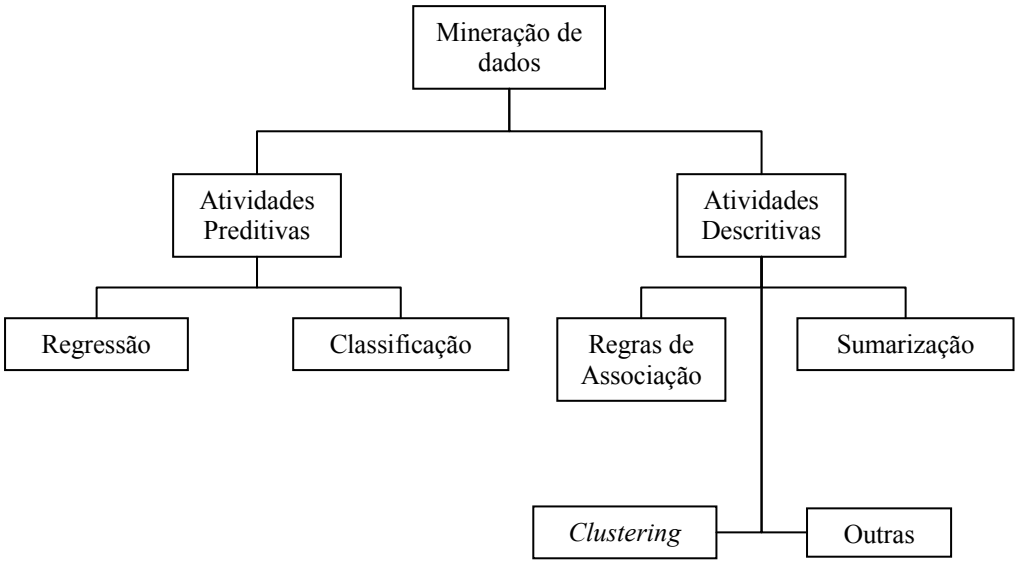


Figura 2.21 – Tarefas de Mineração de Dados

Atividades de predição consistem na generalização de exemplos ou experiências passadas com respostas conhecidas em uma linguagem capaz de reconhecer a classe de um novo exemplo desconhecido. Os dois principais tipos de tarefas para predição são classificação e regressão. Na classificação é feita a predição de um valor categórico ou discreto, já na regressão, o atributo a ser predito consiste em um valor contínuo [61].

Atividades de descrição consistem na identificação de comportamentos ou tendências intrínsecos do conjunto de dados, sendo que estes dados não possuem uma classe especificada. Algumas tarefas são *clustering*, regras de associação e sumarização [55].

Podem-se utilizar diferentes formas de produzir os padrões extraídos através da mineração de dados, a técnica utilizada irá depender do tipo de estrutura que se deseja produzir na saída dos dados. Neste trabalho, criou-se um conjunto de regras baseadas no conhecimento de especialistas sobre o domínio. Tendo como objetivo predizer variáveis que estão implícitas no banco de dados original, mas que podem ser extraídas ou inferidas utilizando técnicas de mineração de dados baseada em regras de classificação. Através desse procedimento criou-se uma base de conhecimento mais fiel ao domínio em estudo, tornando mais completa e consistente a base de treinamento de sistemas inteligentes.

2.5.3. Classificação Baseada em Regras

A classificação baseada em regras é uma técnica para classificar dados que utiliza um conjunto de regras "se...então...". As regras para o modelo são representadas de forma disjunta, ou seja, somente uma regra é disparada por vez, $R = (r_1 \vee r_2 \vee ...r_k)$, onde *R* é o conjunto de regras e r_i são as regras de classificação [62].

Tabela 2.4 – Exemplo de conjunto de regras para o problema de classificação de vertebrados [62].

r_1 : (Dá a luz = não) $\wedge$ (Criatura Voadora = sim) $\rightarrow$ Ave
r_2 : (Dá a luz = não) $\wedge$ (Criatura Aquática = sim) $\rightarrow$ Peixe
r_3 : (Dá a luz = sim) $\wedge$ (Temp. Corporal = sangue quente) $\rightarrow$ Mamífero
r_4 : (Dá a luz = não) $\wedge$ (Criatura Voadora = não) $\rightarrow$ Réptil
r_5 : (Criatura Aquática = semi) $\rightarrow$ Anfíbio

Cada regra de classificação pode ser expressa da seguinte forma:

$$r_i : (Condição_i) \rightarrow y_i \qquad 2.31$$

O lado esquerdo da regra é chamado antecedente da regra ou precondição e contém uma conjunção de testes de atributos:

$$Condição_i = (A_1 \ op \ v_1) \wedge (A_2 \ op \ v_2) \wedge ...(A_k \ op \ v_k) \qquad 2.32$$

onde (A_j, v_j) é um par atributo-valor e *op* é um dos seguintes operadores lógicos: $\{=, \neq, <, >, \leq, \geq\}$. Cada teste de atributo $(A_j \ op \ v_j)$ é conhecido como uma conjunção. O lado direito da regra é chamado conseqüente da regra, que contém a classe predita por y_i.

A classificação baseada em regras foi utilizada para implementação do motor de inferência do processo de *KDD* realizado neste trabalho. As regras utilizadas para classificação serão descritas em detalhes no item 3.3.4, que aborda a etapa de Mineração de Dados.

2.6. Conceitos Básicos de Confiabilidade em Sistemas de Distribuição de Energia

Esta dissertação está focada na identificação da causa de desligamentos não programados que ocorrem na rede de distribuição de energia. A correta identificação da causa de um desligamento pode tornar mais eficiente a alocação de recursos materiais e de equipes de trabalho, visando cobrir áreas que influenciam de forma significativa os índices de confiabilidade da empresa. Por esse motivo, alguns conceitos de confiabilidade, bem como a classificação das causas que levam a um desligamento não programado, devem ser abordados nesta seção. Esses dois temas são considerados requisitos para compreensão do problema em estudo.

2.6.1. Conceitos Básicos da Teoria da Confiabilidade

Entre os conceitos que estão envolvidos na teoria de confiabilidade está a noção de probabilidade, que pode ser entendida a partir da noção de experimento aleatório. Os conceitos básicos de teoria de probabilidades já foram introduzidos na seção 2.1. A seguir serão apresentados algumas definições básicas da teoria de confiabilidade [35]:

- **Confiabilidade**: probabilidade de um componente (aparelho, sistema, equipamento) cumprir suas funções pré-fixadas, dentro de um período de tempo desejado e sob certas condições operativas;

- **Capacidade de Trabalho**: propriedade do componente pela qual ele tem condições de cumprir as suas funções, segundo parâmetros estabelecidos na sua documentação técnica;

- **Vida Útil**: propriedade do componente em conservar sua capacidade de funcionamento, com as interrupções necessárias para reparos, até um valor limite fornecido na sua documentação técnica. Este limite pode ser de ruptura ou outro tipo de avaria, bem como um decréscimo do rendimento, na precisão, na potência, etc;

- **Prazo de Funcionamento**: período estipulado pela documentação técnica para o uso do componente dentro das especificações de projeto;

- **Qualidade**: entende-se por qualidade de um componente, o conjunto de propriedades que determinam seu grau de aptidão para o serviço a que ele for destinado;

- **Confiabilidade Estimada**: confiabilidade de determinado componente medida através de ensaios específicos efetuados segundo um programa de ensaios inteiramente definido;

- **Confiabilidade Prevista**: é a confiabilidade calculada a partir de um modelo matemático definido, levando-se em conta dados de projeto e da confiabilidade estimada de componentes, bem como condições operativas predeterminadas;

- **Confiabilidade Operacional**: é a confiabilidade observada durante a operação normal de componentes. Ela depende das condições reais de utilização, do meio ambiente, do pessoal de manutenção, etc;

- **Desempenho**: caracteriza o comportamento de um componente (aparelho, sistema, equipamento) perante os ensaios de recepção e também pelo seu comportamento no campo, tendo em vista que o componente pode estar sujeito a outras solicitações que não as especificadas;

- **Falha**: caracteriza o término do desempenho satisfatório de um determinado componente (aparelho, sistema, equipamento) fazendo com que esse não consiga executar sua função;

- **Defeito**: caracteriza uma imperfeição no estado do componente, que pode resultar em uma falha do próprio componente ou de um outro;

- **Saída**: o conceito de saída está intimamente ligado à idéia de o componente não estar disponível para a operação, sendo que este fato poderá ser causado por falha ou manutenção preventiva. Portanto pode-se dizer que nem toda saída é obrigatoriamente provocada por uma falha, porém toda falha acarreta em uma saída.

2.6.2. Índices de Confiabilidade

Para medir o desempenho e a confiabilidade de sistemas de distribuição de energia, existem alguns índices estabelecidos pela ANEEL [63], os quais serão apresentados a seguir:

- **Duração Equivalente de Interrupção por Unidade Consumidora (DEC)**: intervalo de tempo que, em média, no período de observação, em cada unidade consumidora do conjunto considerado ocorreu descontinuidade da distribuição de energia elétrica.

$$DEC = \frac{\sum_{i=1}^{k} Ca(i) \times t(i)}{Cc} \qquad 2.33$$

- **Freqüência Equivalente de Interrupção por Unidade Consumidora (FEC)**: número de interrupções ocorridas, em média, no período de observação, em cada unidade consumidora do conjunto considerado.

$$FEC = \frac{\sum_{i=1}^{k} Ca(i)}{Cc} \qquad 2.34$$

Onde,

Ca(i) = Número de unidades consumidoras atingidas em um evento (i), no período de apuração.

t(i) = Duração de cada evento *(i)*, no período de apuração.

i = Índice de eventos ocorridos no sistema que provocam interrupções em uma ou mais unidades consumidoras.

k = Número máximo de eventos no período considerado.

Cc = Número total de unidades consumidoras, do conjunto considerado, no final do período de apuração.

- **Duração de Interrupção Individual por Unidade Consumidora (DIC)**: intervalo de tempo que, no período de observação, em cada unidade consumidora ocorreu descontinuidade da distribuição de energia elétrica.

$$DIC = \sum_{i=1}^{n} t(i) \qquad 2.35$$

- **Freqüência de Interrupção Individual por Unidade Consumidora (FIC)**: número de interrupções ocorridas, no período de observação, em cada unidade consumidora.

$$FIC = n \qquad 2.36$$

- **Duração Máxima de Interrupção Contínua por Unidade Consumidora (DMIC)**: tempo máximo de interrupção contínua, da distribuição de energia elétrica, para uma unidade consumidora qualquer.

$$DMIC = t(i)\max \qquad 2.37$$

Onde,

i = Índice de interrupções da unidade consumidora ou do ponto de conexão, no período de apuração, variando de 1 a n.

n = Número de interrupções da unidade consumidora ou do ponto de conexão considerada, no período de apuração.

$t(i)$ = Tempo de duração da interrupção *(i)* da unidade consumidora ou do ponto de conexão considerada, no período de apuração.

$t(i)\max$ = valor correspondente ao tempo da máxima duração de interrupção *(i)*, no período de apuração, verificada na unidade consumidora ou no ponto de conexão considerado em horas e centésimos de horas.

2.6.3. Classificação das Causas de Desligamentos

Deve-se considerar como a causa de um desligamento o motivo primário que levou a interrupção do fornecimento de energia, uma vez que causas secundárias

originadas pela causa raiz podem "mascarar" a verdadeira causa e originar uma ação de manutenção e operação inadequada.

É importante que a empresa tenha um padrão de classificação das causas de interrupção, a fim de proporcionar uma maneira organizada de armazenar o histórico de desligamentos, como forma de orientar ações de operação e manutenção e subsidiar o planejamento da expansão, colaborando com os processos de engenharia e econômico-financeiros de alocação dos recursos nos ativos de rede.

Na literatura existem algumas classificações de causas bem definidas. Em [13] as causas são classificadas em função de sua natureza, como:

1. **Interrupção Programada** (*Schedule Interruption*) – Interrupção de fornecimento devido ao desligamento programado para manutenção preventiva ou construção.
2. **Perda de Suprimento** (*Loss of Supply*) – Interrupção de fornecimento devido a problemas no sistema supridor por diminuição da freqüência por aumento de carga, tensão fora dos limites aceitáveis de operação, transitório no sistema de transmissão ou excursão de freqüência.
3. **Contato com Árvores** (*Tree Contacts*) – Interrupção ocorrida por desligamentos causados pelo contato de árvores no circuito elétrico.
4. **Descarga Atmosférica** (*Lightning*) – Interrupção de fornecimento devido a descargas atmosféricas.
5. **Equipamento Defeituoso** (*Defective Equipment*) – Interrupção causada por falha de equipamentos devido ao tempo de uso ou manutenção incorreta.
6. **Clima Adverso** (*Adverse Weather*) – Interrupção de fornecimento de energia causada por chuva, gelo, tempestades, neve, vento, temperaturas extremas, ou outra condição climática extrema.
7. **Ambiente Adverso** (*Adverse Environment*) – Interrupção causada por exposição dos equipamentos a condições anormais, como maresia, contaminação industrial, umidade, corrosão, vibração, fogo ou enchente.
8. **Elemento Humano** (*Human Element*) – Interrupção de fornecimento devido à interferência de funcionários da empresa como uso incorreto de

equipamentos, instalação ou construção incorreta, erro de configuração de proteção, erro de manobras de chaveamento, estragos propositais.

9. **Interferência Externa** (*Foreign Interference*) – Interrupção de energia devido a fatores fora do controle da concessionária, como pássaros e animais diversos, veículos, escavações, vandalismo, sabotagem e objetos estranhos.

10. **Outras** (*Unknow/Other*) – Interrupções sem uma causa ou razão aparente que possa ter contribuído para o desligamento.

A referência [64] cita as causas mais comuns de desligamentos em redes (aéreas ou subterrâneas) de distribuição de energia, como mostra a Figura 2.22. O autor ainda cita os animais que mais prejudicam as redes de distribuição no hemisfério norte (EUA e Canadá), considera que é importante levar em consideração a taxa de crescimento de cada tipo de planta que divide espaço com a rede, e ainda mostra um gráfico comparando a contribuição de cada causa em três distribuidoras de energia dos EUA.

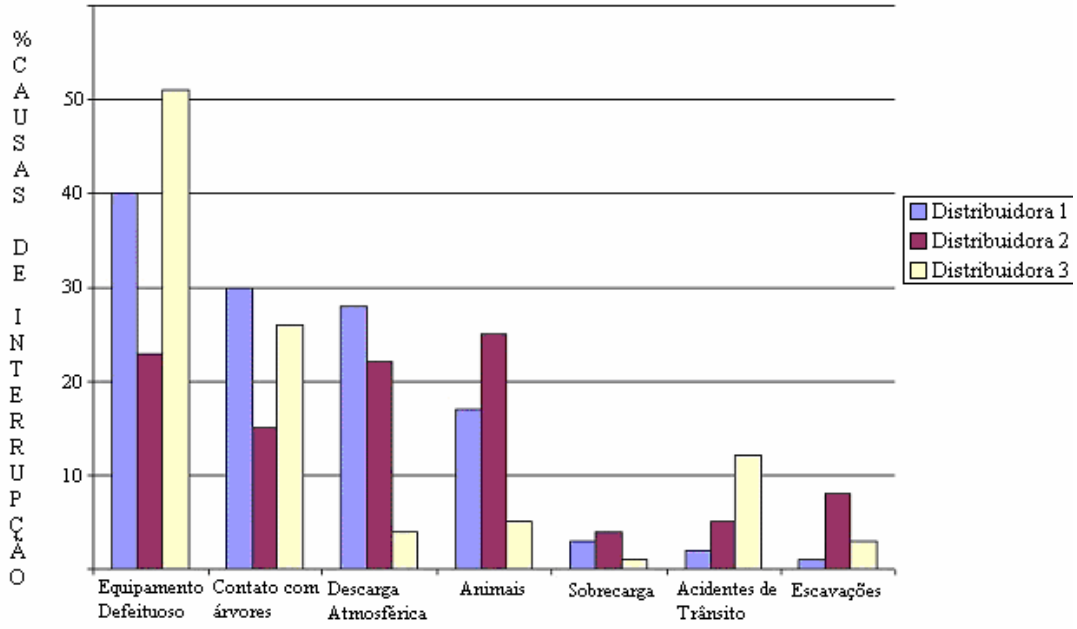


Figura 2.22 – Causas de desligamento.[64]

Na Figura 2.22 percebe-se que escavações aparecem como uma causa significativa para a distribuidora dois. Esta causa é tipicamente um problema de redes subterrâneas. Quando se trata de redes aéreas, ela pode ficar implícita em Interferência Externa, como o autor em [13] a classificou.

3. Metodologia Desenvolvida

A metodologia desenvolvida tem por objetivo tratar informações a partir de bases de dados pré-existentes, isto é, bases de dados que não foram projetadas com um fim específico e orientadas para algum aplicativo computacional.

Para os desenvolvimentos, utiliza-se uma base de dados real, na qual são armazenados milhares de ocorrências vinculadas a desligamentos não programados na rede de distribuição de energia. Após analisar cuidadosamente os campos e o conteúdo da base disponível, verificou-se que a mesma apresentava campos incompletos, informações contraditórias, campos sem dados e muitos outros problemas. Esses fatores colaboraram para concluir que a base, no estado inicial, era inapropriada para utilização direta. Dessa forma, o processo de *KDD* foi escolhido para regularizar os problemas existentes.

As próximas seções descrevem as etapas que envolvem o processo de *KDD* utilizado na metodologia. A descrição do trabalho segue a lógica do desenvolvimento aplicada a este trabalho, onde algumas decisões de cunho prático que envolvem experiência de campo são consolidadas junto a profissionais que atuam diretamente na rede ou na operação de sistemas de distribuição. Algumas etapas são mais pormenorizadas devido a sua importância no processo geral. Um exemplo pode ser observado na etapa de mineração de dados, a qual será apresentada no item 3.3.4 com uma descrição mais detalhada.

3.1. Introdução

No sistema desenvolvido por Pretto [2], os eletricistas utilizam um software desenvolvido para *PDA (Personal Digital Assistant)* na coleta de informações a respeito do local do evento de desligamento. Este é considerado o primeiro módulo do sistema. Uma visão geral do sistema pode ser visto na Figura 3.1. O segundo módulo está dividido em sub-módulos: um para armazenamento e outro para tratamento posterior de eventos de desligamentos não programados. Esta base de dados presente no segundo módulo pode servir como fonte de informação para estudos estatísticos e análises probabilísticas, permitindo a correta identificação das causas mais prováveis de um desligamento não programado.

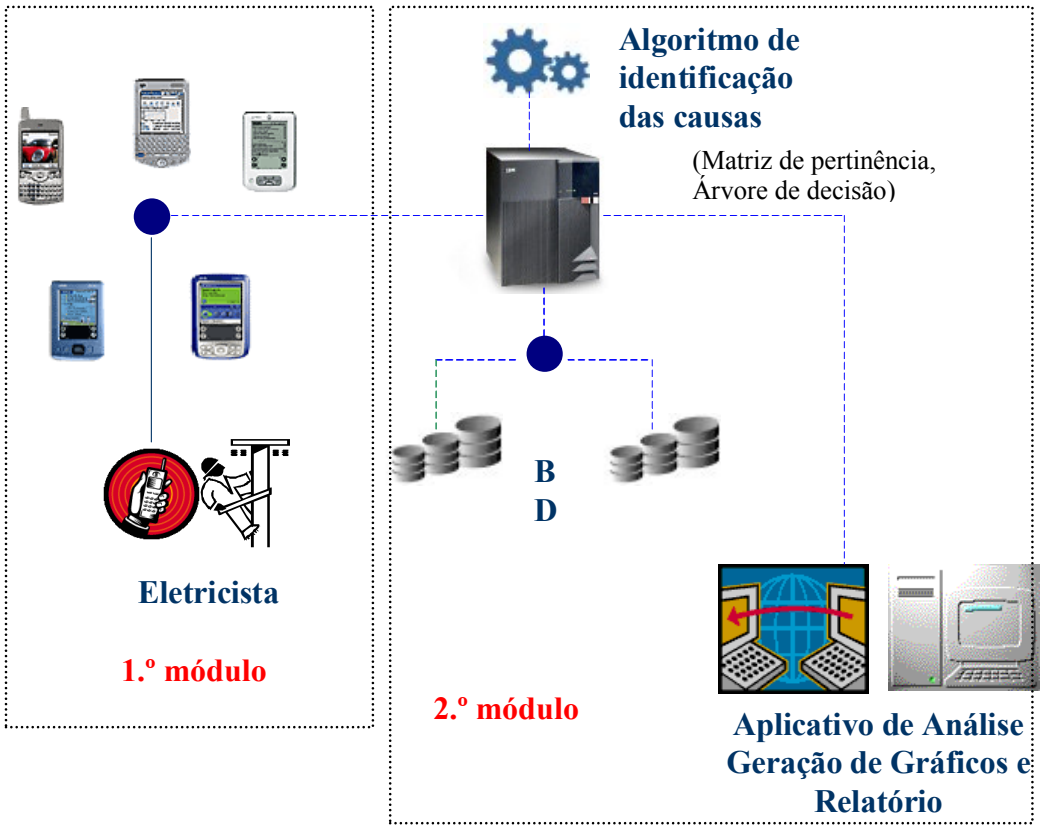


Figura 3.1 – Esquema do trabalho desenvolvido por Pretto [2].

Neste trabalho o primeiro módulo do sistema, como apresentado em [5] e [6], está ausente. O banco de dados utilizado como histórico de desligamentos já existia na empresa e foi construído com as informações coletadas pelos eletricistas através da

metodologia convencional utilizando planilhas preenchidas manualmente, ao longo de um ano, portanto o sistema de coleta de Pretto não foi utilizado como proposto.

Os bancos de dados das empresas concessionárias costumam ser inconsistentes e com elevado grau de incertezas, o tratamento dessas informações é a abordagem proposta nesta dissertação. Isso é causado principalmente pela ausência de uma metodologia adequada na coleta dos dados no local do evento de desligamento e também pelas próprias características dos sistemas de energia, que apresentam muitas variáveis difíceis de serem observadas. A base de dados utilizada, contendo informações sobre o histórico de eventos de desligamentos é a disponível na empresa, onde os dados brutos coletados pelos eletricistas não sofreram nenhum tipo de tratamento, não estando organizados de uma forma adequada para o aprendizado de máquina.

Dessa forma, fez-se necessário realizar um processo de extração de conhecimento a partir desses dados, tornando a base de conhecimento adequada ao treinamento de sistemas inteligentes. Através de um processo de extração de conhecimento em banco de dados, ou *Knowledge Discovery in Databases* (*KDD*), foi possível obter a partir de um banco de dados inconsistente, incompleto e sem uma padronização adequada, um conjunto de dados representando o conhecimento necessário para realizar o aprendizado e mantendo o mesmo formato de representação do conhecimento utilizado em [2]. Esse processo de extração de conhecimento será descrito em detalhes neste trabalho.

Os métodos de identificação utilizados no segundo módulo do sistema foram substituídos por uma Rede Bayesiana e uma Rede Neural Artificial. No entanto, para que o treinamento das duas redes utilizadas como método de identificação pudesse ser feito adequadamente, os dados deveriam antes ser organizados pelo processo de *KDD*. A Figura 3.2 mostra o fluxo de informações do sistema proposto.

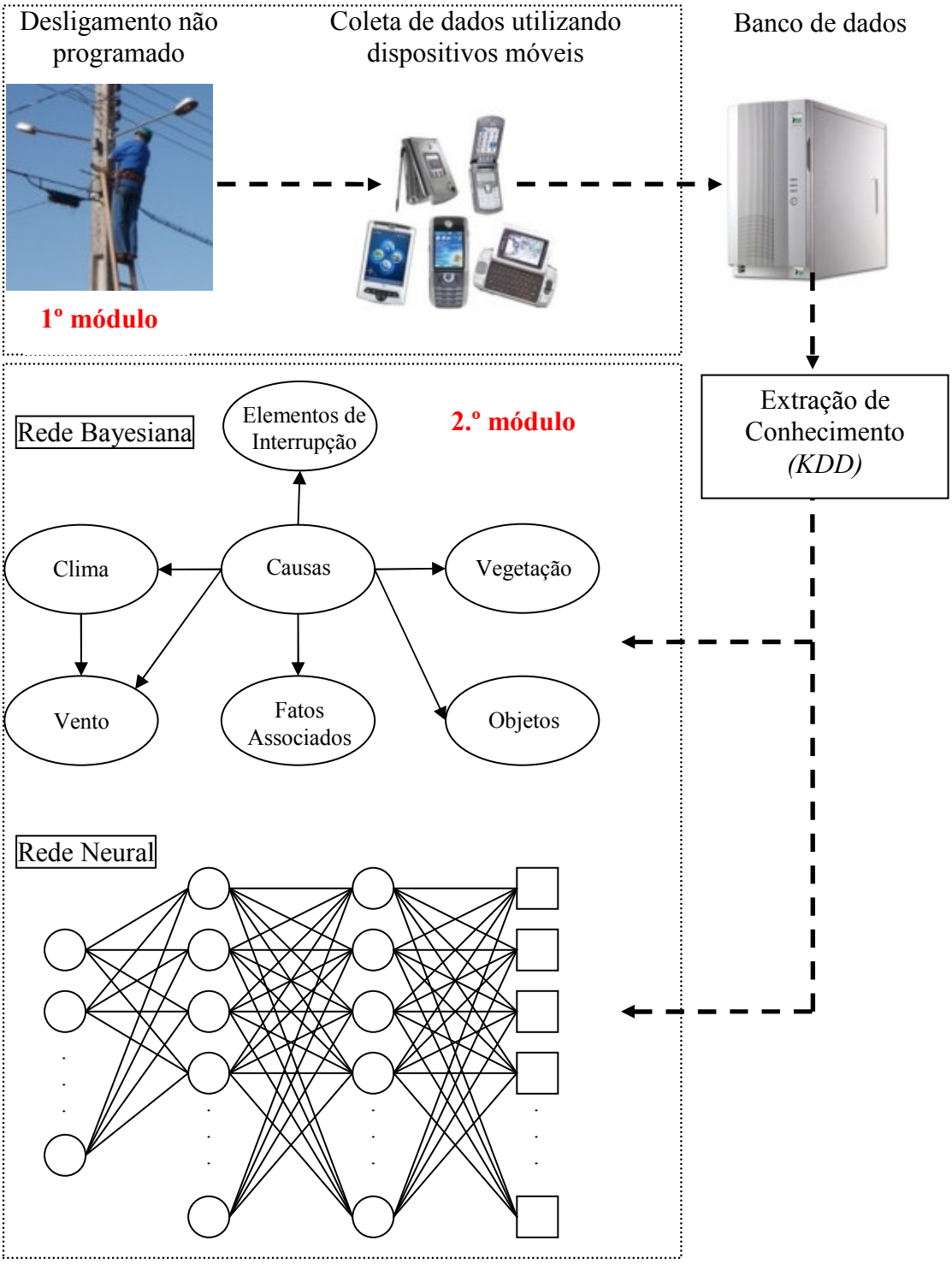


Figura 3.2 – Fluxo de informações do sistema.

Com base na observação das variáveis relacionadas a um evento de desligamento é possível utilizar raciocínio lógico para se obter um diagnóstico da causa provável de um desligamento na rede de distribuição. No entanto, muitas vezes existe um grau de incerteza associado ao o que é observado ou simplesmente nada pode ser observado a respeito de uma determinada variável, o que pode gerar entradas de dados

incompletos num banco de dados sobre eventos de desligamento. Nesse caso, a utilização de RB e RNA são uma boa opção como método de diagnóstico de falhas, já que ambas heurísticas são apropriadas para lidar com manipulação de dados sob incerteza e/ou que apresentam campos ausentes na base de dados.

As seções seguintes deste Capítulo apresentam de forma detalhada a metodologia desenvolvida.

3.2. Definição das Variáveis em Estudo

3.2.1. Definição das Causas

Para definir as causas de eventos que causam desligamentos, e que serão utilizadas neste trabalho, foram analisadas e tomadas como base a classificação de causas da empresa RGE, mostrado na Tabela 3.1 e as referências [2], [13] e [64], onde as causas são agrupadas de acordo com a natureza do evento e sua contribuição para a ocorrência da falta de energia.

Tabela 3.1 - Codificação de Causas para Falta de Energia

Código	Descrição
	Grupo1 - Intervenções Programadas
101	Alteração para Melhorias
102	Alteração para Ampliação
201	Manutenção Preventiva
202	Manutenção Corretiva
203	Manobra prévia para Desligamento Programado
204	Necessidade de Operação
	Grupo 2 - Meio Ambiente
301	Poluição
303	Descarga Atmosférica
304	Vegetal
306	Animais/Pássaros
307	Vento
308	Erosão
309	Inundações
	Grupo 3 - Terceiros
401	Pandorga
402	Vandalismo
403	Abalroamento de Postes
404	Redes de Comunicação (telefonia / TV a cabo ...)
405	Bola na Rede Elétrica
406	Objetos Estranhos na Rede (cordas, tapetes, arames, ferr
407	Queimadas / Incêndio
408	Defeito Interno Clientes
409	Danos causados por terceiros
410	Serviços públicos
	Grupo 4 - Outros
501	Sobrecarga
506	Defeito Transitório
507	Coordenação da Proteção
508	Corte de Carga
509	Improcedente
510	Baixa Tensão de Fornecimento

Código	Descrição
	Grupo 5 - Falha em Componentes
601	Isolador
602	Para-Raios
604	Condutor
605	Capacitor
606	Religador
607	Regulador
608	Seccionalizador
610	Conexões(todo ponto de ligação)
612	Medidor
613	Elo Fusível Queimado
614	Ramal de Ligação
616	Chave Fusível
617	Chave Faca
618	Chave à Óleo / SF6
619	Condutor Desregulado
620	Emenda
621	Poste
622	Poste Podre
623	Cruzeta
624	Ferragens
625	Disjuntor
626	TP - Transformador de Potencial
627	TP - Transformador de Corrente
628	Gerador
629	Serviços Auxiliares
630	Transformador
631	Proteção da SE
632	Padrão de Entrada
633	Amarrilho
	Grupo 6 - Falha Humana
701	Falha de Operação
703	Falha de Construção
704	Falha de Projeto
705	Falha de Manutenção
706	Falta de Manutenção

Dessa forma, as classes de causa de desligamento definidas para este trabalho serão as seguintes:

1. **Falha no Componente** - Interrupção de fornecimento causada por falha do componente devido aos mais diversos motivos;
2. **Sobrecarga** - Interrupção causada por sobrecarga do sistema. Em dias de muito consumo, pode ocorrer a abertura de uma chave em função do carregamento;
3. **Clima Adverso** - Interrupção de fornecimento de energia causada por chuva, granizo tempestades, ventos, temperaturas extremas, ou outras condições climáticas adversas que possam influenciar num desligamento;
4. **Descarga Atmosférica** - Interrupção causada por uma descarga atmosférica na rede elétrica;
5. **Interferência do Meio Ambiente** - Interrupção ocasionada por intervenção do meio ambiente, objetos presos na rede elétrica, animais, pássaros, entre outros;
6. **Vegetal** - Interrupção causada por influência de árvores e galhos de árvores perto da rede elétrica;
7. **Interferência Humana** - Quando ocorrem desligamentos ocasionados por vandalismo, furto ou outra interferência humana é atribuída esta causa. Interrupções ocasionadas por empresas que trabalham perto da rede, obras, empresas de telefonia entre outras, também são atribuídas a este item;
8. **Acidente** - Interrupções ocasionadas por acidentes envolvendo veículos;
9. **Incêndio/Queimada** - Interrupções ocasionadas por incêndio ou queimada perto da rede elétrica.

Na próxima seção é apresentado o estudo das variáveis que serão utilizadas para compor o sistema de identificação de causa de desligamentos proposto.

3.2.2. Definição das Variáveis

Em [13] foi desenvolvido um sistema para armazenamento de dados sobre interrupção de energia para cálculo do desempenho da rede. Neste sistema, para cada

evento ocorrido, os dados são armazenados de forma a vincular o desligamento a uma das sete grandes categorias de equipamentos:

1. Linha de Distribuição;
2. Cabo de Distribuição;
3. Transformador de Distribuição;
4. Transformador de Potência;
5. Chaves;
6. Reguladores;
7. Capacitores.

Neste trabalho, as variáveis que associam os fatos ocorridos e utilizadas para compor o sistema de identificação de causas de interrupção de energia foram as seguintes:

1. Elementos de Interrupção;
2. Clima;
3. Vento;
4. Fatos Associados;
5. Vegetação;
6. Objetos.

A seguir descreve-se cada uma destas variáveis de forma detalhada.

Elementos de Interrupção

Baseado nessa classificação, foi criado em [2] o conceito de Elemento de Interrupção, que indica em que parte da estrutura física da rede ocorreu a falha. Os elementos de interrupção são os seguintes:

1. Postes;
2. Equipamentos;
3. Condutores;
4. Isoladores;
5. Cruzeta.

Clima

As condições climáticas e o vento têm influência direta nas falhas que ocorrem nos equipamentos da rede elétrica. Os estados selecionados para essas variáveis foram adaptados de [64] para a realidade da região sul do Brasil:

1. Chuva;
2. Temporal;
3. Neblina;
4. Neve/Granizo;
5. Bom.

Vento

1. Excessivo;
2. Moderado;
3. Sem Vento.

Fatos Associados

Informações sobre atividades que são realizadas próximas à rede elétrica e que podem influenciar em um desligamento. Os fatos associados a desligamentos estão descritos abaixo:

1. Acidente;
2. Queimada/Incêndio;
3. Empresa trabalhando;
4. Vandalismo;
5. Inundação;
6. Erosão.

Vegetação

De acordo com [64], contato com árvores é uma das três causas mais comuns de interrupção de energia, portanto a observação das condições da vegetação no local do evento é muito importante. Os estados elencados são os seguintes:

1. Sem Poda;
2. Podada.

Objetos

É comum a ocorrência de desligamentos ocasionados por objetos presos na rede elétrica, portanto essa variável deve ser observada no local do desligamento. Nesse caso é indicada somente a presença ou não de objetos presos na rede. Os estados possíveis são:

1. Sim;
2. Não.

3.3. Tratamento dos Dados

A concessionária de energia cedeu para análise 570.409 registros de desligamentos não programados ocorridos no período entre abril de 2005 a maio de 2006. Esta concessionária atende aproximadamente um milhão de consumidores em sua rede de distribuição nas tensões de 13,8 kV e 33 kV, possui 62 subestações de distribuição, totalizando 80.910 km de extensão de rede de distribuição.

As tabelas de dados, AEVEN e EVENT, fornecidas em formato Excel, foram armazenadas num banco de dados Oracle®, dessa forma foi possível valer-se de uma ferramenta mais poderosa para tratamento dos dados. A consulta ao banco Oracle® foi realizada utilizando uma versão demo do software PL/SQL Developer®.

O tratamento dos dados será feito seguindo as etapas descritas anteriormente no processo de *KDD*, não necessariamente na ordem exposta.

3.3.1. Seleção dos dados

O primeiro passo na utilização da técnica de *KDD* foi à seleção dos campos de maior interesse oriundos do banco de dados original, os quais poderiam auxiliar no processo de mineração dos dados.

Os campos selecionados para o estudo estão representados de forma simplificada na Tabela 3.2.

Tabela 3.2– Campos selecionados para estudo

NUM_1	X_CORD	Y_CORD	DEV_TYPE_NAME	CLIMA	CAUSA

- NUM_1: Número da Ordem de Serviço que atendeu o desligamento;
- X_CORD: Coordenada X (*UTM*) do local onde ocorreu o desligamento;
- Y_CORD: Coordenada Y (*UTM*) do local onde ocorreu o desligamento;
- DEV_TYPE_NAME: Equipamento que foi afetado no desligamento;
- CLIMA: Clima no momento do desligamento;
- CAUSA: Causa apontada segundo a coleta do eletricista em campo, seguindo o padrão da empresa.

Os campos X_CORD e Y_CORD encontravam-se na tabela EVENT e os demais campos na tabela AEVEN. O campo EID, comum às duas tabelas, foi utilizado como chave primária para a consulta ao banco de dados, ver Figura 3.3, tornando possível a união de todas informações desejadas em uma única tabela chamada de EVENTOS.

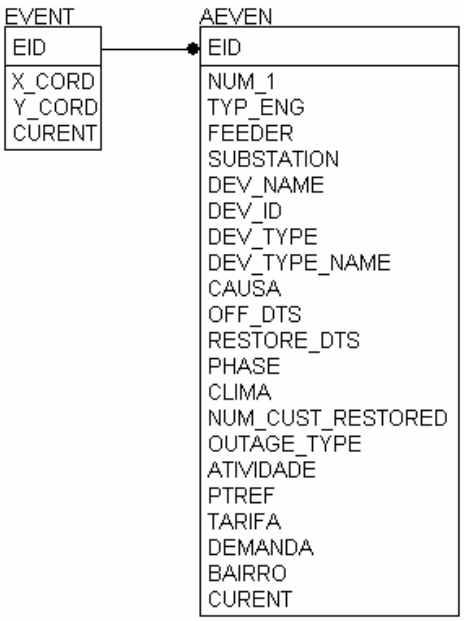


Figura 3.3 – Tabelas EVENT e AEVENT

3.3.2. Limpeza e Integração dos Dados

De acordo com a Figura 2.19, a etapa de limpeza e integração dos dados forma a primeira fase do processo de *KDD*. Utilizando o histórico de eventos fornecido pela empresa, identificou-se a estrutura relacional aplicada ao banco de dados através do seu diagrama de entidades e relacionamentos (ER), e verificaram-se os tipos de ferramentas necessárias para proceder à etapa em questão. As consultas ao banco de dados foram realizadas utilizando a linguagem *Structured Query Language* (*SQL*). Dessa forma, foi possível extrair da estrutura original as variáveis definidas na seção anterior, abandonando a estrutura original e adotando uma nova estrutura relacional, adequada ao armazenamento das variáveis de interesse. Na Figura 3.4 apresenta-se um exemplo descritivo de uma seqüência de filtros criados com o objetivo de limpar e integrar as variáveis de interesse na nova estrutura utilizada.

Assim, do conjunto original de eventos de desligamentos armazenados no banco de dados, constatou-se que apenas uma pequena porcentagem desses dados se adequava ao domínio de estudo. Portanto, grande parte desses registros foi descartada da nova base de dados. Conforme a Figura 3.4, os primeiros registros a serem eliminados da base de eventos foram os que apontavam como causa um desligamento que havia sido programado pela companhia distribuidora de energia. Cita-se como exemplo, desligamentos para manutenção preventiva, alteração para melhorias, desligamento a pedido do cliente, entre outros. Em seguida, foram retirados os registros contendo os valores ‘000’, por ser um valor inexistente na lista de códigos de causa da companhia e valores do tipo NULL, indicando que o campo não foi preenchido pelo eletricista na hora do atendimento da ordem de serviço.

Desta forma, foi possível separar apenas registros de desligamentos não-programados, os quais interessam ao estudo em questão. Porém, havia a necessidade de limpar e integrar este novo conjunto de dados, o qual apresentava problemas, tais como: condições contraditórias, como por exemplo, a indicação da causa “Descarga Atmosférica” em clima “Bom”. Nesse ponto, buscou-se identificar e remover contradições registradas. A Figura 3.4 apresenta uma visão da ordem de grandeza (quantificação) de dados eliminados do conjunto inicial utilizando os critérios descritos acima. Após essa limpeza e integração, dos 570.409 registros fornecidos, apenas 12% dos dados foram considerados aderentes aos eventos registrados para serem utilizados

no sistema, totalizando 69.222 registros. O número reduzido de dados qualificados presentes nesta base de dados – nesse caso em torno de 12% ressalta a importância do processo de coleta de informações desses eventos. Em [3] são expostos alguns exemplos de sistemas de coleta de dados de interrupções, em que a qualidade da informação é o foco principal. Neste ponto, é importante elaborar uma reflexão acerca dos registros coletados em campo. Uma das maiores preocupações em [3] foi retirar a responsabilidade de apontamento de uma causa de interrupção do eletricista (coletor), pois observou-se que durante a recomposição da rede, muitas vezes não é possível identificar diretamente uma causa, e a grande maioria das vezes, os eletricistas são direcionados para recompor rapidamente a interrupção, com o objetivo de mitigar indicadores de duração e freqüência.

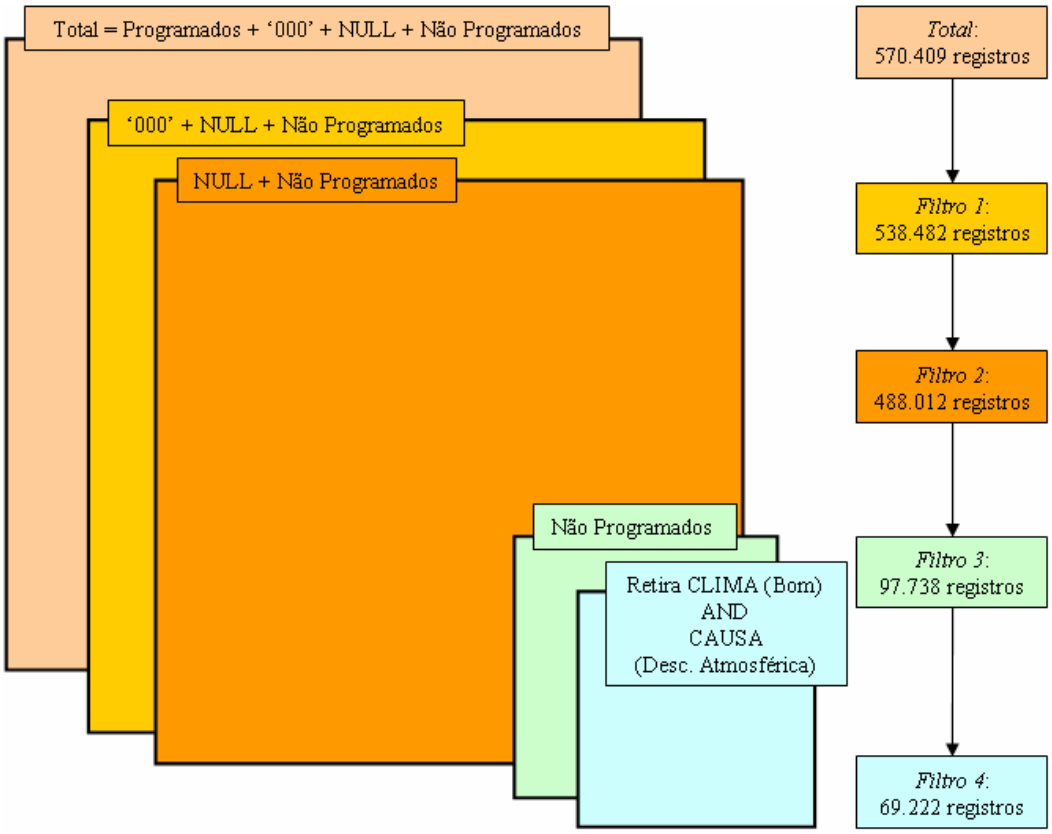


Figura 3.4 – Filtragem dos dados

Promovendo uma nova análise dos dados após a primeira fase de Limpeza e Integração, observou-se que grande parte das interrupções foi registrada com o campo "CLIMA" preenchido como "bom". A Figura 3.5 apresenta essa característica. Após examinar cuidadosamente os registros, e verificar os procedimentos de operação seguidos pela empresa, constatou-se que a nova divisão continha uma tendência registrada, ou seja, durante a recomposição das redes, o campo "CLIMA" da ficha de

recomposição preenchida pelo eletricista era um campo ignorado, sendo posteriormente preenchido com seu valor padrão. Obviamente, essa análise deveria considerar essa condição.

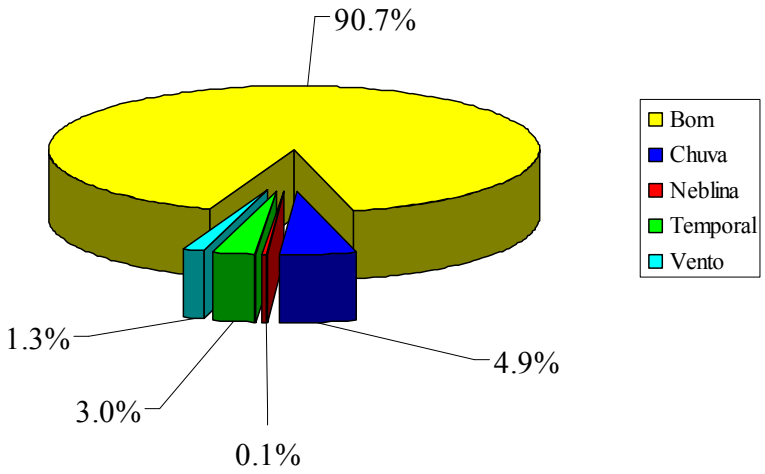


Figura 3.5 – Desligamentos considerados válidos

O fato verificado na nova base de dados contraria uma afirmação de Brown em [64], onde o autor afirma que em clima adverso as interrupções são mais freqüentes que em clima bom, para a grande maioria das empresas de distribuição de energia. Com base nessa referência e nos procedimentos de operação da distribuidora, constatou-se que a maioria dos registros de interrupções de interesse a este trabalho, contidos na base de dados, deveria ocorrer em clima adverso, fato que conduziu a uma correção da tendência verificada e apresentada na Figura 3.5. Assim, com intuito de tornar a base de dados mais fiel à realidade da operação dos sistemas de distribuição, foi realizada mais uma fase de Limpeza e Integração dos Dados, com a adição dos seguintes critérios:

- 70% da nova base são compostas por informações pertencentes às condições de clima adverso, conforme apresentado na Figura 3.6;
- 30% da nova base são compostas pelo resultado da Limpeza e Integração de Dados proposta para os 90.7% dos dados vinculados à categoria de “Clima Bom”, conforme apresentado na Figura 3.6.

Os dados filtrados foram divididos em um conjunto contendo eventos ocorridos durante clima bom, com registros não tendenciosos e outro conjunto em clima adverso, como ilustra a Figura 3.6.

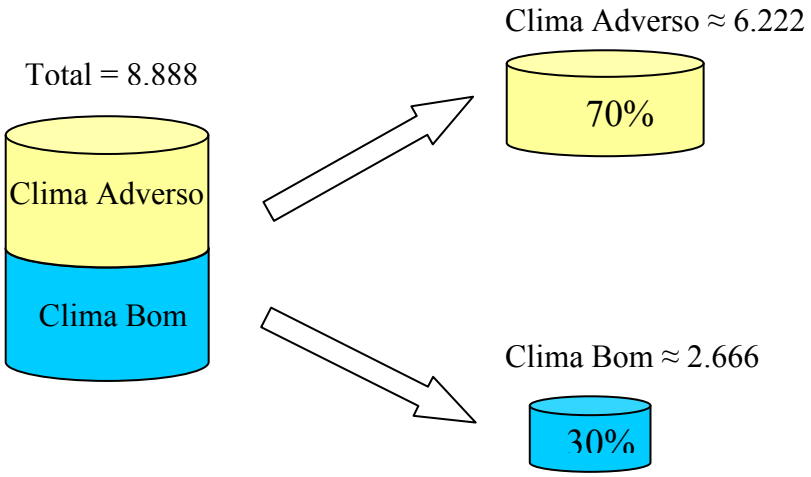


Figura 3.6 – Conjuntos de desligamento em clima bom e clima adverso

Dessa forma, a redução proposta para a base de dados traduz uma forte aderência aos aspectos de identificação das variáveis definidas para o problema. A nova base foi acomodada em uma estrutura com as seguintes quantificações: 8.888 eventos registrados. É importante observar que a base iniciou com 570.409 registros e passou para 8.888 eventos de interrupções com condições de utilização.

3.3.3. Transformação dos dados

Após a etapa de Limpeza e Integração, os dados foram organizados em um formato adequado para facilitar sua utilização, ou seja, foram aplicadas regras vinculadas à etapa de Transformação dos Dados, onde regras escritas em uma linguagem de programação pudessem acessar os campos escolhidos de forma a facilitar a Transformação dos dados desejada. Um exemplo do formato definido para esta etapa pode ser visualizado na Tabela 3.3.

Tabela 3.3 – Formato de entrada do motor de inferência

DEV_TYPE_NAME	CLIMA	CAUSA	X_CORD	Y_CORD
Chave Fusível	Temporal	Poluição	40227267	679542682
Transformador	Chuva	Poluição	57937709	677247941
Chave Fusível	Temporal	Descarga Atmosférica	46971602	676555018
Chave Fusível	Temporal	Descarga Atmosférica	56973869	678813192

Esse conjunto de regras, que tem como objetivo não somente a inferência de campos ausentes na base de dados, mas também a classificação correta dos eventos, sendo descrita em detalhes a seguir.

3.3.4. Mineração de dados (MD)

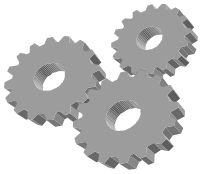
A etapa de Mineração de Dados consiste em extração de informação interessante, não trivial, implícita, previamente desconhecida e potencialmente útil nas informações armazenadas em grandes massas de dados [58]. Podem-se utilizar diferentes formas para produzir padrões extraídos através da mineração de dados, e cada uma irá ditar o tipo de técnica que será utilizada para produzir a estrutura de saída dos dados. Para essa tarefa existe uma grande variedade de técnicas que costumam ser utilizadas, tais como regras de associação, regras de produção, padrões de seqüências, agrupamento (*clustering*) e padrões em séries temporais. Neste trabalho, a extração de padrões (conhecimento) foi realizada através de regras de classificação, tendo em vista o formato de representação de conhecimento desejado para a formação do banco de dados necessário ao treinamento de sistemas inteligentes. Para uma melhor compreensão desta etapa, apresenta-se a seguir, em maior detalhe, as decisões que formam as regras de classificação utilizadas neste trabalho.

Motor de Mineração de Dados

Com o objetivo de obter-se uma boa base de conhecimento a partir de uma fonte de informação pobre, os dados foram tratados utilizando um conjunto de regras, a qual se denomina a motor de inferência do sistema. A Figura 3.6 mostra a informação na entrada da máquina de inferência e a informação desejada na saída, utilizada neste trabalho.

ENTRADA

DEV_TYPE_NAME	CLIMA	CAUSA	X_CORD	Y_CORD
Chave Fusível	Temporal	Poluição	40227267	679542682
Transformador	Chuva	Poluição	57937709	677247941
Chave Fusível	Temporal	Descarga Atmosférica	46971602	676555018
Chave Fusível	Temporal	Descarga Atmosférica	56973869	678813192



Motor de Mineração de Dados (regras)

SAÍDA

ELEMENTOS	FATOS	CLIMA	VENTO	OBJETOS	VEGETAÇÃO	CAUSAS

Figura 3.7 – Formato dos dados na entrada e na saída do motor de mineração de dados.

Onde:

- ELEMENTOS: elemento em que ocorreu a falha; possíveis estados: poste, equipamentos, condutores, isoladores, cruzeta;
- FATOS: fatos associados ao evento de desligamento que foram observados no local; possíveis estados: acidente, queimada/incêndio, empresa, vandalismo, inundação, erosão;
- CLIMA: condições climáticas no momento do desligamento; possíveis estados: bom, chuva, temporal, neve/granizo, neblina;
- VENTO: intensidade do vento no momento desligamento;
- OBJETOS: se existem objetos presos aos condutores; possíveis estados: sem vento, moderado, excessivo;

- VEGETAÇÃO: condições da vegetação; possíveis estados: sem poda, podada;
- CAUSA: causa do desligamento; possíveis estados: vegetal, descarga atmosférica, falha no componente, clima adverso, sobrecarga, interferência do meio, incêndio, acidente, interferência humana;

O nome dos campos foi reduzido com o propósito de diminuir o espaço utilizado para formalização das regras, como mostra a Tabela 3.4 e Tabela 3.5.

Tabela 3.4 – Legenda para campos de entrada.

DEV_TYPE_NAME	CLIMA	CAUSA	X_CORD	Y_CORD
DEV	CLe	CSe	-	-

Tabela 3.5 – Legenda para campos de saída.

ELEMENTO	FATOS	CLIMA	VENTO	OBJETOS	VEGETAÇÃO	CAUSAS
EL	FT	CLs	VT	OBJ	VEG	CSs

A seguir será descrito o conjunto de regras utilizado para classificação dos dados.

Campo ELEMENTO

Utilizando o conceito de Elemento de Interrupção, baseado na classificação apresentada em [6], e o campo 'CAUSA' presente no histórico de eventos original, é possível concluir que elemento físico da rede falhou para causar o desligamento. Assim, baseado em informações sobre a causa do evento e nas classes de elemento afetadas durante uma falha no fornecimento de energia [64], um novo campo chamado elementos foi criado. Portanto, para cada evento de desligamento analisado no banco de dados original, uma nova classe de elementos foi atribuída de acordo com os dados disponíveis nesse banco de dados. Todos esses elementos foram classificados com o auxílio de um especialista, como descrito a seguir:

1. Poste: Poste Podre, Poste, Abalroamento de Poste, Erosão.
2. Equipamentos: Religador, Descarga Atmosférica, Animais/Pássaros, TP - Transformador de Potência, Ferragens, Pára-Raios, Vandalismo, Regulador, Transformador, Elo Fusível Queimado, Chave Fusível, Chave Faca.
3. Condutores: Conexões (todo ponto de ligação), Ramal de Ligação, Condutor Desregulado, Condutor, Vento, Emenda, Amarrilho, Animais/Pássaros.
4. Isoladores: Isoladores, Animais/Pássaros.
5. Cruzeta: Cruzeta.

Estados – Poste, Equipamentos, Condutores, Isoladores, Cruzeta.

Os elementos que não foram classificados diretamente a partir da sua causa original passaram por um processo de classificação através de regras, e assim chegou-se a uma conclusão sobre todas as informações constantes do banco de dados original. Definiram-se, então, um conjunto de 22 regras para relacionar essas causas a um elemento de interrupção:

$$r_1: (\text{CSe} = \text{vegetal}) \wedge [(\text{DEV} = \text{chave fusível}) \vee \ldots (\text{DEV} = \text{transformador})] \rightarrow \text{equipamentos}$$

$$r_2: (\text{CSe} = \text{vegetal}) \wedge [(\text{DEV} \neq \text{chave fusível}) \wedge \ldots (\text{DEV} \neq \text{transformador})] \rightarrow \text{condutores}$$

$$r_3: (\text{CSe} = \text{sobrecarga}) \wedge [(\text{DEV} = \text{chave fusível}) \vee \ldots (\text{DEV} = \text{transformador})] \rightarrow \text{equipamentos}$$

$$r_4: (\text{CSe} = \text{sobrecarga}) \wedge [(\text{DEV} \neq \text{chave fusível}) \wedge \ldots (\text{DEV} \neq \text{transformador})] \rightarrow \text{condutores}$$

$$r_5: (\text{CSe} = \text{queimada/incêndio}) \wedge [(\text{DEV} = \text{chave fusível}) \vee \ldots (\text{DEV} = \text{transformador})] \rightarrow \text{equipamentos}$$

$$r_6: (\text{CSe} = \text{queimada/incêndio}) \wedge [(\text{DEV} \neq \text{chave fusível}) \wedge \ldots (\text{DEV} \neq \text{transformador})] \rightarrow \text{poste}$$

$$r_7: (\text{CSe} = \text{danos causados por terceiros}) \wedge [(\text{DEV} = \text{chave fusível}) \vee \ldots (\text{DEV} = \text{transformador})] \rightarrow \text{equipamentos}$$

$$r_8: (\text{CSe} = \text{danos causados por terceiros}) \wedge [(\text{DEV} \neq \text{chave fusível}) \wedge \ldots (\text{DEV} \neq \text{transformador})] \rightarrow \text{condutores}$$

$$r_9: (\text{CSe} = \text{poluição}) \wedge [(\text{DEV} = \text{chave fusível}) \vee \ldots (\text{DEV} = \text{transformador})] \rightarrow \text{equipamentos}$$

$$r_{10}: (\text{CSe} = \text{poluição}) \wedge [(\text{DEV} \neq \text{chave fusível}) \wedge \ldots (\text{DEV} \neq \text{transformador})] \rightarrow \text{condutores}$$

r_{11} : (CSe = falta de manutenção) $\wedge$ [(DEV = chave fusível) $\vee$ ...
(DEV = transformador)] $\rightarrow$ equipamentos

r_{12} : (CSe = falta de manutenção) $\wedge$ [(DEV $\neq$ chave fusível) $\wedge$ ...
(DEV $\neq$ transformador)] $\rightarrow$ isoladores

r_{13} : (CSe = bola na rede elétrica) $\wedge$ [(DEV = chave fusível) $\vee$ ...
(DEV = transformador)] $\rightarrow$ equipamentos

r_{14} : (CSe = bola na rede elétrica) $\wedge$ [(DEV $\neq$ chave fusível) $\wedge$ ...
(DEV $\neq$ transformador)] $\rightarrow$ condutores

r_{15} : (CSe = pandorga) $\wedge$ [(DEV = chave fusível) $\vee$ ...
(DEV = transformador)] $\rightarrow$ equipamentos

r_{16} : (CSe = pandorga) $\wedge$ [(DEV $\neq$ chave fusível) $\wedge$ ...
(DEV $\neq$ transformador)] $\rightarrow$ condutores

r_{17} : (CSe = redes de comunicação) $\wedge$ [(DEV = chave fusível) $\vee$ ...
(DEV = transformador)] $\rightarrow$ equipamentos

r_{18} : (CSe = redes de comunicação) $\wedge$ [(DEV $\neq$ chave fusível) $\wedge$ ...
(DEV $\neq$ transformador)] $\rightarrow$ condutores

r_{19} : (CSe = inundações) $\wedge$ [(DEV = chave fusível) $\vee$ ...
(DEV = transformador)] $\rightarrow$ equipamentos

r_{20} : (CSe = inundações) $\wedge$ [(DEV $\neq$ chave fusível) $\wedge$ ...
(DEV $\neq$ transformador)] $\rightarrow$ poste

r_{21} : (CSe = objetos estranhos na rede) $\wedge$ [(DEV = chave fusível) $\vee$ ...
(DEV = transformador)] $\rightarrow$ equipamentos

r_{22} : (CSe = objetos estranhos na rede) $\wedge$ [(DEV $\neq$ chave fusível) $\wedge$ ...
(DEV $\neq$ transformador)] $\rightarrow$ condutores

Campo FATOS

Este campo possui informações sobre atividades que são realizadas próximas à rede elétrica e que podem influenciar em um desligamento.

Com base nas causas apontadas no conjunto de dados utilizado como entrada para as regras foi possível retirar algumas conclusões a respeito de fatos associados ao desligamento. Dessa forma, algumas as seguintes regras foram adotadas:

r_1 : (CSe = queimada/incêndio) $\rightarrow$ queimada/incêndio

r_2 : (CSe = danos causados por terceiros) $\rightarrow$ vandalismo

r_3 : (CSe = abalroamento de poste) $\rightarrow$ acidente

r_4 : (CSe = inundações) $\rightarrow$ inundação

r_5 : (CSe = erosão) $\rightarrow$ erosão

Pode-se considerar como saída do motor de inferência para este campo os seguintes estados:

Estados - acidente, queimada/incêndio, empresa, vandalismo, inundação e erosão.

Para qualquer outra causa apontada atribuiu-se ao campo "FATOS" o símbolo '*', indicando que a variável não foi observada naquele evento.

Campo CLIMA

No banco de dados original, "Vento" era um dos estados que a variável "CLIMA" podia assumir. No entanto, caso estivesse chovendo e ventando ao mesmo tempo, não seria possível descrever adequadamente as condições climáticas do local do evento. Por isso, o estado "Vento" foi eliminado da variável "CLIMA", sendo utilizado agora para inferir a respeito da intensidade do vento no novo campo "VENTO".

Assim, criou-se a seguinte regra, com o objetivo de substituir o antigo estado "Vento" por um novo estado no campo "CLIMA":

$r_1 : (\text{CLe} = \text{vento}) \rightarrow \text{bom}$

Os demais estados dessa variável foram extraídos diretamente a partir do banco original, pois já estavam em um formato adequado para o estudo realizado.

Estados - bom, chuva, temporal, neve/granizo, neblina.

Campo VENTO

A definição do estado desse campo baseia-se no que pode ser observado no campo "CLIMA", já que as condições climáticas dão uma idéia sobre a intensidade do vento. Por exemplo, num dia de tempestade, a chance de se ter ventos fortes é alta, então se pode dizer que num dia de tempestade, o vento é excessivo. A partir disso derivaram-se as seguintes regras para definir o estado da variável.

$r_1 : (\text{CLe} = \text{bom}) \rightarrow \text{sem vento}$
$r_2 : (\text{CLe} = \text{chuva}) \rightarrow \text{moderado}$
$r_3 : (\text{CLe} = \text{tempestade}) \vee (\text{CLe} = \text{vento}) \vee (\text{CLe} = \text{granizo}) \rightarrow \text{excessivo}$

Estados - sem vento, moderado, excessivo.

Campo OBJETOS

É comum a ocorrência de desligamentos ocasionados por objetos presos na rede elétrica, portanto, essa variável deve ser observada no local do desligamento. Nesse caso é indicada somente à presença ou não de objetos presos na rede.

O estado desse campo foi definido a partir da causa apontada nos dados de entrada. A regra é a seguinte:

$$r_1: (\text{CSe} = \text{vegetal}) \vee (\text{CSe} = \text{pandorga}) \vee (\text{CSe} = \text{animais/pássaros}) \vee \ldots$$
$$(\text{CSe} = \text{objetos estranhos na rede}) \vee (\text{CSe} = \text{bola na rede elétrica}) \rightarrow \text{sim}$$
$$r_2: (\text{CSe} \neq \text{vegetal}) \wedge (\text{CSe} \neq \text{pandorga}) \wedge (\text{CSe} \neq \text{animais/pássaros}) \wedge \ldots$$
$$(\text{CSe} \neq \text{objetos estranhos na rede}) \wedge (\text{CSe} \neq \text{bola na rede elétrica}) \rightarrow \text{não}$$

Aplicando as regras, a variável irá assumir um dos dois estados definidos para este campo:

Estados - sim, não.

Campo VEGETAÇÃO

De acordo com Brown [64], contato com árvores é uma das três causas mais comuns de interrupção de energia, portanto a observação das condições da vegetação no local do evento é muito importante.

À primeira vista, nenhuma conclusão a respeito das condições da vegetação pode ser tirada dos dados brutos. No entanto, se as informações disponíveis sobre a causa e a coordenada geográfica do local do desligamento forem cruzadas, pode-se inferir algo sobre o estado da vegetação para o evento, como, por exemplo, estado podado ou não-podado.

Os eventos de desligamento não-programados foram agrupados por município utilizando o algoritmo *k-means* [65]. A idéia deste algoritmo é classificar em agrupamentos (clusters) as informações com características similares, baseada na análise comparativa entre as informações e valores numéricos dos dados, de forma automática e sem a necessidade de uma pré-classificação. Por causa desta característica, o k-means é considerado como um algoritmo de mineração de dados não supervisionado.

Estados - podada, não podada.

O funcionamento do algoritmo *k-means* é bastante simples. O número de classes ou agrupamentos deve ser informado ao algoritmo. Os passos do algoritmo são descritos brevemente a seguir, apoiados pela Figura 3.8.

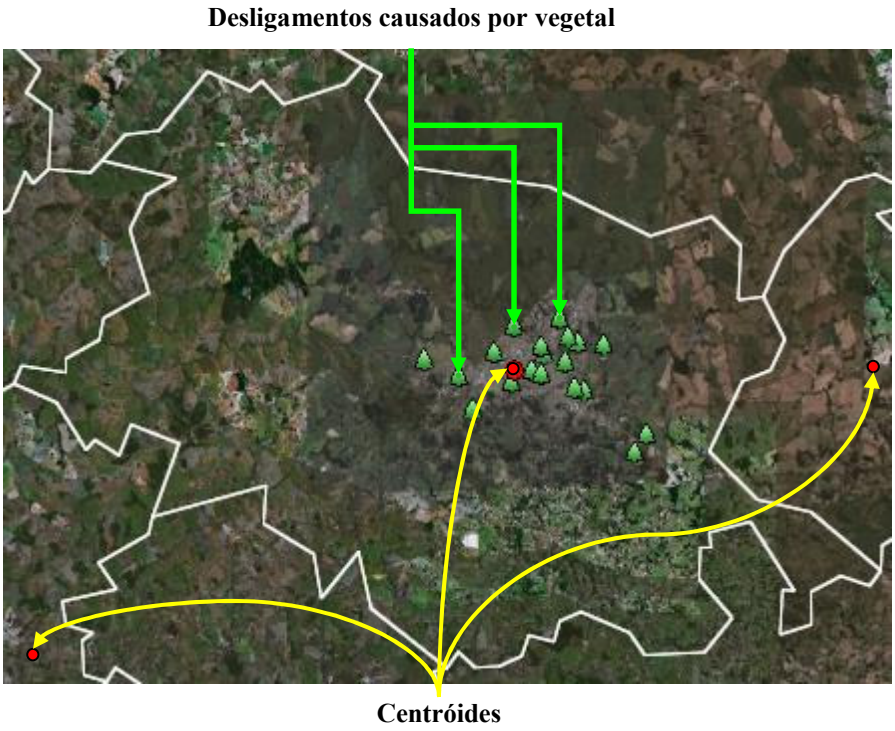


Figura 3.8 - Ilustração para o campo vegetal

Passo 1: Os centróides passados ao algoritmo são inicializados. Neste trabalho, os centróides são os pontos correspondentes às coordenadas geográficas dos municípios da área de atuação da empresa de distribuição, conforme apresentado na Figura 3.8, porém é possível utilizar qualquer seqüência aleatória de pontos;

Passo 2: Calcular a distância euclidiana entre cada ponto (coordenada geográfica onde ocorreu o desligamento) e os centróides (coordenada geográfica dos municípios);

Passo 3: Colocar cada ponto nas classes de acordo com a sua distância do centróide da classe, ou seja, os pontos que estiverem mais próximos ao centróide de uma determinada classe, serão incluídos nela própria;

Passo 4: Calcular os novos centróides para cada classe. O novo centróide da classe é calculado através da média de todos os pontos pertencentes à classe;

Passo 5: O algoritmo volta para o Passo 2 repetindo iterativamente o refinamento do cálculo das coordenadas dos centróides. Repetir até a convergência, que

ocorrerá quando não houver mais a necessidade de trocar de classe os pontos a serem agrupados.

Os parâmetros de entrada da função do algoritmo são os seguintes:

- m – matriz de dados que contém os pontos correspondentes às coordenadas geográficas X e Y (*UTM*) dos eventos de desligamento;
- k – número de grupos (clusters), nesse caso é 272;
- israndi – matriz de dados que contém os pontos correspondentes às coordenadas geográficas X e Y (*UTM*) dos 272 municípios, que representam os centróides.

A função retorna uma matriz de dados idêntica à matriz m, porém com uma coluna adicional que representa os grupos (municípios) onde cada evento de desligamento ocorreu.

Dessa forma, foi possível saber quantos desligamentos causados por vegetação ocorreram cada município, permitindo a extração de dados estatísticos que serão utilizados como critério na definição do estado da vegetação no local do evento, como demonstra a Tabela 3.6.

Tabela 3.6 – Dados estatísticos extraídos

Desligamento Causa: Vegetal	N.o de Municípios	Média μ	Desvio Padrão σ	Variância σ^2
1759	272	6,47	4,98	24,76

O processo de mineração de dados passou então a ter mais uma fonte de informação, o campo *número de desligamentos por município*. Esse número de desligamentos por município foi utilizado na criação da regra que define se aquela região onde ocorreu o evento possui uma vegetação abundante ou não. A Tabela 3.7mostra como ficou a base dados.

Tabela 3.7 – Nova base de dados

Evento	DEV_TYPE_NAME	CLIMA	CAUSA	ÁREA	Nº VEGETAIS
1	Chave Fusível	Chuva	Descarga Atmosférica	1	9
2	Transformador	Chuva	Descarga Atmosférica	1	9
3	Transformador	Chuva	Descarga Atmosférica	1	9
⋮	⋮	⋮	⋮	⋮	⋮
2420	Transformador	Vento	Ramal de Ligação	67	4
2421	Transformador	Bom	Conexões	67	4
2422	Chave Fusível	Bom	Vegetal	67	4
⋮	⋮	⋮	⋮	⋮	⋮
6204	Chave Fusível	Temporal	Poste Podre	173	21
6205	Chave	Temporal	Poste Podre	173	21
6206	Transformador	Temporal	Poste Podre	173	21
6207	Transformador	Temporal	Transformador	173	21
6208	Chave Fusível	Vento	Descarga Atmosférica	173	21
6209	Transformador	Vento	Descarga Atmosférica	173	21
⋮	⋮	⋮	⋮	⋮	⋮
10145	Chave Fusível	Bom	Animais/Pássaros	272	0
10146	Transformador	Bom	Vento	272	0
10147	Chave Fusível	Bom	Chave Fusível	272	0
10148	Transformador	Bom	Animais/Pássaros	272	0

A regra leva em consideração o número médio de desligamentos causados por vegetal em cada município da concessão da RGE. A média μ=6,47 foi truncada para μ=6. Assim, sabendo que um evento ocorreu num determinado município e que nesse município ocorreu um número *x* de desligamentos causados por vegetal, é possível concluir algo a respeito do estado da vegetação. Neste caso a regra elaborada foi descrita assim:

r_1 : $(x > \mu) \rightarrow$ não podado

r_2 : $(x < \mu) \rightarrow$ podado

Estados: não podado, podado.

Campo CAUSAS

Este é o campo que deve ser preenchido com a causa do desligamento. As causas presentes no banco de dados original foram classificadas dentro de uma das nove classes de causas de desligamento utilizadas nesse trabalho:

1. Falha no componente;
2. Sobrecarga;
3. Clima adverso;
4. Descarga atmosférica;
5. Interferência do meio;
6. Vegetal;
7. Interferência humana;
8. Acidente;
9. Queimada/incêndio.

Com o objetivo de preencher o campo "CAUSAS" da nova base de dados, utilizou-se um novo conjunto de regras, as quais são exemplificadas a seguir:

r_1 : (CSe = vegetal) $\rightarrow$ vegetal

r_2 : (CSe = descarga atmosférica) $\rightarrow$ descarga atmosférica

r_3 : (CSe = condutor desregulado) $\vee$ (CSe = condutores) $\vee$ (CSe = poste podre) $\vee$ ...
(CSe = poste) $\vee$ (CSe = ferragens) $\vee$ (CSe = regulador) $\vee$ ...
(CSe = transformador) $\vee$ (CSe = TP - transformador de potência) $\vee$ ...
(CSe = TC - transformador de corrente) $\vee$ ...
(CSe = TP - transformador de potência) $\rightarrow$ falha do componente

r_4 : (CSe = vento) $\rightarrow$ clima adverso

r_5 : (CSe = sobrecarga) $\rightarrow$ sobrecarga

r_6 : (CSe = animais/pássaros) $\vee$ (CSe = poluição) $\rightarrow$ interferência do meio

r_7 : (CSe = queimada/incêndio) $\rightarrow$ queimada/incêndio

r_8 : (CSe = abalroamento de poste) $\rightarrow$ acidente

r_9 : (CSe = danos causados por terceiros) $\vee$ (CSe = vandalismo) $\vee$ ...
(CSe = pandorga) $\vee$ (CSe = objetos estranhos na rede) $\vee$ ...
(CSe = redes de comunicação) $\vee$ (CSe = bola na rede elétrica) $\vee$ ...
(CSe = falta de manutenção) $\rightarrow$ interferência humana

r_{10} : (CSe = conexões) $\vee$ (CSe = isolador) $\vee$ (CSe = ramal de ligação) $\vee$ ...
(CSe = pára - raios) $\vee$ (CSe = religador) $\vee$ (CSe = cruzeta) $\vee$ ...
(CSe = emenda) $\vee$ (CSe = elo fusível queimado) $\vee$ (CSe = amarrilho) $\vee$ ...
(CSe = chave fusível) $\vee$ (CSe = chave faca) $\wedge$ [(CLe = temporal) $\vee$ ...
(CLe = vento) $\vee$ (CLe = granizo)] $\rightarrow$ clima adverso

r_{11} : (CSe = conexões) $\vee$ (CSe = isolador) $\vee$ (CSe = ramal de ligação) $\vee$ ...
(CSe = pára - raios) $\vee$ (CSe = religador) $\vee$ (CSe = cruzeta) $\vee$ ...
(CSe = emenda) $\vee$ (CSe = elo fusível queimado) $\vee$ (CSe = amarrilho) $\vee$ ...
(CSe = chave fusível) $\vee$ (CSe = chave faca) $\wedge$ [(CLe $\neq$ temporal) $\wedge$ ...
(CLe $\neq$ vento) $\wedge$ (CLe $\neq$ granizo)] $\rightarrow$ falha no componente

r_{12} : (CSe = inundações) $\vee$ (CSe = erosão) $\wedge$ ...
(CLe $\neq$ bom) $\rightarrow$ clima adverso

r_{13} : (CSe = inundações) $\vee$ (CSe = erosão) $\wedge$ ...
(CLe = bom) $\rightarrow$ interferência do meio

3.3.5. Avaliação e Representação do Conhecimento

Em [67], o autor descreve Redes Bayesianas como um tipo de representação de conhecimento que permite o aprendizado não só a partir de dados estatísticos, como também do conhecimento de especialistas. Assim como as Redes Bayesianas, as RNA permitem a representação do conhecimento adquirido através do aprendizado a partir de um conjunto de amostras. Por esse motivo, ambas são utilizadas juntamente com o conjunto de dados extraídos através do processo de *KDD*, para validação da metodologia e visualização dos diagnósticos. Essa validação será abordada em detalhe na seção 3.6.

3.3.6. Resultados do Processo de *KDD*

Um dos objetivos desse trabalho foi aplicar a técnica de *KDD* para a extração de conhecimento de uma base de dados bruta relacionada à interrupção de energia em redes de distribuição. Uma fase importante foi relacionada à etapa de mineração de

dados que consistiu em extração de informação interessante, relacionada aos aspectos que envolvem uma interrupção, não trivial vinculada à associação de condições climáticas da hora das interrupções, implícitas vinculadas aos apontamentos incompletos dos coletores (eletricistas), previamente desconhecida vinculadas aos aspectos operacionais dos sistemas de distribuição e potencialmente útil como a identificação das causas que podem ser diretamente aplicadas ao planejamento, operação e manutenção dos sistemas de distribuição. Neste caso, um exemplo do resultado do processo de *KDD* pode ser visto na Tabela 3.8, onde uma nova base de dados foi composta para ser utilizada como entrada para um aplicativo que identifica causas de desligamentos mediante algumas observações de campo.

Tabela 3.8 – Resultado do processo de *KDD*, conhecimento extraído para o treinamento de sistemas especialistas

ELEMENTO	FATOS	CLIMA	VENTO	OBJETO	VEGETAÇÃO	CAUSAS
Equipamentos	*	Chuva	Moderado	Não	SemPoda	DescAtm
Poste	Acidente	Bom	SemVento	Não	SemPoda	Acidente
Condutores	*	Bom	SemVento	Sim	SemPoda	Vegetal
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Equipamentos	*	Temporal	Excessivo	Não	Podada	DescAtm
Isoladores	*	Chuva	Moderado	Não	SemPoda	FalhaComp
Equipamentos	*	Chuva	Moderado	Não	SemPoda	DescAtm
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Condutores	*	Bom	SemVento	Sim	SemPoda	Vegetal
Poste	Acidente	Bom	SemVento	Não	Podada	Acidente
Equipamentos	Empresa	Bom	SemVento	Não	Podada	InterfHum
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Condutores	*	Chuva	Moderado	Não	SemPoda	FalhaComp
Equipamentos	Vandalismo	Bom	SemVento	Não	SemPoda	InterfHum
Isoladores	*	Chuva	Moderado	Não	SemPoda	FalhaComp

3.4. Implementação da Rede Bayesiana

As Redes Bayesianas mostradas nos exemplos foram implementadas utilizando o software *Netica* [42], por esse apresentar uma boa interface gráfica permitindo que a visualização, o aprendizado e as inferências fossem realizadas de forma bastante simples. No entanto, a versão Demo disponibilizada apresenta uma série de limitações. Uma delas, e que foi o motivo determinante para se buscar uma ferramenta *freeware*, foi o fato de que a versão Demo do *Netica* limita o aprendizado de parâmetros através do algoritmo *EM* a somente 1000 amostras. A alternativa encontrada foi utilizar o pacote

Bayes Net Toolbox for Matlab (*BNT*). O *BNT* [69] é um pacote de código aberto para *Matlab* para modelagem de grafos orientados. *BNT* suporta vários tipos de nós (distribuições de probabilidade), inferência exata e aproximada, aprendizado de estrutura e de parâmetros, modelos estáticos e dinâmicos. O *BNT* é bastante utilizado para ensino e em pesquisa aplicada.

O desenvolvimento de uma Rede Bayesiana está dividido basicamente em duas partes: (i) definição da estrutura e (ii) definição dos parâmetros numéricos. Ambos podem ser definidos por indução a partir de uma amostra de dados ou através do conhecimento de um especialista. A seguir será mostrado como a Rede Bayesiana proposta foi implementada utilizando o *BNT*.

3.4.1. Estrutura da Rede Bayesiana

É a parte qualitativa da rede, representa as relações de dependência entre as variáveis através de um grafo acíclico orientado (GAO). Os nós representam as variáveis e os arcos, as relações de dependência entre elas.

A referência [67] descreve uma abordagem simples para a construção de estruturas de Redes Bayesianas:

"A abordagem é baseada em duas observações: (1) as pessoas podem na maioria das vezes determinar as relações causais entre as variáveis, e (2) as relações causais correspondem tipicamente às relações de dependência condicional".

O primeiro passo na construção da estrutura da rede é a identificação correta das variáveis envolvidas no domínio em estudo, ver seção 3.2. Então, as variáveis que representam o cenário são as seguintes:

1. Causa do Desligamento;
2. Clima;
3. Fatos Associados;
4. Vegetação;
5. Objetos Estranhos;
6. Elemento de Interrupção;

7. Vento.

A estrutura utilizada na rede é bastante semelhante à de um classificador naïve bayes [40], esse tipo de topologia mostra-se muito eficaz em aplicações para diagnóstico em geral e por esse motivo será utilizado no trabalho. A topologia adotada se diferencia da naïve bayes convencional apenas pela relação de dependência adicionada do nó "CLIMA" para o nó "VENTO", como mostra a Figura 3.9.

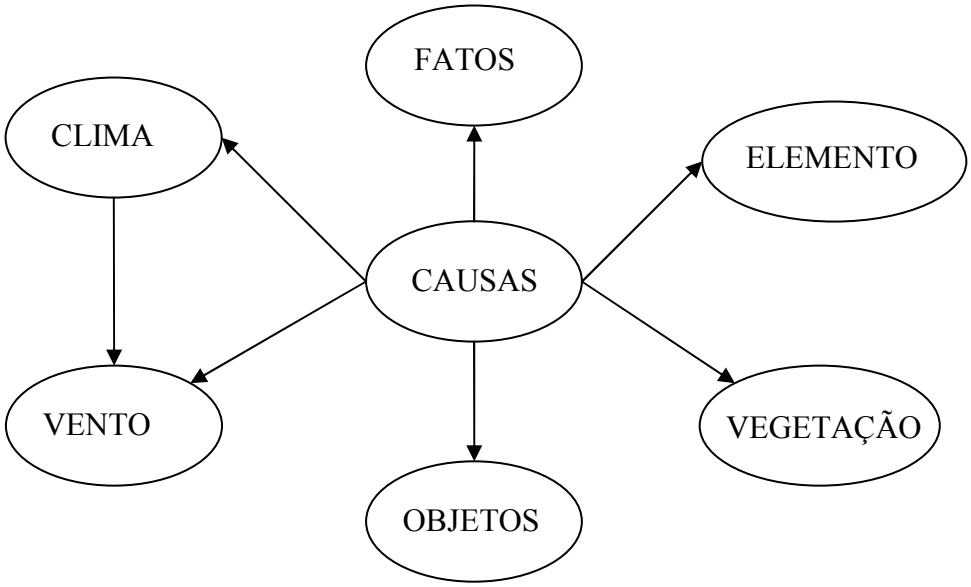


Figura 3.9 – Estrutura da Rede Bayesiana

Implementação da Estrutura da Rede Bayesiana no *BNT*

Para especificar o GAO utilizando o *BNT*, deve-se criar uma matriz *n* x *n*, onde *n* é o número de nós da rede, logo a matriz que representará a estrutura da rede será uma matriz de sete por sete elementos. Deve-se numerar os nós em ordem topológica, isto é, pais devem ser numerados antes de seus filhos. Então,

CAUSAS = 1

CLIMA = 2

FATOS = 3

VEGETAÇÃO = 4

OBJETOS = 5

ELEMENTOS = 6

VENTO = 7

Tendo definido a ordem dos nós, agora é preciso estabelecer a relação de dependência que eles têm entre si, para isso atribui-se à posição correspondente da matriz o valor 1 para indicar que existe um arco entre aqueles dois nós (correspondentes à linha e à coluna da posição da matriz), a direção do arco é definida de acordo com a posição da matriz em que foi atribuído o valor. Por exemplo, deseja-se estabelecer uma relação entre o nó 1 (CAUSA) e o nó 2 (CLIMA), então o valor 1 deve ser atribuído à posição (1,2) da matriz, indicando que o nó 1 é pai do nó dois, se não existe relação entre os nós, então é atribuído o valor zero àquela posição. A matriz que representa a Rede Bayesiana da Figura 3.9 é a seguinte:

$$\mathbf{DAGNETA} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Após a definição do número de estados de cada nó e do tipo de nó (que neste caso é discreto) foi utilizada a função *mk_bnet* para criar a Rede Bayesiana RBNETA. A seguir, foi construída uma *CPT* para cada nó da rede com a função *tabular_CPD*. Através dessa função pode-se inserir diretamente nas *CPT* as probabilidades da rede, mas aqui essas *CPT* ficarão em branco, pois seus parâmetros serão obtidos com base no histórico de eventos de desligamento. Com essas informações, está definida a estrutura da RB. Sabe-se o número de estados de cada nó, a relação de dependência entre eles e as *CPT* prontas para receber os parâmetros numéricos da rede.

3.4.2. Treinamento da Rede Bayesiana

É a parte quantitativa da rede, especifica as probabilidades entre as variáveis, ou seja, insere-se nas *CPT* as probabilidades condicionais entre os nós que têm dependência direta.

O aprendizado da rede será feito utilizando o algoritmo *EM*, pois o conjunto de dados obtido através do processo de mineração de dados apresenta muitos campos incompletos, já que algumas vezes o estado de uma variável não pode ser observado nem inferido. Como exemplo, pode-se citar o nó "Fatos associados", que apresenta dados faltosos em quase todos os casos. Isso pode ser explicado pelo fato de que acidente, incêndio, empresas trabalhando no local, vandalismo, enchente e erosão nem sempre são ocorrências normalmente observadas.

Implementação dos Parâmetros Numéricos da Rede Bayesiana no *BNT*

Sendo conhecida a estrutura da Rede Bayesiana, deve-se realizar o aprendizado dos parâmetros com base numa amostra de dados conhecida. Primeiro é necessário definir o algoritmo de inferência que será utilizado. Esse irá servir não somente para realizar inferências na rede depois de treinada, mas também durante o aprendizado da rede, pelo algoritmo *EM*, mais especificamente no passo E. O algoritmo escolhido foi o de eliminação de variáveis, descrito na subseção 2.2.3, a função *var_elim_inf_engine* cria uma máquina de inferência para a rede. Essa máquina de inferência será utilizada sempre que se quiser estimar o estado de um nó, dado o que é observado nos demais nós da rede. Assim, de posse do conjunto de dados de treinamento foi utilizada a função *learn_params_em* (que tem o algoritmo *EM* implementado) para o aprendizado da rede. Após a execução de todos passos necessários para a construção da Rede Bayesiana, o sistema terá a topologia e os parâmetros mostrados na Figura 3.10. Apesar da estrutura da rede ser sempre a mesma, deve-se lembrar que isso não ocorre com os seus parâmetros, que serão distintos para cada conjunto de treinamento utilizado na etapa de validação dos resultados.

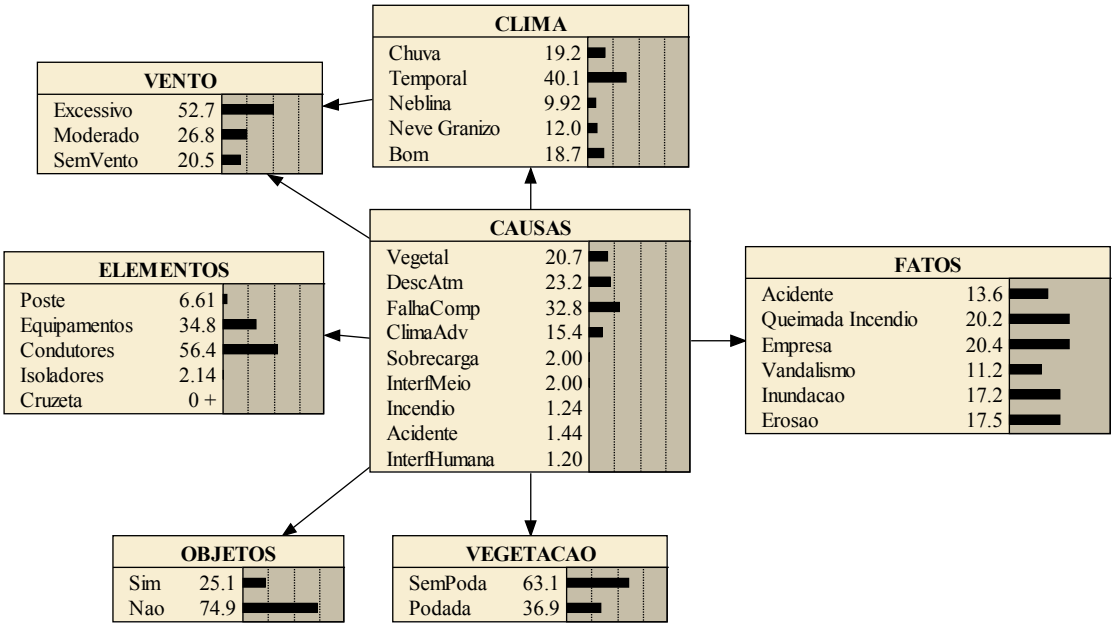


Figura 3.10 – Topologia e parâmetros da RB. Os valores mostrados estão em percentual.

3.5. Implementação da Rede Neural

RNA é um método muito utilizado na solução de problemas na área de sistemas de potência [4][68] e considerando a semelhança existente com Redes Bayesianas, será feita uma comparação entre as duas heurísticas. Tanto Redes Bayesianas quanto Redes Neurais, têm sua representação baseada em atributos, são representadas através de tuplas de atributos. Além disso, podem lidar com entradas discretas e contínuas [40].

Foi implementada uma rede do tipo *multilayer feedforward* contendo duas camadas escondidas com quinze neurônios cada. A especificação dos pesos sinápticos que interconectam os neurônios da rede foi feita utilizando os algoritmos *back-propagation* e *resilient back-propagation*. A determinação do número de camadas escondidas, bem como o número de neurônios em cada uma das camadas, foi realizada comparando quatro topologias distintas utilizando os dois algoritmos para o treinamento. Para escolha da topologia utilizou-se como critério de avaliação o desempenho da rede, isto é, a que obteve o menor erro no menor espaço de tempo após o treinamento. Os seguintes tipos de estrutura de RNA foram comparados:

- uma camada escondida com quinze neurônios – chamadas Bp15 e Rp15;
- duas camadas escondidas com quinze neurônios em cada camada – chamadas Bp30_15 e Rp30_15;
- três camadas escondidas com trinta neurônios na primeira camada e quinze neurônios na segunda e terceira camada - chamadas Bp30_15_15 e Rp30_15_15;
- quatro camadas escondidas com trinta neurônios na primeira e segunda camada e quinze neurônios na terceira e quarta camada - chamadas Bp30_30_15_15 e Rp30_30_15_15.

As Figura 3.11 e Figura 3.12 mostram o erro obtido com cada tipo de estrutura utilizando como algoritmos de treinamento o *back-propagation* e o *resilient back-propagation* respectivamente. A Tabela 3.9 faz uma comparação de desempenho entre as diferentes topologias treinadas para ambos os algoritmos.

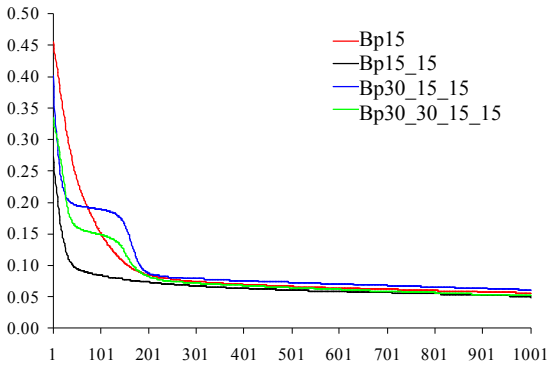


Figura 3.11 – Desempenho da RNA utilizando o algoritmo *Back propagation* com taxa de aprendizado = 0.5 e momento = 0.7.

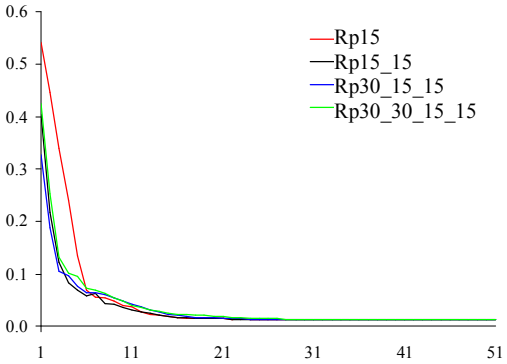


Figura 3.12 – Desempenho da RNA utilizando o algoritmo *Resilient Back propagation*, $+\Delta_{ij}^{(t)} = 1.2.$ e $-\Delta_{ij}^{(t)} = 0.5.$

Tabela 3.9 – Comparação entre as diferentes topologias treinadas.

	Topologia	Tempo (s)	MSE
1	Bp15	31,11	0,055335
2	Bp15_15	60,98	0,049119
3	Bp30_15_15	75,39	0,060524
4	Bp30_30_15_15	102,73	0,051573
5	Rp15	2,92	0,011991
6	Rp15_15	2,95	0,011921
7	Rp30_15_15	4,08	0,011982
8	Rp30_30_15_15	5,78	0,011988

Assim, uma RNA de duas camadas com quinze neurônios cada foi treinada e validada com base no mesmo conjunto de treinamento e validação utilizados na Rede Bayesiana. O algoritmo de treinamento utilizado foi o *Resilient Back propagation*, e a função de ativação escolhida foi a sigmoidal, pois este método de aprendizado necessita do cálculo do gradiente da função, então esta deve ser contínua e diferenciável.

É importante salientar que existe uma diferença fundamental entre os dois métodos, que faz com que as Redes Bayesianas sejam mais indicadas em determinadas aplicações. As Redes Neurais podem consistir de várias camadas de nós e normalmente todos nós de uma camada estão conectados a todos nós da camada subseqüente. Um nó presente numa camada escondida da Rede Neural não tem significado algum para o domínio do sistema, o que não ocorre com os nós das Redes Bayesianas que, juntamente com suas *CPT*, podem representar conceitos bem definidos a respeito do domínio.

O software *Matlab* foi utilizado para a modelagem da Rede Neural. A ferramenta *nntool*, que faz parte da *Toolbox* de Redes Neurais do *Matlab* permite a criação da estrutura, treinamento e validação da rede.

3.5.1. Estrutura da Rede Neural

A modelagem da estrutura da Rede Neural deve ser feita a partir da estrutura definida para Rede Bayesiana. Portanto, as variáveis utilizadas serão as mesmas. As entradas da rede serão as variáveis observadas no local do evento e a saída da rede será a causa identificada.

As entradas da Rede Neural estão divididas em seis categorias:

1. Elementos de Interrupção: condutores, cruzeta, equipamentos, isoladores e poste.
2. Fatos Associados: acidente, empresa, erosão, inundação, queimada/incêndio e vandalismo.
3. Clima: bom, chuva, neblina, neve/granizo e temporal.
4. Vento: excessivo, moderado e sem vento.
5. Objetos: sim e não.
6. Vegetação: podada e sem poda.

Na saída da rede estão as nove possíveis causas para o desligamento: acidente, clima adverso, descarga atmosférica, falha no componente, incêndio, interferência humana, interferência do meio, sobrecarga e vegetal. A Figura 3.13 mostra a estrutura da RNA.

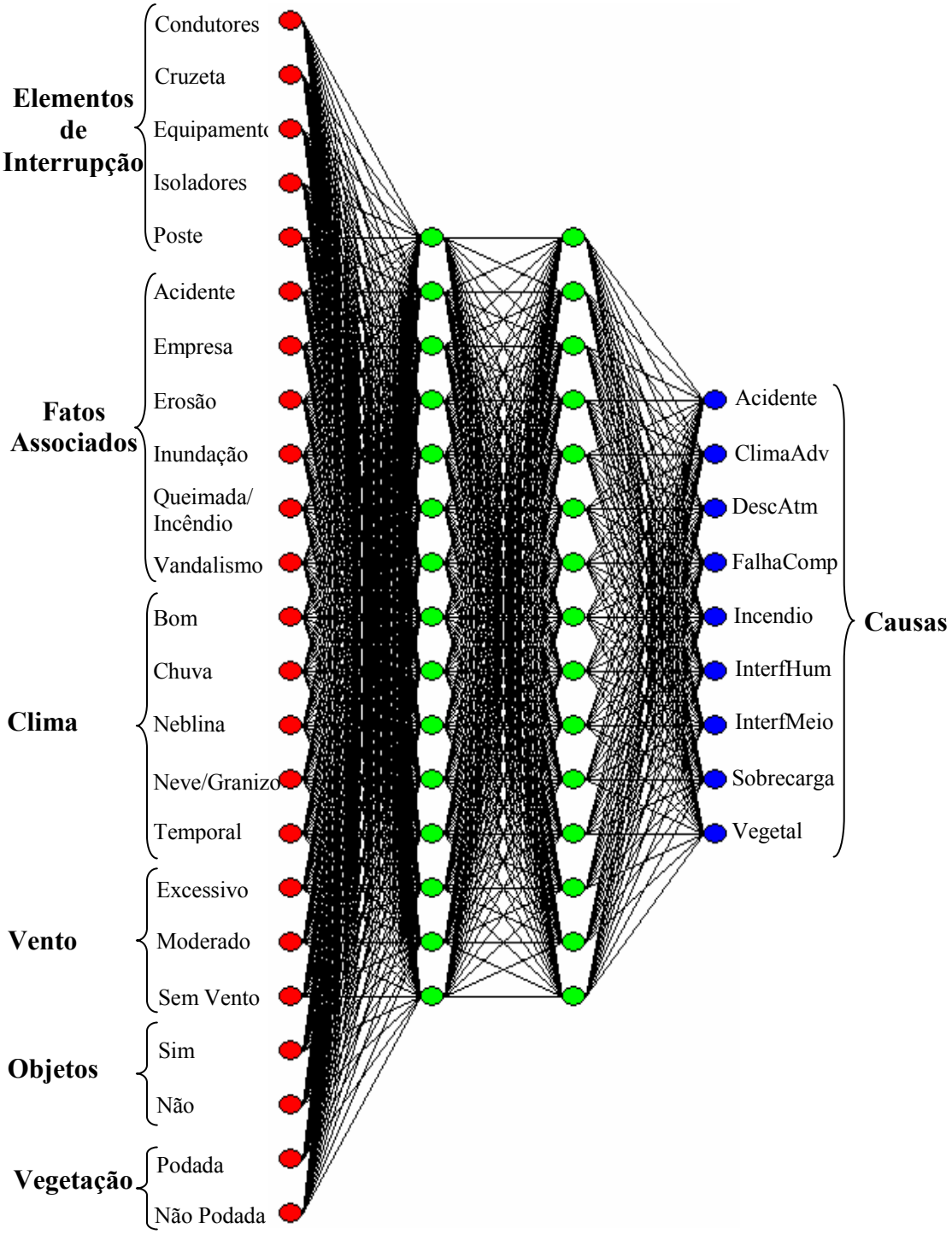


Figura 3.13 – Estrutura da Rede Neural

3.5.2. Treinamento da Rede Neural

O treinamento da Rede Neural foi realizado com base na mesma amostra de dados utilizada para o aprendizado da Rede Bayesiana. Cada entrada da Rede Neural pode assumir valor verdadeiro (1) ou falso (0). Da mesma forma as saídas da rede irão assumir um desses valores, dependendo da causa do desligamento. Como exemplo, a Figura 3.14 mostra como um dos registros utilizado para o treinamento da RB foi adaptado para ser aplicado na entrada e saída da rede neural. Esse procedimento foi repetido para todos os pares entrada / saída que compõem o conjunto de eventos de desligamento.

Elementos	Equipamentos
Fatos	*
Clima	Chuva
Vento	Moderado
Objetos	Não
Vegetação	SemPoda
Causas	DescAtm

Elementos	Condutores	0
	Cruzeta	0
	Equipamentos	1
	Isoladores	0
	Poste	0
Fatos	Acidente	0
	Empresa	0
	Erosao	0
	Inundaçao	0
	Queimada_Incendio	0
	Vandalismo	0
Clima	Bom	0
	Chuva	1
	Neblina	0
	Neve_Granizo	0
	Temporal	0
Vento	Excessivo	0
	Moderado	1
	SemVento	0
Objetos	Sim	0
	Nao	1
Vegetação	Podada	0
	SemPoda	1
Causas	Acidente	0
	ClimaAdv	0
	DescAtm	1
	FalhaComp	0
	Incendio	0
	InterfHumana	0
	InterfMeio	0
	Sobrecarga	0
	Vegetal	0

Figura 3.14 – Adaptação dos dados de entrada da RB para Rede Neural

Assim, um arquivo contendo o conjunto de treinamento da Rede Neural foi criado. O aprendizado foi realizado utilizando o algoritmo *Resilient back propagation*.

3.6. Resultados

3.6.1. Validação da Rede Bayesiana

O método da prova bipartida (*split-half method*) foi aplicado na validação da rede. Ele consiste em dividir o conjunto de dados em duas partes iguais, uma para treinamento e uma para validação. O conjunto obtido após o tratamento dos dados será chamado de *conjunto original* e totaliza 8888 amostras. A partir do *conjunto original*, foram selecionados aleatoriamente dez conjuntos contendo 4444 amostras cada, formando cinco pares de conjuntos de treinamento e validação. Seguindo o mesmo procedimento outros cinco pares de conjuntos de treinamento e validação foram criados contendo, cada um, 1000 amostras de dados selecionados aleatoriamente. O objetivo disso é mostrar que aumentando o número de amostras para o treinamento da Rede Bayesiana, pode-se reduzir o erro, além de possibilitar verificar se o erro apresentado para um par de conjuntos de treinamento e validação não está sendo tendencioso para aquele domínio em estudo.

Para cada evento existem variáveis ou nós que descrevem as condições no local de desligamento, e um nó que indica as possíveis causas de um desligamento forçado. O seguinte procedimento foi realizado, após o treinamento da rede. Os estados observados em cada evento do conjunto de validação são instanciados em seus respectivos nós, com exceção do nó 'CAUSAS', que irá indicar a causa provável do desligamento. Para cada evento, a rede apontou uma causa. Essa causa estimada é comparada com a causa real. O processo se repetiu para todos os eventos da amostra. Assim foi possível calcular o número de vezes que a rede teve sucesso no diagnóstico da causa e o número de vezes em que falhou. Foi possível determinar o erro da Rede Bayesiana dividindo-se o número de vezes em que o diagnóstico falhou pelo número total de eventos, como mostra a equação 3.1:

$$erro = \frac{n.o\ de\ diagnósticos\ errados}{n.o\ total\ de\ eventos} \qquad 3.1$$

Tabela 3.10 – Erros de diagnóstico da Rede Bayesiana para os conjuntos de 1000 amostras.

	Conjunto 1	Conjunto 2	Conjunto 3	μ	σ
RB 1	9.60 %	8.50 %	9.00 %	9.00 %	0.56
RB 2	9.20 %	8.20 %	8.70 %	8.70 %	0.50
RB 3	8.80 %	8.20 %	8.50 %	8.50 %	0.30
Erro médio	9.20 %	8.30 %	8.73 %	8.74 %	0.45

Tabela 3.11 – Erros de diagnóstico da Rede Bayesiana para os conjuntos de 4444 amostras.

	Conjunto 1	Conjunto 2	Conjunto 3	μ	Σ
RB 1	7.70 %	7.31 %	7.64 %	7.55 %	0.21
RB 2	7.66 %	7.41 %	7.68 %	7.58 %	0.15
RB 3	7.58 %	7.23 %	7.60 %	7.47 %	0.21
Erro médio	7.65 %	7.32 %	7.53 %	7.53 %	0.19

A Figura 3.15 mostra os estados observados em um evento de desligamento da amostra de dados e a causa apontada pela rede.

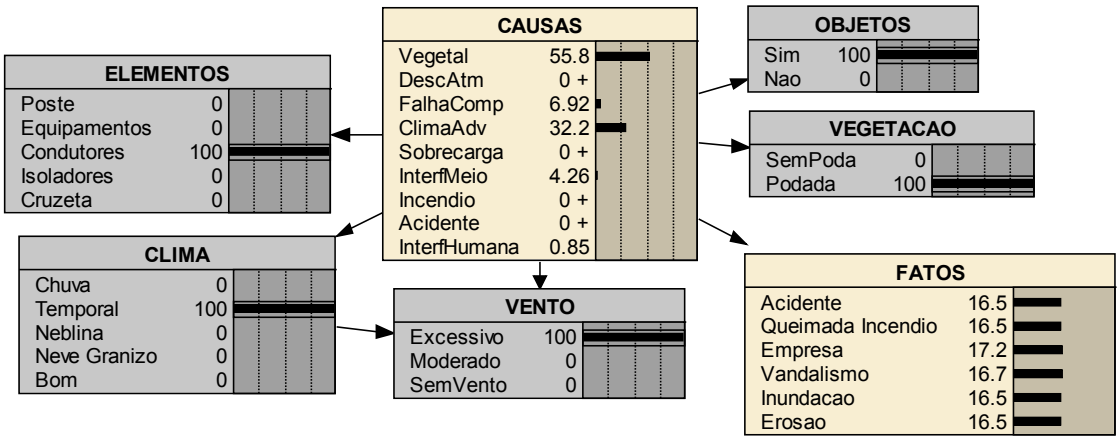


Figura 3.15 – Rede Bayesiana instanciada. Os valores mostrados estão em percentual.

3.6.2. Validação da Rede Neural

O mesmo procedimento utilizado na validação da Rede Bayesiana foi aplicado na validação do modelo criado para a Rede Neural. Portanto, os mesmos conjuntos de dados utilizados para treinamento e para a validação da Rede Bayesiana foram aplicados na Rede Neural.

$$erro = \frac{n.o\ de\ diagnósticos\ errados}{n.o\ total\ de\ eventos} \qquad 3.2$$

Tabela 3.12 – Erros de diagnóstico da Rede Neural para os conjuntos de 1000 amostras.

	Conjunto 1	Conjunto 2	Conjunto 3	μ	σ
RNA 1	7.60 %	6.70 %	7.60 %	7.30 %	0.52
RNA 2	7.50 %	6.90 %	7.50 %	7.30 %	0.34
RNA 3	7.90 %	7.60 %	7.90 %	7.80 %	0.17
Erro médio	7.67 %	7.06 %	7.67 %	7.46 %	0.29

Tabela 3.13 – Erros de diagnóstico da Rede Neural para os conjuntos de 4444 amostras.

	Conjunto 1	Conjunto 2	Conjunto 3	μ	σ
RNA 1	6.60 %	5.97 %	6.38 %	6.32 %	0.32
RNA 2	6.48 %	6.02 %	6.52 %	6.34 %	0.27
RNA 3	6.46 %	5.97 %	6.52 %	6.32 %	0.30
Erro médio	6.51 %	5.99 %	6.47 %	6.32 %	0.29

3.6.3. Análise dos Resultados

Muitas vezes, não é possível obter analiticamente uma solução para o tipo de problema proposto, especialmente por envolver diversos parâmetros e devido à alta não linearidade do modelo. Assim, uma aproximação satisfatória para a solução do problema foi encontrada através de heurísticas computacionais, como Redes Bayesianas e Redes Neurais Artificiais. O aprendizado dessas heurísticas se dá através de algoritmos, que são processos iterativos onde a cada ciclo, ou iteração, atualizam-se os valores dos parâmetros, para o caso das Redes Bayesianas, e os valores dos pesos sinápticos, para o caso das RNA, respeitando as regras de atualização dos algoritmos e buscando sempre uma melhora na performance em relação ao ciclo anterior. O processo iterativo continua até que os parâmetros (RB) e pesos sinápticos (RNA) tenham convergido para um conjunto ótimo. O critério de parada pode ser o número de iterações, ou a mínima mudança permitida nos valores entre duas iterações sucessivas.

Existem diferentes algoritmos utilizados na busca de solução de problemas, podendo-se variar a forma como são conduzidas as rotinas de atualização. Para as Redes Bayesianas, foi utilizado o algoritmo *Expectation Maximization (EM)*, que estima os parâmetros mais coerentes para serem o conjunto de dados da amostra que maximiza a função de verossimilhança. No caso das RNA, procura-se minimizar uma função de custo, ξ, definida em termos do sinal de erro ε^p, através da modificação dos pesos sinápticos da rede. O algoritmo utilizado para alcançar o objetivo de projeto desejado foi o *Resilient Back-propagation*.

É importante salientar que os algoritmos não garantem que uma solução de máximo ou de mínimo global da função será encontrada. A única garantia que existe é que o algoritmo tenta melhorar os resultados iniciais, que podem ser fornecidos ou escolhidos aleatoriamente. Uma forma de testar se os parâmetros estão convergindo para o mesmo valor, ou para uma região próxima, é realizar diversas vezes o procedimento de aprendizado para diferentes pontos de inicialização, incluindo perturbações no conjunto de dados (ruído) e verificar se soluções semelhantes são obtidas repetidas vezes. Assim foi realizada a validação da Rede Bayesiana e da RNA. Foram consideradas para comparação as duas heurísticas e conjuntos com quantidades de amostras distintas. A RB apresentou erro médio de 8.74% e 7.46%; a RNA

apresentou erro médio de 7.53% e 6.32%; ambas para conjuntos de 1000 e 4444 amostras, respectivamente. O fato dos erros encontrados no diagnóstico das falhas serem tão próximos mostra que apesar das soluções não terem convergido para o mesmo ponto, sabe-se que as mesmas convergem ao menos para uma região próxima.

Esses erros, apesar de serem relativamente altos, são plenamente satisfatórios dentro do objetivo proposto, uma vez que se busca a identificação das causas mais prováveis para uma determinada interrupção. Cabe salientar que o principal objetivo de uma correta identificação das causas de um evento tem o objetivo macro de guiar o processo de investimento e melhoria na distribuição de energia. Sendo assim, o valor do erro em si, não compromete a função maior do sistema.

Apesar do erro obtido através do uso de RNA ser inferior ao das Redes Bayesianas, essa última apresenta uma vantagem significativa por mostrar com mais clareza as relações de causa e efeito entre os indícios coletados em campo e a causa do desligamento. Além disso, as Redes Bayesianas têm uma apresentação mais intuitiva que permite realizar uma sintonia por especialistas, podendo melhorar os resultados que foram obtidos somente através de treinamento.

4. Conclusão

A velocidade de desenvolvimento dos sistemas de informação tem transformado alguns setores da sociedade. O setor elétrico, em especial, tem avançado continuamente no sentido da busca por eficiência nos seus processos. Apontar causas de interrupções de energia é uma tarefa difícil para as equipes de restabelecimento de redes, e até secundário frente à necessidade do rápido restabelecimento, entretanto, os apontamentos incorretos prejudicam uma alocação adequada de investimentos e impactam no diagnóstico correto para a melhoria do desempenho dos indicadores de qualidade do sistema.

As empresas concessionárias de energia possuem enormes quantidades de informação armazenadas em seus bancos de dados, porém, muitas vezes, com dados faltosos, inconsistentes e com elevado grau de incerteza. Ainda assim, existe uma informação valiosa contida nesses bancos de dados, que pode ser melhor utilizada para estudos e análises após etapas de qualificação dos dados.

Dessa forma, a aplicação do processo de *KDD* pode ser útil nesta tarefa, possibilitando a qualificação das informações utilizadas para planejar e operar os sistemas. Neste trabalho, buscou-se a construção de uma base de dados que representasse a correta identificação de causas de desligamentos forçados em sistemas de distribuição, a partir de um universo de aproximadamente 570.000 eventos. Para alcançar esse objetivo, foram aplicadas técnicas que envolvem a manipulação e o tratamento de dados, bem como a criação de regras de classificação para a extração de conhecimento útil e interessante dessas fontes de informação, garantindo um número mínimo de informações que irá compor o conjunto de dados vinculados a um desligamento. No final do processo de *KDD* aplicado nesta dissertação restaram 8.888 eventos de interrupção não programados com condições de utilização. Isso tornou a base de dados mais confiável e adequada ao treinamento de sistemas de classificação e identificação de causas de desligamentos não programados na rede de distribuição.

Para avaliar a identificação de causas de eventos não programados foram pesquisadas e implementadas metodologias baseadas em Redes Bayesianas e Redes Neurais Artificiais, uma vez que estas técnicas são apropriadas ao aprendizado com base em amostras de dados incompletas. Ambas as metodologias não garantem que uma solução de máximo ou de mínimo global da função será encontrada, entretanto garantem uma melhoria nos resultados iniciais, que podem ser fornecidos ou escolhidos aleatoriamente. Para testar se os parâmetros estavam convergindo para o mesmo valor, ou para uma região próxima, o procedimento de aprendizado foi executado para diferentes pontos de inicialização, incluindo perturbações no conjunto de dados (ruído) e verificado se soluções semelhantes são obtidas repetidas vezes.

Como forma de avaliar a pertinência da escolha das metodologias pesquisadas, foram realizados testes de comparação das duas heurísticas com conjuntos e quantidades de amostras distintas.

Em relação aos erros, a Rede Bayesiana apresentou erro médio de 8,74% para um conjunto de 1.000 amostras e 7,46% para um conjunto de 4.444 amostras. Por sua vez, a Rede Neural apresentou erro médio de 7,53% e 6,32% para os conjuntos de 1.000 e 4.444 amostras. O fato dos erros encontrados no diagnóstico das falhas serem tão próximos mostra que, apesar das soluções não terem convergido para o mesmo ponto, sabe-se que as mesmas convergem para uma região próxima. Esses erros, apesar de serem relativamente altos, são plenamente satisfatórios dentro do objetivo proposto, uma vez que se busca a identificação das causas mais prováveis para uma determinada interrupção, e não de uma causa única ou específica.

Dessa forma, pode-se concluir que o erro apresentado na identificação da causa correta dos desligamentos utilizando Redes Bayesianas e Redes Neurais são bastante parecidos, mostrando que os resultados são coerentes apesar de se utilizar duas heurísticas distintas. Deve-se observar, no entanto, que ambas metodologias de análise de causas não substituem totalmente a indicação da causa pelo eletricista, uma vez que elas apenas transferem a identificação de causas para o operador do centro de operação, no momento do estudo de um desligamento ou no momento de uma pesquisa ao histórico. Contudo, tanto a Rede Bayesiana quanto a Rede Neural, guiam estes processos eliminando causas pouco prováveis.

Uma vantagem significativa das Redes Bayesianas é de mostrar com mais clareza as relações de causa e efeito entre os indícios coletados em campo e a causa do

desligamento. A possibilidade de inserção de conhecimento de profissionais experientes diretamente nas *CPT* dos nós da Rede Bayesiana torna essa característica especialmente útil quando o conjunto de dados utilizado para o treinamento do sistema é inconsistente ou apresenta um grau de incerteza associado, permitindo que a rede seja sintonizada manualmente. Adicionalmente, a representação gráfica de uma Rede Bayesiana é explícita, intuitiva e compreensível para uma pessoa não especialista, facilitando o entendimento do modelo.

No caso do modelo da Rede Neural utilizada nesta dissertação, o algoritmo para o aprendizado tem a função de modificar os pesos sinápticos da rede a fim de atingir um objetivo de projeto desejado. Uma vez treinada a rede, esse conhecimento é representado através de pesos e ficará armazenado como se fosse uma caixa preta no sistema, não permitindo qualquer modificação de forma manual pelo usuário.

Outra importante conclusão desta dissertação está relacionada a utilização da definição de elemento de interrupção. O conceito de elemento de interrupção vem colaborar com a identificação dos componentes que estão sofrendo maior influência do meio, ou que apresentam uma menor qualidade, e que acabam participando, direta ou indiretamente de um desligamento. As causas escolhidas para a classificação, que serão avaliadas através das metodologias implementadas, juntamente com o elemento de interrupção, constituem uma ferramenta poderosa para o operador que deseja identificar as contribuições mais significativas para os índices de qualidade de fornecimento de energia.

A implementação deste sistema em uma distribuidora de energia elétrica proporciona a possibilidade de criação de uma base de dados consistente sobre desligamentos forçados e, ao mesmo tempo, tornando possível a criação de ferramentas para diagnóstico e identificação de causas, como Redes Bayesianas e Redes Neurais Artificiais, para auxiliar no direcionamento dos investimentos em manutenção e a gestão de recursos financeiros, bem como o acompanhamento do desempenho da operação do sistema.

4.1. Trabalhos Futuros

Esta dissertação explorou um tema cada vez mais importante para as empresas de energia. Esta área de pesquisa relacionada a identificação de causas de desligamentos não programados em redes de distribuição tem um caráter multidisciplinar, envolvendo aspectos computacionais, como por exemplo banco de dados, aspectos estatísticos e probabilísticos, confiabilidade de sistemas elétricos, inteligência computacional, meta-heurísticas e aspectos de planejamento, operação e manutenção de sistemas de distribuição.

Apresentam-se a seguir alguns tópicos de trabalhos futuros para pesquisa e melhoria da metodologia desenvolvida:

- Criação de um novo modelo para classificação de causas, estratificando e organizando faltas permanentes e faltas transitórias;
- Aprofundamento da pesquisa sobre identificação de causas baseada em informações incompletas e de múltiplas bases de dados, utilizando novas técnicas de descoberta de conhecimento em base de dados, aprendizagem automática e mineração de dados;
- Utilização de fontes de dados de tempo real, suportadas pelo sistema SCADA, para melhoria na qualidade do conjunto de informações para alimentar sistema automáticos de identificação de causas de desligamentos;
- Desenvolvimento de aplicativos para a criação de novas topologias de redes Bayesianas que permitam a inserção de nós na rede com acesso a sistemas SCADA, AMR e trouble call;
- Pesquisa e implementação de outras topologias e treinamento de redes neurais, bem como estruturas que permitam, a semelhança das redes Bayesiana, inserir informaçòes de sistemas SCADA, AMR e trouble call;
- Desenvolvimento de metodologias que permitam estabelecer uma correlação entre os desligamentos e outras rotinas relacionadas ao processo da empresa, não somente manutenção, como também operação, suprimentos e comercial;

- Comparar os resultados obtidos com outras heurísticas de identificaçào, como por exemplo, árvores binárias, utilizando uma mesma base de informação validada;
- Implementação e integração das metodologias desenvolvidas para construção de um sistema de apoio a decisão de operação e manutenção.

Referências Bibliográficas

[1] Site da ANEEL – Disponível em: <www.aneel.gov.br>

[2] PRETTO, Carlos Oliva. **Sistema de Coleta e Tratamento de Informações sobre Desligamentos Não Programados Baseados em Computadores Móveis**. 2005. 87 f. Dissertação (Mestrado em Engenharia Elétrica) – Programa de Pós-Graduação em Engenharia Elétrica, Faculdade de Engenharia, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2005.

[3] PRETTO, C.O.; ROSA, M.A.; LEMOS, F.A. B.; SANTOS, T.T. Utilização De Computação Móvel para Qualificação das Rotinas de Operação e Manutenção em Redes de Distribuição. **Revista Brasileira de Controle & Automação (SBA)**, v. 17, p. 446-458, 2006.

[4] PRETTO, C.O.; RANCICH, G.V.; LEMOS F.A.B.; ROSA, M.A. Forced Outages Information Treatment System and Cause Identification Based on Mobile Computing and Neural Networks. In: 2003 IEEE Bologna Powertech Conference, 2003, Bologna – Italy. **Proceedings…** v. 1, 6 p. [in CD]

[5] ROSA, M.A.; PRETTO, C.O.; LEMOS, F.A.B.; HAFFNER, S. Forced Outage Cause Identification Using a Membership Matrix. In: IEEE/PES T&D 2004 Latin America, 2004, São Paulo. **Proceedings of IEEE/PES T&D 2004 Latin America, 2004**. v. 1, p. 1 – 6, (in Portuguese)

[6] PRETTO, C.O.; ROSA, M.A.; LEMOS, F.A.B. Data Acquisition Using Mobile Computing Technology to Enhance Operation and Maintenance Planning. In: 18th CIRED - International Conference and Exhibition on Electricity Distribution, 2005, Turin. **Proceedings...** v. 1, p. 1 - 6.

[7] CHIEN, Chen-Fu; CHEN, Shi-Lin; LIN, Yih-Shin. Using Bayesian Network for Fault Location on Distribution Feeder. **IEEE Transactions on Power Delivery**. v. 17, n. 3, p. 785 – 793, 2002.

[8] NASSAR, S.M.; COELHO, J.; WRONSCKI, V.R.; QUEIROZ, H.; GAUCHE,

E. Identificação de condições climáticas adversas através de redes bayesianas. In: IX SEPOPE - Symposium of Specialists in Electrical Operational and Expansion Planning, 2004, Rio de Janeiro. **Proceedings…** v. 1, p. 1-5.

[9] TRONCHONI, A.B.; PRETTO, C.O.; LICKS, V.; ROSA, M.A.; LEMOS, F.A.B. Forced Outage Cause Identification Based on Bayesian Networks. In: 2007 IEEE Lausanne Powertech, 2007, Lausanne - Swiss. **Proceedings...** v. 1, p. 1-6.

[10] PRZYTULA, K.W.; THOMPSON, D. Construction of Bayesian Networks for Diagnostics. In: 2000 IEEE Aerospace Conference, 2000, Big Sky – USA. **Proceedings...** v. 5, p. 193 - 200.

[11] NIKOVSKI, D. Constructing Bayesian Networks for Medical Diagnosis from Incomplete and Partially Correct Statistics. **IEEE Transactions on Knowledge and Data Engineering**, v.12, n. 4, p. 509 - 516, 2000.

[12] AGRE, G. Diagnostic Bayesian Networks. **Computers & Artificial Intelligence**. v. 16, n. 1, p. 47 - 67, 1997.

[13] OPRISAN, M.; FILIPPELLI, F.; CLARK, I.M.; BILLINTON, R. A Reliability data system for the reporting of forced outages of distribution equipment. In: WESCANEX '91 'IEEE Western Canada Conference on Computer, Power and Communications Systems in a Rural Environment', 29-30 May, 1991, p. 267-270.

[14] BILLINTON, R. System and Equipment Performance Assessment, Reliability, Security and Power Quality of Distribution Systems. 1992. In: IEE North Eastern Centre Power Section Symposium on the, 5 Apr. 1995.

[15] BILLINTON, R.; BILLINTON, J. Distribution System Reliability Indices. **IEEE Transactions on Power Delivery**. v. 4, n. 1, p. 561 – 568, 1989.

[16] WARREN, C.A. Distribution Reliability: What Is It? **IEEE Industry Applications Magazine**. v. 2, n. 4, p. 32 - 37, July/August 1996.

[17] CHOW, MO-Y.; TAYLOR, L.S. Analysis and Prevention of Animal-Caused Faults in Power Distribution Systems. **IEEE Transactions on Power Delivery**. v. 10, n. 2, p. 995 – 1001, 1995.

[18] FUKUI, C.; KAWAKAMI, J. An expert system for fault section estimation using information from protective relays and circuit breakers. **IEEE Transactions on**

Power Delivery, v. 1, n. 4, p. 83 – 90, 1986.

[19] XU, L.; CHOW, MO-Y. A Classification Approach for Power Distribution Systems Fault Cause Identification. **IEEE Transactions on Power Delivery**, v. 21, n. 1, p. 53 – 60, 2006.

[20] XU, L.; CHOW, M.-C.; GAO, X.Z. Comparisons of Logistics Regression and Artificial Neural Network on Power Distribution Systems Fault Cause Identification. In: IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications, 2005. SMCia/05. **Proceedings...** v., n., p. 128-131.

[21] VAZQUEZ, E.; CHACON, O.L.; ALTUVE, H.J. An online expert system for fault section diagnosis in power systems. **IEEE Transactions on Power Systems,** v. 12, n. 1, p. 357 – 362, 1997.

[22] LIU, Yan; SCHULZ, N.N. Intelligent System Application in Distribution Outage Management. In: 2002 IEEE Power Engineering Society Winter Meeting, 2002. **Proceedings…** v. 2, p. 833 - 837.

[23] LIU, Yan; SCHULZ, N.N. Knowledge–Based System for Distribution System Outage Locating Using Comprehensive Information. **IEEE Transactions on Power Systems**. v. 17, n. 2, p. 451 – 456, 2002.

[24] COELHO, J; GAUCHE, E.; NASSAR, S.M.; WRONSCKI, V.R.; QUEIROZ, H.; LIMA, M. de; LOURENÇO, M.C. Reliability diagnosis of distribution system under adverse weather conditions. In: 2003 IEEE Bologna Powertech, 2003, Bologna – Italy. **Proceedings…** In CD.

[25] WRONSCKI, V.R.; NASSAR, S.M.; COELHO, J.; GAUCHE, E.; QUEIROZ, H.; LIMA, M. de; LOURENÇO, M.C. Metodologias para identificar associação entre padrões climáticos e qualidade de fornecimento de energia elétrica. V Seminário Brasileiro de Qualidade de Energia, agosto de 2003, Aracaju, SE. In CD.

[26] CHOW, M.-Y.; YEE, S.O.; TAYLOR, L.S. Recognizing animal-caused faults in power distribution systems using artificial neural networks. **IEEE Transactions on Power Delivery**. v. 8, n. 3, p. 1268 – 1274, 1993.

[27] OLARU, C.; GEURTS, P.; WEHENKEL L. Data mining tools and applications in power system engineering. In: 13th Power System Computational Conference, 1999, Trondheim – Norway. **Proceedings…** v. 1, p. 324 – 330.

[28] MADAN, S.; SON, Won-Kuk; BOLLINGER, K.E. Applications of data mining for power systems. In: IEEE 1997 Canadian Conference on Electrical and Computer Engineering, 1997. **Proceedings…** v. 2, p. 403-406.

[29] HOLSHEIMER, M.; SEIBES, A. **Data Mining**, Report CSR9406, CWI, Amsterdam, 1994.

[30] PENG, J.T.; CHIEN, C.F.; TSENG, T.L.B. Rough set theory for data mining for fault diagnosis on distribution feeder. **Proceedings of IEE Generation, Transmission and Distribution, 2004**. v. 151, n. 6, p. 689-697.

[31] MORI, H. State-of-the-art overview on data mining in power systems. In: 2006 IEEE PES Power Systems Conference and Exposition, 2006, Atlanta – USA. **Proceedings...** p. 33 – 34.

[32] XU, L.; CHOW, M.-Y.; TAYLOR, L.S. Power Distribution Fault Cause Identification with Imbalanced Data Using the Data Mining-Based Fuzzy Classification E-Algorithm. **IEEE Transactions on Power Delivery**. v. 22, n. 1, p. 164 – 171, 2007.

[33] XU, L.; CHOW, M.-Y.; TIMMIS, J.; TAYLOR, L.S. Power Distribution Outage Cause Identification with Imbalanced Data Using Artificial Immune Recognition System (AIRS) Algorithm. **IEEE Transactions on Power Systems**. v. 22, n. 1, p. 198 – 204, 2007.

[34] MEYER, Paul L. **Probabilidade: aplicações à estatística**. 2º ed. Rio de Janeiro: LTC, 2000. 476 p.

[35] CAMARGO, C. Celso de B. **Confiabilidade Aplicada a Sistemas de Distribuição**. Rio de Janeiro: LTC, 1981. 206 p.

[36] TRIVEDI, KISHOR, SHRIDHARBHAI. **Probability and statistics with reliability, queuing, and computer science applications**. 2º ed. New York: John Wiley & Sons, 2002. 848 p.

[37] JENSEN, F.V. **Bayesian Networks and Decision Graphs**. New York: Springer - Verlag, 2001. 268 p.

[38] HSU, Hwei P. **Schaum´s Outline of Theory and Problems of Probability, Random Variables, and Random Processes**. New York, NY: McGraw-Hill, 1997. 306 p.

[39] BAYES, T. An essay towards solving a problem in the doctrine of chances. **Philosophical Transactions of the Royal Society of London**, 53:370–418, 1763.

[40] NORVIG, Peter; RUSSELL, Stuart. **Artificial Intelligence: a modern approach**. 2. ed. Upper Saddle River, NJ: Prentice Hall, c2003. 1080 p.

[41] COZMAN, F.G. Generalizing Variable Elimination in Bayesian Networks. In: Workshop on Probabilistic Reasoning in Artificial Intelligence, 2000, Atibaia - Brazil.

[42] Norsys Software Corporation. Netica - Application for working with belief networks and influence diagrams.

[43] MONTGOMERY, Douglas C.; RUNGER, George C. **Applied Statistics and Probability for Engineers**. John Wiley & Sons Inc, 2003. 784 p.

[44] MYUNG, J. – Tutorial on maximum likelihood estimation. **Journal of Mathematical Psychology**. v. 47, n.1, p. 90–100, 2003.

[45] MITCHELL, T.M. **Machine Learning**. New York, NY: McGraw – Hill, 1997. 414 p.

[46] HAYKIN, S. **Redes Neurais: princípios e prática**. Tradução: Paulo Martins Engel. 2. ed. Porto Alegre: Bookman, 2001. Tradução de Neural Networks: a comprehensive foundation.

[47] CORTEZ, P.; NEVES J. *Redes Neuronais Artificiais*. Universidade do Minho, Braga, Portugal, 2000.

[48] CYBENKO, G. Approximations by superpositions of a sigmoid function. *Mathematics of Control, Signals and Systems*, 2:303 – 314,1989.

[49] CYBENKO, G. Continuous valued neural networks with two hidden layers are sufficient. Technical report, Department of Computer Science, Tufts University, 1988.

[50] RIEDMILLER, M. Rprop – description and implementation details. Technical report. University of Karlsruhe, 1994.

[51] WIDROW, B.; HOFF, M.E. Adaptative switching circuits. *Institute of Radio Engineers.* In: Wescon Convention Record, Part 4, p. 96-104, 1960.

[52] JACOBS, R. Increase rates of convergence through learning rate adaptation. **Neural Networks**. vol. 1, n.o 4, pp. 295-307, 1988.

[53] TOLLENAERE, T. Supersab: Fast adaptive back propagation with good scaling properties. **Neural Networks**. vol. 3, n.o 5, pp. 561-573, 1990.

[54] ZADEH, L.A. Fuzzy Sets. **Information and Control**. v. 8, p. 338-353, 1965.

[55] REZENDE, S.O. **Sistemas Inteligentes: Fundamentos e Aplicações**. Barueri, SP: Manole, c2005. 525 p.

[56] FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R.. **Advances in Knowledge Discovery and Data Mining**. Menlo Park, CA: AAAI PRESS, 1996. 611 p.

[57] HAN, J. **Data Mining: Concepts and Techniques**. San Francisco, CA: Morgan Kaufmann, c2001. 550 p.

[58] FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine.** v. 17, n. 3, p. 37 - 54, 1996.

[59] BUCHANAN, B.G. e SHORTLIFFE, E.H. Rule-Based Expert Systems: The MYCIN Experiments of the Standford Heuristic Programming Project. Massachusetts: Addison-Wesley Publishing Company, Reading, 1984

[60] ARARIBÓIA, G. **Inteligência Artificial: Um Curso Prático**. Rio de Janeiro: Livros Técnicos e Científicos Editora Ltda, 1987.

[61] WEISS, S.M. **Predictive data mining: A practical guide**. San Francisco, CA: Morgan Kaufmann, 1998. 228 p.

[62] TAN, PANG-NING. **Introduction to data mining**. Boston: Addison-Wesley, 2006.769 p.

[63] Agência Nacional de Energia Elétrica – ANEEL. **Resolução ANEEL n. 24, de 27 de janeiro de 2000**. Estabelece as disposições relativas à Continuidade da Distribuição de energia elétrica às unidades consumidoras.

[64] BROWN, Richard E. **Electric Power Distribution Reliability**. New York: Marcel Dekker, c2002.

[65] MACQUEEN, J.B. Some Methods for classification and Analysis of Multivariate Observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA. **Proceedings…** p. 281-297.

[66] TEKNOMO, KARDI. **K-Means Clustering Tutorials**. Disponível em: <http:\\people.revoledu.com\kardi\tutorial\kMean\>.

[67] HECKERMAN, D. Bayesian Networks for Data Mining. **Data Mining and Knowledge Discovery**, v. 1, p. 79-119, 1997.

[68] WRONSCHI, V.R.; NASSAR, S.M.; COELHO, J.; GAUCHE, E.; QUEIROZ, H.L.; LIMA, M. de. Influence of Weather Variables in Continuity Levels of Electrical Power Supply – An Analysis Thought Artificial Neural Networks. In: VIII SEPOPE, 2002, Brasília. **Anais…** v. 1. p. 1-6.

[69] MURPHY, K.P. **The Bayes Net Toolbox for *Matlab***. Department of Computer Science, University of California, Berkeley, CA, 2001.