UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ

CAMPUS DE FOZ DO IGUAÇU

PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA ELÉTRICA E COMPUTAÇÃO

DISSERTAÇÃO DE MESTRADO

ALGORITMOS DE INTELIGÊNCIA COMPUTACIONAL UTILIZADOS NA DETECÇÃO DE FRAUDES NAS REDES DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

ALTAMIRA DE SOUZA QUEIROZ

FOZ DO IGUAÇU
2016

Altamira de Souza Queiroz

Algoritmos de Inteligência Computacional Utilizados na Detecção de Fraudes nas Redes de Distribuição de Energia Elétrica

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e Computação como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica e Computação. Área de concentração: Sistemas Dinâmicos e Energéticos.

Orientador: Prof. Dr. Edgar Manuel Carreño Franco

Foz do Iguaçu
2016

Algoritmos de Inteligência Computacional Utilizados na Detecção de Fraudes nas Redes de Distribuição de Energia Elétrica

Altamira de Souza Queiroz

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e Computação e aprovada pela Banca Examinadora:
Data da defesa pública: 19/02/2016.

Prof. Dr. **Prof. Dr. Edgar Manuel Carreño Franco** - (Orientador)
Universidade Estadual do Oeste do Paraná - UNIOESTE

Prof. Dr. **Ricardo Luiz Barros de Freitas**
Universidade Estadual do Oeste do Paraná - UNIOESTE

Prof. Dr. **Arnaldo Candido Junior**
Universidade Tecnológica Federal do Paraná - UTFPR

Resumo

Um dos principais problemas que enfrentam atualmente as empresas concessionárias de energia elétrica é a ocorrência de perdas de energia na rede de distribuição, causadas por fraudes e furtos de energia elétrica. Sendo que tais problemas provocam prejuízos financeiros e também colocam em risco a segurança pública, é de grande interesse das concessionárias encontrar soluções para detectar e combater fraudes nas redes de distribuição de energia elétrica. Neste conceito, o presente trabalho apresenta uma análise dos algoritmos de Inteligência Computacional para extrair conhecimento de bases de dados de informações de consumo mensal de energia elétricas de usuários de uma determinada concessionária, a fim de identificar padrões de consumo com anomalias que representem possíveis fraudes nas redes de distribuição de energia elétrica. Para detectar padrões nas curvas de consumo, foram utilizados algoritmos de Redes Neurais Artificiais e Máquinas de Vetores de Suporte. Após a criação dos modelos, estes foram testados para verificar qual seria o melhor algoritmo para a detecção de padrões de consumo com anomalias, e os resultados obtidos, foram então, comparados com uma base de dados fornecida pela concessionária com a verificação manual dos usuários. Os testes demonstraram que os algoritmos utilizados são capazes de detectar padrões nas curvas de consumo de energia elétrica, inclusive detectando situações especiais de fraudes que técnicas manuais não detectaram.

Palavras-chave: Padrões de Consumo, Perdas de Energia, Inteligência Computacional, Redes Neurais Artificias, Máquinas de Vetores de Suporte, Rede de Distribuição de Energia Elétrica.

Abstract

One of the main problems currently faced by electric utilities is the occurrence of energy losses in the distribution network caused by fraud and electricity theft. Because of the financial losses and risks to public safety, the development of solutions to detect and combat fraud in the distribution networks is of the utmost importance. This work presents an analysis of computational intelligence algorithms to extract knowledge in databases with information from monthly energy consumption to identify consumption patterns with anomalies which could represent fraud. The algorithms Artificial Neural Networks and Support Vector Machines were tested to see which one perform better on the identification consumption patterns with abnormalities. Tests have shown that the algorithms used are able to detect patterns in electricity consumption curves, including special situations of fraud that manual techniques did not detect.

Keywords: Consumption Patterns, Energy Losses, Computational Intelligence, Artificial Neural Networks, Support Vector Machines, Electric Power Distribution Network.

Dedico este trabalho ao meu pai e minha mãe
por todo o esforço e apoio que ambos tiveram comigo
para que eu pudesse concluir o mestrado,
vencendo a distância, as dificuldades e a saudade.
E também pelo carinho, paciência que dedicaram a mim.

Agradecimentos

Primeiramente, agradeço a Deus, por ter me dado força para superar todas as dificuldades encontradas neste caminho que tracei para chegar até aqui.

Quero agradecer a todos que contribuíram, de maneira direta ou indireta, ostensiva ou não, explícita ou implícita, formal ou informal, para que este trabalho pudesse ser finalizado.

Assim, agradeço ao meu pai e minha mãe por todo o apoio que sempre me deram, por terem paciência nos momentos de nervoso e impaciência que as dificuldades do mestrado e a distancia da família causaram durante este caminho. Por me ensinarem a não desistir perante os desafios, e a enfrentar todos os obstáculos com humildade sabedoria, me mostrando que mesmo que o mundo se vire contra mim, eu sempre poderei contar com meu pai e minha mãe.

Agradeço ao meu irmão pela paciência nos momentos em que queria brincar e eu não podia por ter que estudar, pelas brincadeiras, pela alegria que sempre me passou a cada palavra, a cada ligação. Por, mesmo distante, sempre estar presente, sendo meu melhor amigo e confidente. Agradeço a minha irmã e meu cunhado pelo apoio e colaboração de maneira direta e indireta durante o desenvolvimento deste trabalho.

Ao professor Edgar, que me adotou como orientanda e não me deixou desistir de alcançar o título de mestre. Que me apoiou, me ensinou e me orientou neste trabalho com carinho e dedicação. Agradeço também a professora Patricia, que junto com o professor Edgar, colaborou no desenvolvimento deste trabalho.

Aos meus familiares que sempre estiveram me dando força e apoiando, cada um de sua maneira, mas sempre ajudando para que eu pudesse terminar meu mestrado com sucesso. E também aos familiares que criticaram e duvidaram que eu chegasse ate aqui, pois as criticas só serviram para me dar força e continuar na luta.

Agradeço a todos do Programa de Pós-Graduação e UNIOESTE, professores, secretaria, e colegas de trabalho. Unidos, trabalhamos dois anos, repassando conhecimento, aprendendo uns com os outros, pesquisando e estudando muito. Agradeço também o apoio financeiro da CAPES e o espaço oferecido pelo PTI para que todo trabalho pudesse ser realizado

Enfim, agradeço a todos que diretamente ou indiretamente me apoiaram, me ensinaram um pouquinho, pois é com o conjunto de cada pouquinho que vou aprendendo, que me torno uma grande pessoa.

“Não importa quantas vezes eu caia.
Você sempre me verá levantar de novo e com um detalhe:
Cada vez mais forte!”
autor desconhecido

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O mercado de energia elétrica, assim como a competitividade no setor elétrico brasileiro vêm crescendo de forma gradual. Para acompanhar esse crescimento, as empresas concessionárias de energia elétrica buscam novas tecnologias com o objetivo de minimizar as perdas de energia que ocorrem durante o processo de distribuição e diminuir os custos com inspeções em medidores de energia elétrica (Lópes et al., 2014).

De acordo com a Agência Nacional de Energia Elétrica (ANEEL), em média, 15% da energia elétrica produzida no Brasil é perdida por furtos e fraudes nas redes de distribuição (ANEEL, 2000). Tais perdas vêm gerando uma degradação na qualidade dos serviços prestados e prejuízos na receita das empresas concessionárias de energia elétrica (ABRADEE, 2015).

Nos últimos anos, estudos sobre as perdas de energia elétrica e como combatê-las têm ganhando grande destaque nas pesquisas, uma vez que, os furtos de energia elétrica não representam apenas perdas econômicas para as empresas concessionárias de energia elétrica, mas criam também, um impacto social e colocam em risco a segurança pública através dos incêndios e acidentes fatais resultantes de ligações clandestinas e outras fraudes (Eller, 2003) (Aráujo, 2006).

De acordo com Delaiba et al. (2004), o método de detecção de anomalias nas redes de distribuição de energia elétrica é a inspeção do consumo de energia através da leitura dos medidores nas unidades consumidoras e análises dos dados obtidos.

O armazenamento dos valores das leituras mensais do consumo de energia elétrica vêm gerando um crescimento notável na quantidade de dados produzidos e armazenados em meios magnéticos e digitais pelas concessionárias. De acordo com Fayyad et al. (1996), os dados armazenados em grande escala são inviáveis de serem lidos e utilizados pelas organizações através do uso de métodos tradicionais como planilhas e relatórios, ou seja, os dados contidos nessas grandes bases de dados apresentam informações não explícitas. Sendo assim, é necessário usar meios para transformar tais dados em informações que possam ser úteis nas tomadas de decisões das empresas.

A utilização de técnicas de Inteligência Computacional (IC) permite a transformação de grandes bases de dados em informações explícitas, que podem auxiliar nas tomadas de decisões

das organizações (Neves et al., 2012). No caso de fraudes nas redes de distribuição de energia elétrica, as técnicas de IC podem auxiliar na detecção de anomalias no consumo, as quais podem ser causadas por fraudes e furtos nas redes de distribuição de energia elétrica.

1.1 Objetivos

Levando em consideração o crescimento constante das bases de dados de consumo de energia elétrica adquiridos através da leitura dos medidores de energia e o crescimento da utilização de técnicas de IC na transformação de dados em conhecimento, o presente trabalho tem como objetivo principal encontrar padrões de possíveis fraudes e furtos através da análise da curva de consumo dos usuários de energia elétrica.

Os objetivos específicos dessa pesquisa são: compreender, analisar e comparar técnicas de mineração de dados aplicadas à solução dos problemas de fraudes e furtos nas redes de distribuição de energia elétrica. E como complementação, o trabalho também tem o objetivo de comparar e analisar os resultados obtidos através das técnicas de mineração de dados a fim de detectar qual algoritmo melhor classifica os dados da base de dados utilizada e bases de dados semelhantes.

1.2 Motivação

A motivação do presente trabalho se deu pelo grande volume de dados de consumo de energia elétrica armazenados pelas empresas concessionárias de energia elétrica, bem como as limitações humanas em analisar tais dados de maneira ágil, a fim de extrair um conhecimento que possa, posteriormente, ser útil às empresas.

Neste contexto, uma alternativa que pode suprir as limitações humanas em analisar grande quantidade de dados de forma ágil e eficiente é a utilização de técnicas de IC.

Dentro das técnicas de IC, o processo de Descoberta do Conhecimento em Bases de dados (Knowledge Database Discovery - KDD) é uma das alternativas viáveis para a extração do conhecimento de bases de dados de forma rápida e automática. As informações obtidas após o processo de KDD poderão ser utilizadas pelas empresas para detectar anomalias que representem fraudes nas redes de distribuição de energia elétrica (Russell & Norvig, 2013).

A escolha por utilizar técnicas de MD foi motivada pelo fato que, tais técnicas permitem ultrapassar os limites da mente humana através de análises de grandes quantidades de dados e extração de informações relevantes, de forma que possa auxiliar na tomada de decisões das corporações.

A ferramenta WEKA, escolhida para auxiliar na resolução do problema em questão, é

uma ferramenta implementada em Java e de código aberto, que disponibiliza diversos algoritmos de IC, motivando assim, o seu uso.

As técnicas de Redes Neurais Artificiais (RNA) e Máquinas de Vetores de Suporte (MVS) apresentam a capacidade de generalização em casos de dados incompletos ou imprecisos sem sofrerem degradações nos resultados e após construídas, ambas as técnicas podem ser retreinadas com base em novos dados sem a necessidade de alteração em suas arquiteturas. Outro fato que motivou a utilização de tais técnicas de MD, foi a utilização das mesmas em trabalhos similares que obtiveram resultados satisfatórios.

Capítulo 2

Perdas nas Redes de Distribuição de Energia Elétrica

Neste capítulo é apresentada uma introdução sobre as perdas nas redes de distribuição de energia elétrica e como tais perdas podem ser classificadas. É descrito também o custo das perdas para o consumidor e também para as empresas concessionárias de energia elétrica.

Em sequência é feito uma breve descrição de padrões de consumo, apresentando exemplos de curva de consumo e seus padrões. E para complementar o capítulo, são retratadas formas de detecção e prevenção de anomalias nas redes de distribuição de energia elétrica.

2.1 Perdas nas Redes de Distribuição

O conjunto das perdas de energia elétrica é definido pela diferença entre a energia inserida na rede e a energia entregue regularmente aos consumidores finais. Esse conjunto de perdas é chamado de Perdas Globais e podem ocorrer tanto no sistema de transmissão quanto no sistema de distribuição de energia elétrica.

As maiores perdas ocorrem nos sistemas de distribuição, e esse tipo de perda é um dos mais relevantes problemas encontrados pelas empresas concessionárias de energia elétrica, pois a energia é inserida nas redes de distribuição, mas não é comercializada.

Para identificar melhor as perdas ocorridas no sistema, as perdas podem ser divididas em dois grupos (Penin, 2008):

- Perdas Técnicas: ocorrem naturalmente por características próprias do sistema e também por ações internas nos materiais inerentes aos processos de transporte e situações em componentes dos sistemas elétricos, como os condutores, transformadores, medidores e equipamentos;
- Perdas Não-Técnicas: também chamadas de perdas comerciais, ocorrem principalmente

por ações externas na rede e falta de faturamento por parte da empresa distribuidora.

2.1.1 Perdas Técnicas

Sendo que as perdas técnicas são causadas pelas propriedades físicas do sistema e seus componentes, para melhor serem entendidas e identificadas, tais perdas podem ser classificadas em duas classes (Penin, 2008):

- Perdas de demanda: são calculadas a cada momento de uma determinada curva de carga, e são medidas em kW ou MW;
- Perdas de energia: são calculadas para um determinado tempo, normalmente em bases anuais e suas bases de medidas são em kWh ou MWh.

Um exemplo comum de perdas técnicas é a perda nos condutores do sistema elétrico. Tal perda é denominada de Perda Joule e ocorre quando o condutor é aquecido ao ser percorrido por uma corrente elétrica e então ocorre a transformação de energia elétrica em energia térmica (Passos, 2009).

No Brasil, os cálculos das perdas técnicas são efetuados em cada segmento do sistema das redes de energia elétrica, com o objetivo de permitir uma modelagem adequada para cada segmento e resultados com maior precisão (PRODIST, 2015).

2.1.2 Perdas Não-Técnicas ou Perdas Comerciais

As perdas não-técnicas são também conhecidas como perdas comerciais e podem ter diversas causas que resultam em problemas relacionados à falta de faturamento da energia consumida pelas distribuidoras. As causas mais comuns são (Penin, 2008):

- Falhas nos equipamentos: com o passar do tempo, e fatores naturais, como chuvas e sol, os equipamentos sofrem deterioração. Esse tipo de perda não pode ser calculado previamente por ser praticamente impossível modelar o grau de deterioração de cada equipamento e por isso, não são calculadas como perdas técnicas, mesmo sendo causadas por fatores naturais;
- Erros na leitura dos medidores: também conhecido como erro de faturamento, esse tipo de perda acontece quando o leiturista (profissional responsável por fazer a leitura dos medidores) faz uma leitura equivocada no medidor do consumidor. Geralmente a perda é compensada no mês seguinte com a leitura correta da energia consumida;
- Anomalias: medidores com anomalias podem gerar um faturamento irregular. E algumas anomalias interferem diretamente na medição da energia consumida pelo medidor da energia elétrica;

- Fraudes e Furtos de energia: conhecidos como “gato na rede” e “ratos no medidor”, são uma das maiores causas das perdas não-técnicas, os fraudes e furtos de energia consistem basicamente no ato do usuário interferir na rede para que, de alguma forma, a energia consumida não seja faturada. É considerado furto quando o consumidor faz uma ligação direta na rede de distribuição sem o consentimento da concessionária, sem o processo de faturação da energia consumida e outras atitudes similares. No caso da fraude, é considerado quando o medidor de energia é alterado, ou feito um desvio no ramal de entrada antes do medidor, e também é considerado uma fraude quando o consumidor faz religação da energia sem autorização da concessionárias e atitudes semelhantes.

De acordo com um estudo realizado pelo Departamento de Energia Elétrica da Universidade Estadual de Campinas, estima-se que 17% do volume de energia elétrica disponibilizado nos sistemas de distribuição são desviados através de perdas comerciais ocorridas no sistema, e que em números, correspondem à cerca de 7.428.000 MWh (Hernandes et al., 2013). Complementam ainda que, somadas as 59 distribuidoras de energia elétrica do Brasil, a perda decorrente à fraudes e furtos equivalem à cerca de R$ 1,2 bilhões por ano.

2.1.3 O Custo das Perdas nas Redes de Distribuição

O consumidor paga pelas perdas de energia elétrica desde o momento em que a energia sai do gerador e entra nas redes de transmissão, e com isso, o consumidor paga um valor alto por um produto não consumido.

Ressaltando que o sistema de potência elétrico é dividido em geração, transmissão e distribuição de energia elétrica, a ANEEL (2015) afirma que as perdas ocorridas durante o processo de transmissão são distribuídas em 50% para as geradoras de energia elétrica e 50% para o consumidor final. E além desse custo sobre as perdas de transmissão, as perdas não-técnicas que acontecem durante a distribuição de energia elétrica são calculadas nas revisões tarifarias periódicas e cabe a ANEEL definir qual a parcela de perdas não-técnicas de energia poderá ser repassada na tarifa de energia ao consumidor final.

Na Figura 2.1 é possível visualizar de forma simplificada o esquema de geração, transmissão e distribuição nas redes de energia elétrica e como é feito o cálculo de tais perdas:

Em busca de diminuir a taxa na conta de energia por perdas não técnicas, é importante a ajuda dos consumidores através de denúncias e notificações a fim de que as perdas não técnicas sejam reduzidas e os custos nas tarifas sejam menores.

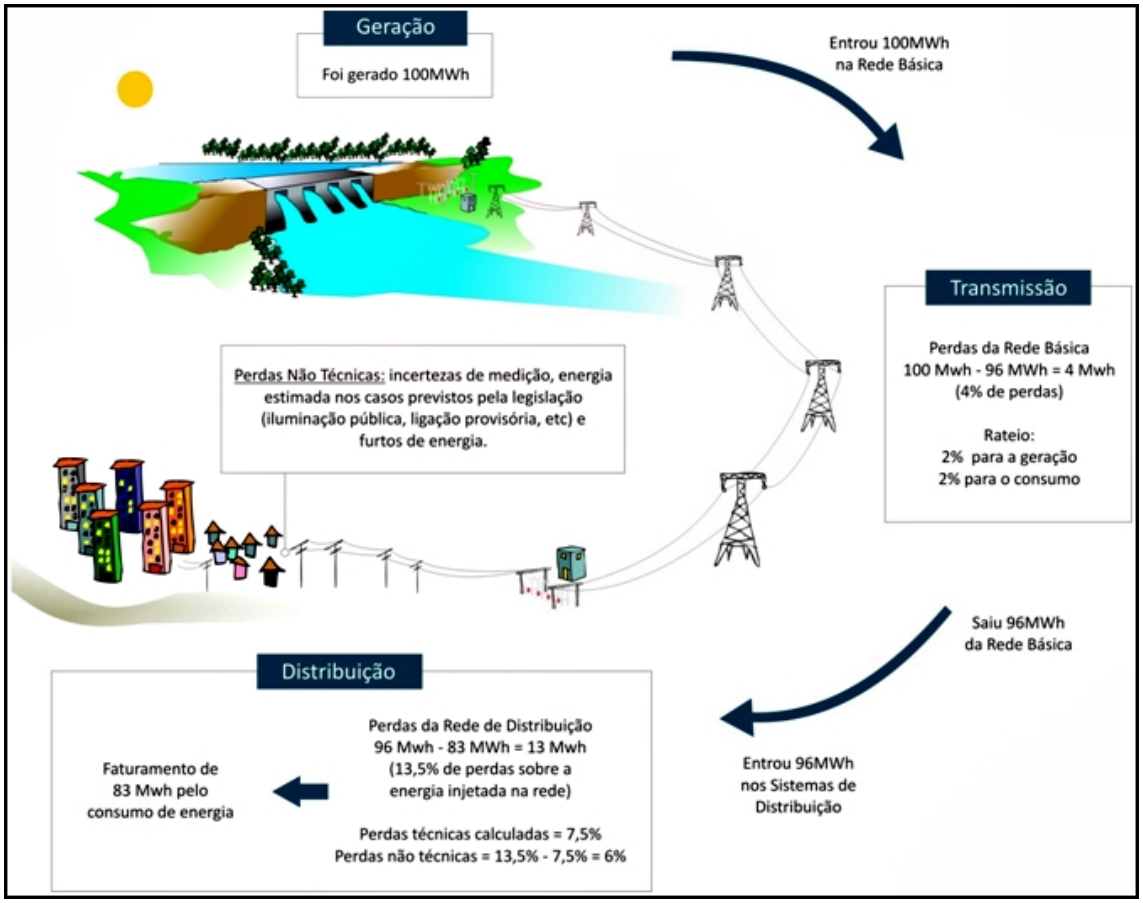

Figura 2.1: Exemplo simplificado do cálculo das perdas de energia elétrica
Fonte: ANEEL (2015)

2.2 Padrões de Consumo de Energia Elétrica

Os valores de consumo de energia elétrica consumido mensalmente pelos usuários são armazenados pelas concessionárias após a leitura nos medidores. Através da análise dos valores de energia consumida é possível classificar os consumidores por padrões de consumo. Esses padrões de consumo variam de acordo com as tipologias de edificações, situações sociais e financeiras de cada consumidor, situação econômica do País, ente outros fatores que inclui clima e determinados acontecimentos em períodos do ano, como férias escolares, viagem, entre outros (Hansen, 2000).

É importante ressaltar que independente do tipo do consumidor, o consumo de energia elétrica possui um comportamento sazonal e cíclico que pode ser observado quando é analisado a curva de consumo anual dos usuários, e tal curva mantém um comportamento de modo regular, chamado de padrão de consumo.

2.2.1 Curvas de Consumo de Energia Elétrica

A curva de consumo de energia elétrica anual pode ser apresentada em um gráfico de linhas no qual os valores de consumos mensais são colocados em uma tabela e em sequência é gerado um gráfico onde pode ser visualizado a curva de consumo de energia e analisado os padrões dessa curva.

Na Figura 2.2 pode-se observar um gráfico com uma curva de consumo de um determinado usuário de energia elétrica durante três anos. É possível observar o comportamento sazonal e cíclico durante os anos representados no gráfico.

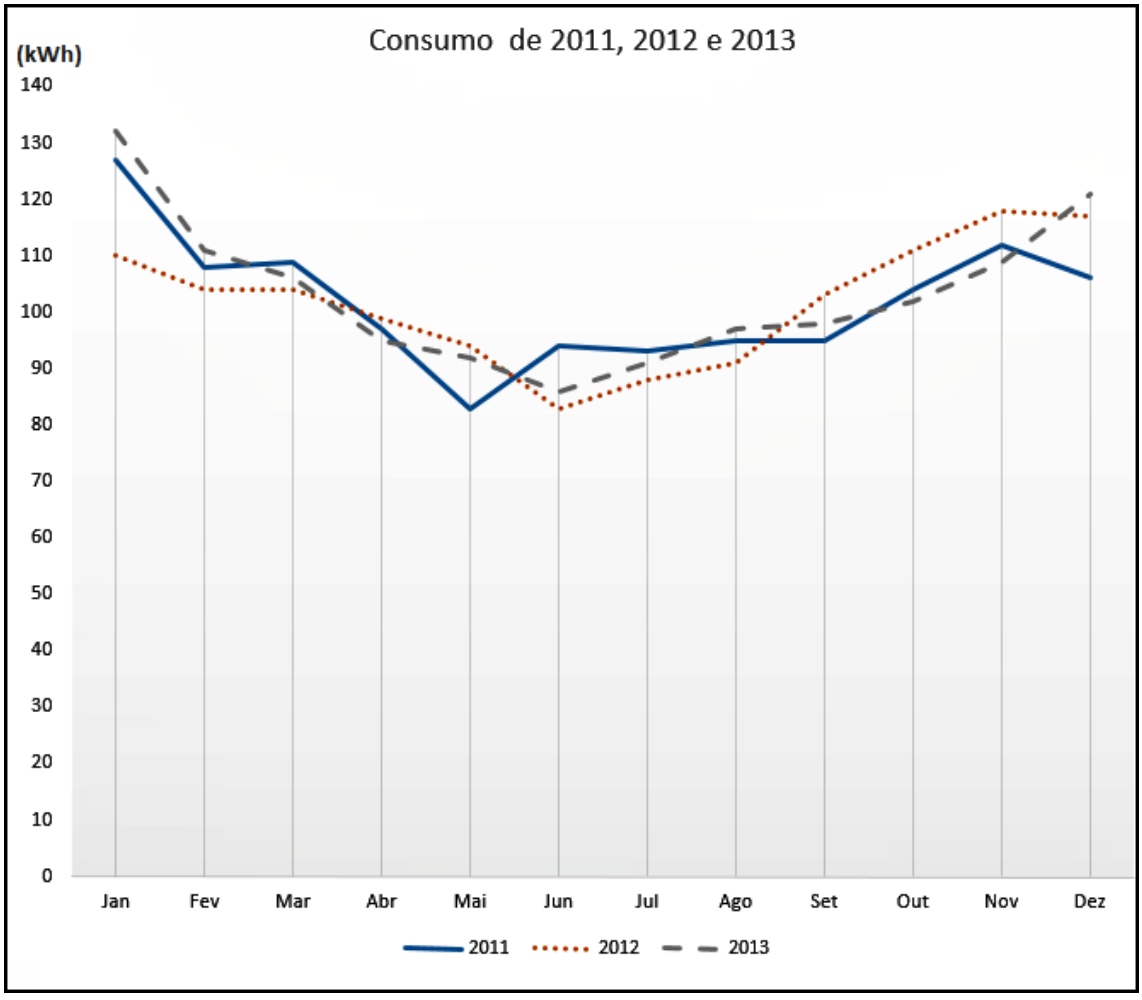

Figura 2.2: Gráfico das curvas de consumo de energia elétrica sem anomalias em um período de três anos

Através de análises das curvas de consumo de energia elétrica e levando em consideração o comportamento de tais curvas, é possível detectar padrões de consumo que podem ser utilizados para classificar consumidores e detectar anomalias nas redes de distribuição de energia elétrica.

O gráfico da Figura 2.3 apresenta a curva de consumo de três anos consecutivos, sendo que em dois anos o consumo não teve anomalias e em um dos anos, o consumo teve anomalias

a partir do mês de Maio. É possível observar que o padrão de consumo alterou totalmente e o consumo deixou de ter um comportamento sazonal e cíclico.

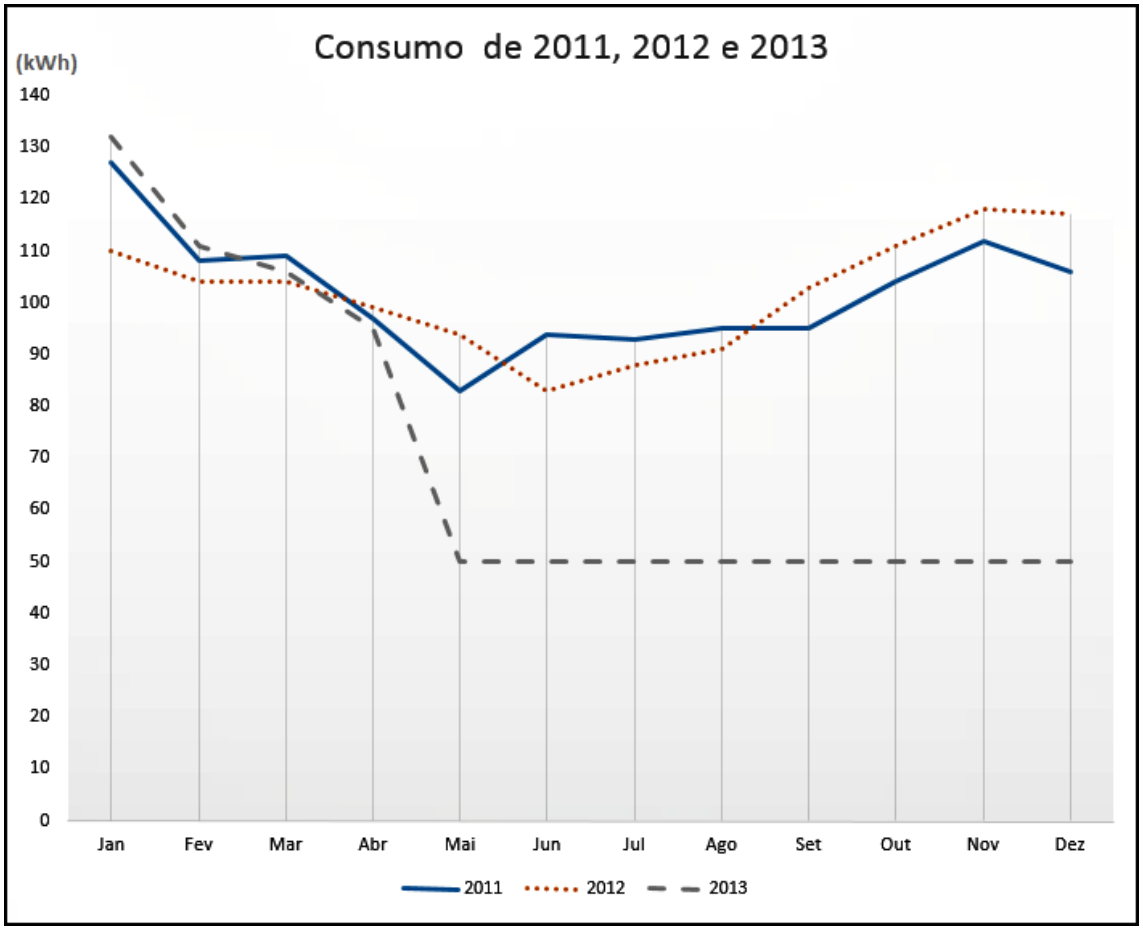

Figura 2.3: Gráfico das curvas de consumo de energia elétrica em um período de três anos sendo que em um ano possui anomalias

Lembrando que o padrão de consumo, assim como a curva de consumo de energia elétrica variam de acordo com fatores sociais, econômicos e climáticos, a Figura 2.4 apresenta uma comparação entre quatro tipos de usuários com padrões de consumo diferentes e todos sem anomalias na rede. É possível observar que o comportamento sazonal e cíclico se mantém em todas as curvas de consumo apresentadas no gráfico. Vale ressaltar as seguintes informações de cada consumidor:

- consumidor-1: família composta por um casal, sem filhos. Possui padrão social alto com diversos equipamentos elétricos em casa, incluindo ar condicionado;
- consumidor-2: família composta por um casal e 2 filhos. Possui padrão social alto com diversos equipamentos elétricos em casa, incluindo ar condicionado e jogos eletrônicos;
- consumidor-3: família composta por um casal, sem filhos. Possui padrão social médio, com equipamentos elétricos, incluindo ar condicionado;

- consumidor-4: família composta por um casal e 2 filhos. Possui padrão social baixo, com poucos equipamentos elétricos, e não possuem ar condicionado.

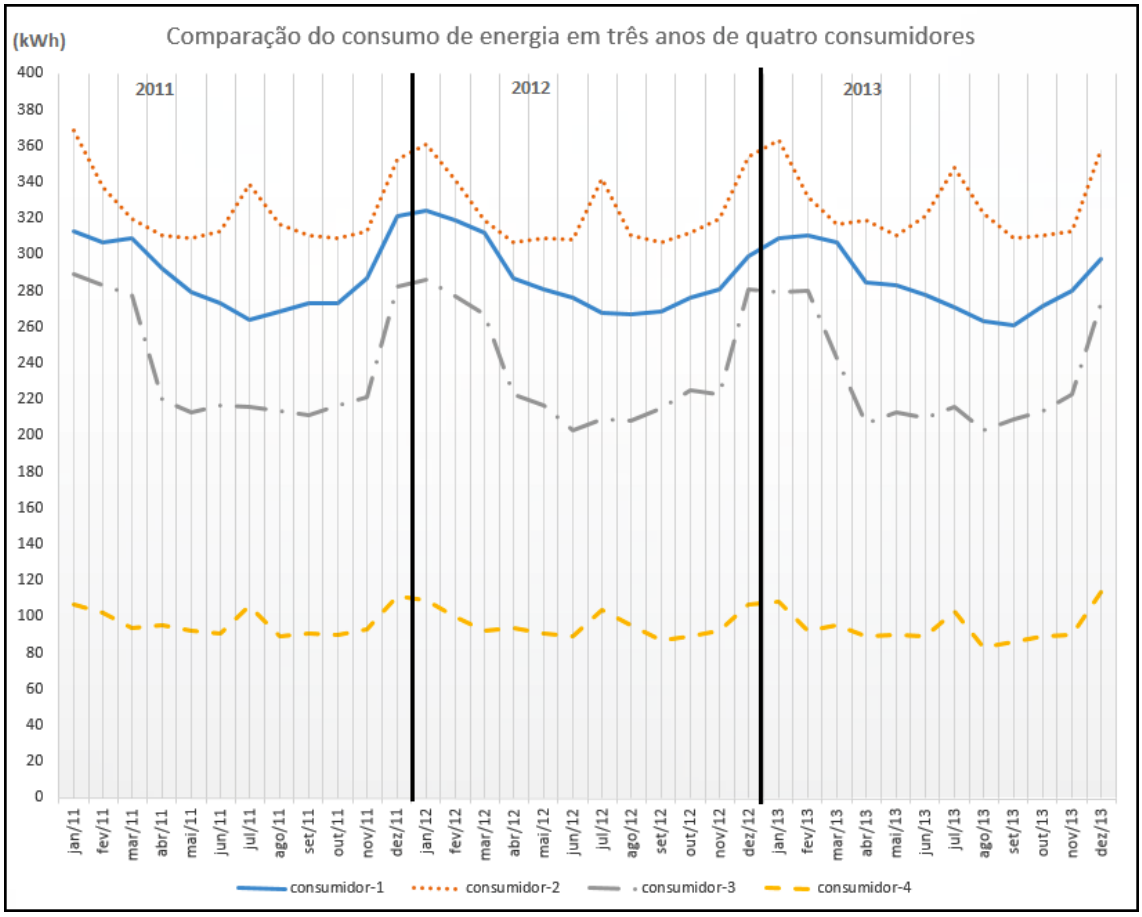

Figura 2.4: Gráfico das curvas de consumo de energia elétrica de quatro usuários diferentes, sem anomalias

Levando em consideração que as temperaturas mais elevadas ocorrem nos meses de Dezembro, Janeiro, Fevereiro e Março, que as férias escolares abrangem os meses de Dezembro, Janeiro e Julho e as informações acima citadas, é possível dizer que o consumidor-1 aumenta seu consumo de energia durante um período do ano em que o clima é mais quente pois aumentam o consumo do ar condicionado e a quantidade de banhos diários, mantendo um padrão de consumo durante os três anos apresentados no gráfico. No caso do consumidor-2, o consumo também sofre um aumento durante os meses de calor, mas a alteração no consumo também ocorre nos meses de férias escolares das crianças. O consumidor-3 possui uma curva de consumo que varia de acordo com o clima, e se mantém o padrão da curva durante os três anos analisados, e como não possuem filhos, não há aumento de consumo no mês de julho (férias escolares). E no ultimo caso, é possível observar que o consumo durante todo o ano é relativamente mais baixo que os demais consumidores, com pequeno aumento no consumo de energia elétrica nos meses mais quentes e de férias escolares.

Neste contexto, fica claro que existem padrões de curva de consumo que variam de acordo

com cada consumidor, variando de acordo com fatores externos, que incluem o clima, classe social, e outros fatores.

2.2.2 Principais Padrões de Curvas com Anomalias

Analisando as principais causas de anomalias nas redes de distribuição de energia elétrica é fácil observar o padrão de tais fraudes. Na Figura 2.5 é possível observar um caso de desvio na rede elétrica por um vizinho do usuário representado no gráfico. Nota-se que o consumo aumenta em degraus constantemente durante todo o ano. Este é um caso de fraude que não causa prejuízo financeiro para a concessionária de energia elétrica, pois a energia está sendo passada no medidor de um usuário e consumida por outro, ou seja, de qualquer forma, um usuário está pagando pela energia consumida, mas este tipo de fraude causa transtorno a sociedade, além de colocar em risco a segurança pública.

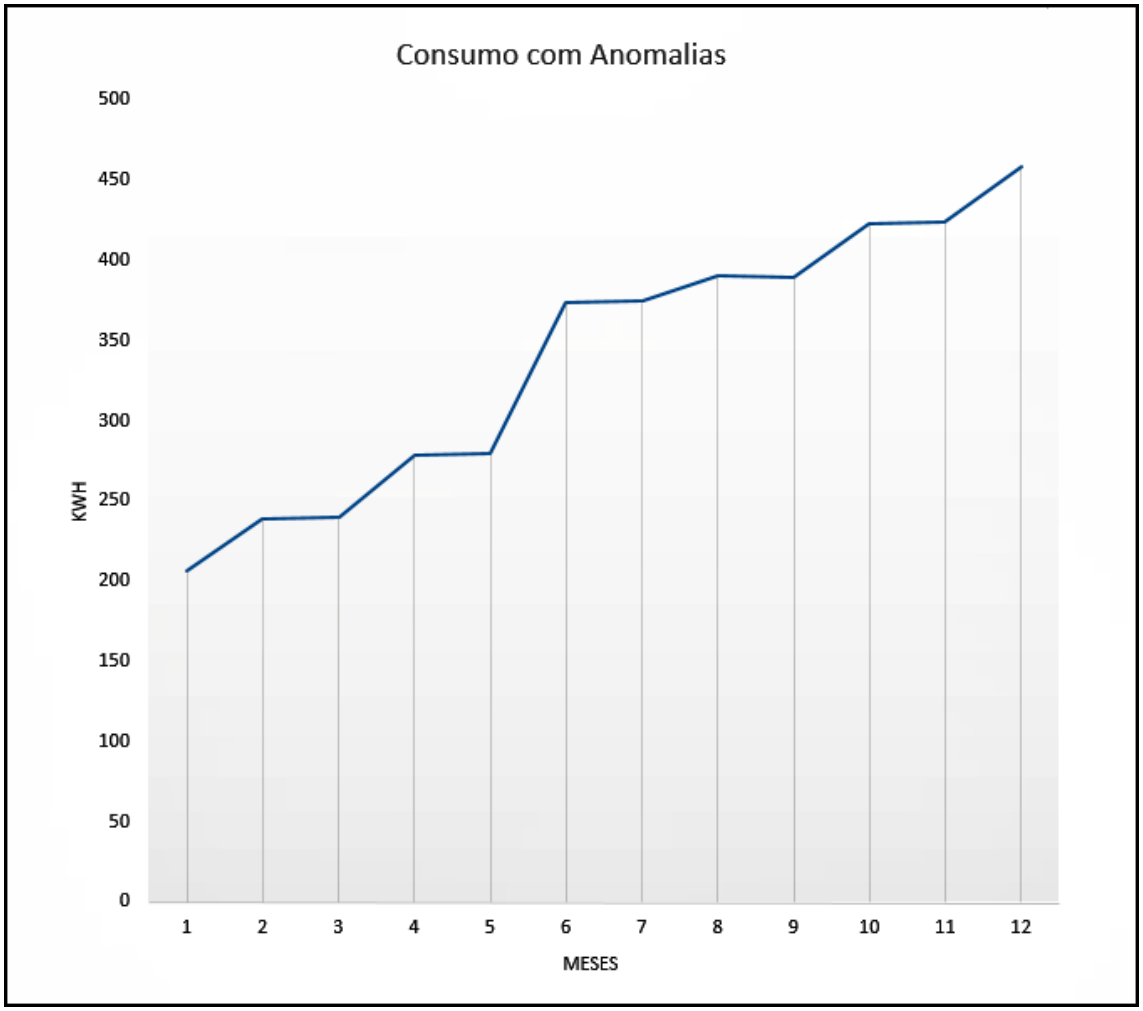

Figura 2.5: Gráfico das curvas de consumo de usuário com anomalia causada por ligação direta por outro usuário após a medição

Outro caso comum de fraudes nas redes de distribuição é medidores fraudados com freios

(furos, e/ou agulhas) no disco de medição. Esse tipo de fraude, representado na Figura 2.6, faz com que o disco medidor de energia fique travado em um determinado valor, e não meça toda a energia consumida. Este padrão de curva de consumo também acontece quando é realizado ligação direta antes do medidor de energia e o usuário utiliza a energia passada pelo medidor apenas para determinados fins, diminuindo assim, o consumo registrado pelo medidor.

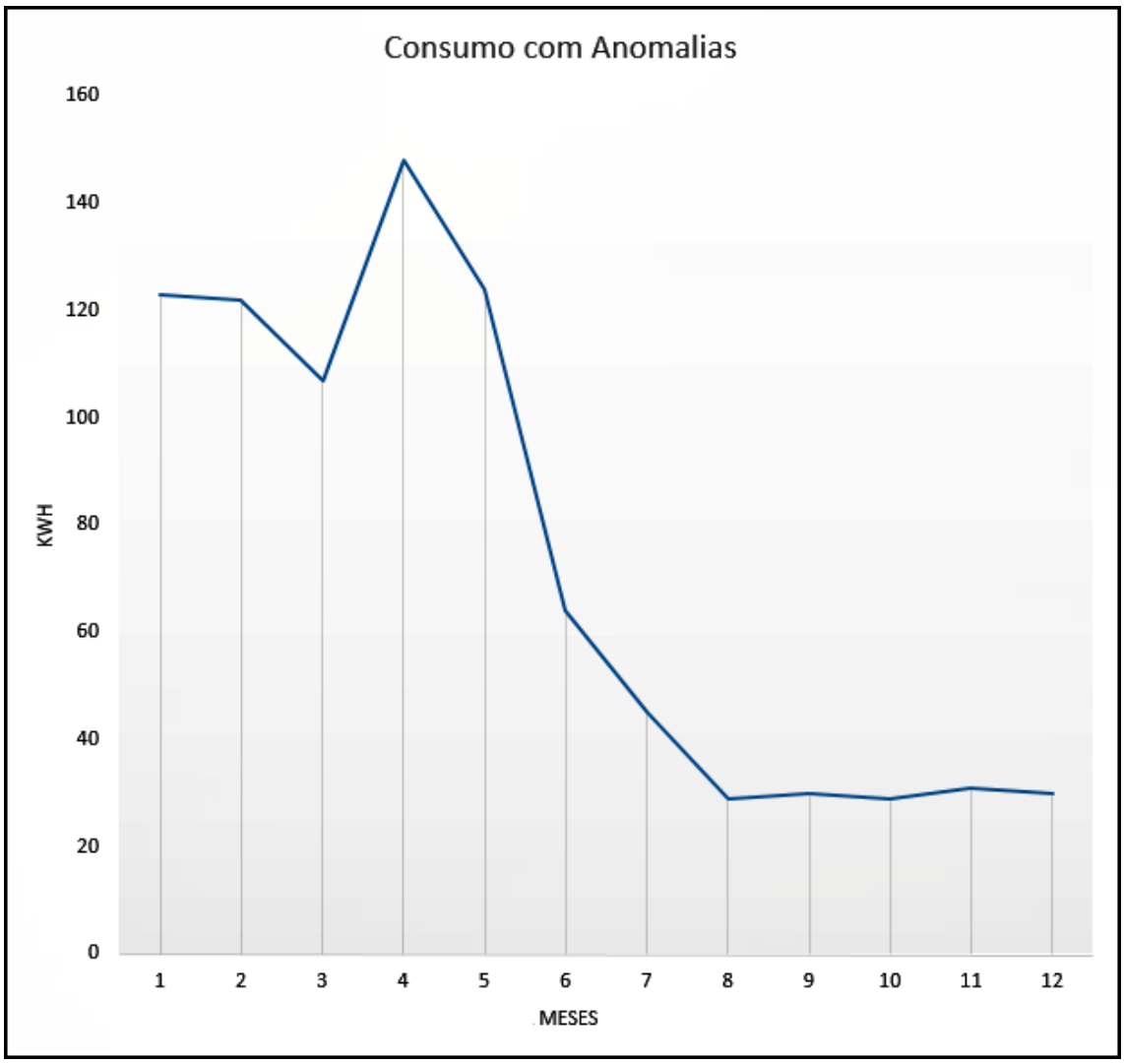

Figura 2.6: Gráfico das curvas de consumo de energia elétrica de usuário com medidor fraudado

Um tipo de padrão de consumo com anomalia, mas que analisando a longo prazo não traz prejuízos financeiros a concessionária, é o caso de falta de leitura nos medidores. Supondo que por algum motivo a leitura não é realizada mensalmente em um determinado usuário, e o mês não contabilizado é passado pelo consumidor para a concessionária, ou apenas feito uma expectativa do consumo. Pode acontecer de, o mês que não teve a leitura, receber um valor mais alto ao valor consumido, e no mês seguinte, quando realizado a leitura correta, o consumo ficar inferior devido ao consumidor já ter pagado no mês anterior, por uma energia ainda não consumida. Este padrão pode também ocorrer de modo inverso, ou seja, o valor de consumo gerado através de expectativas ser menor que o valor consumido, e o mês seguinte, quando

realizado a leitura correta, o consumo ser superior ao valor consumido no mês. O padrão da curva de consumo desse tipo de acontecimento, que também é notado como anomalia nas redes está representado na Figura 2.7.

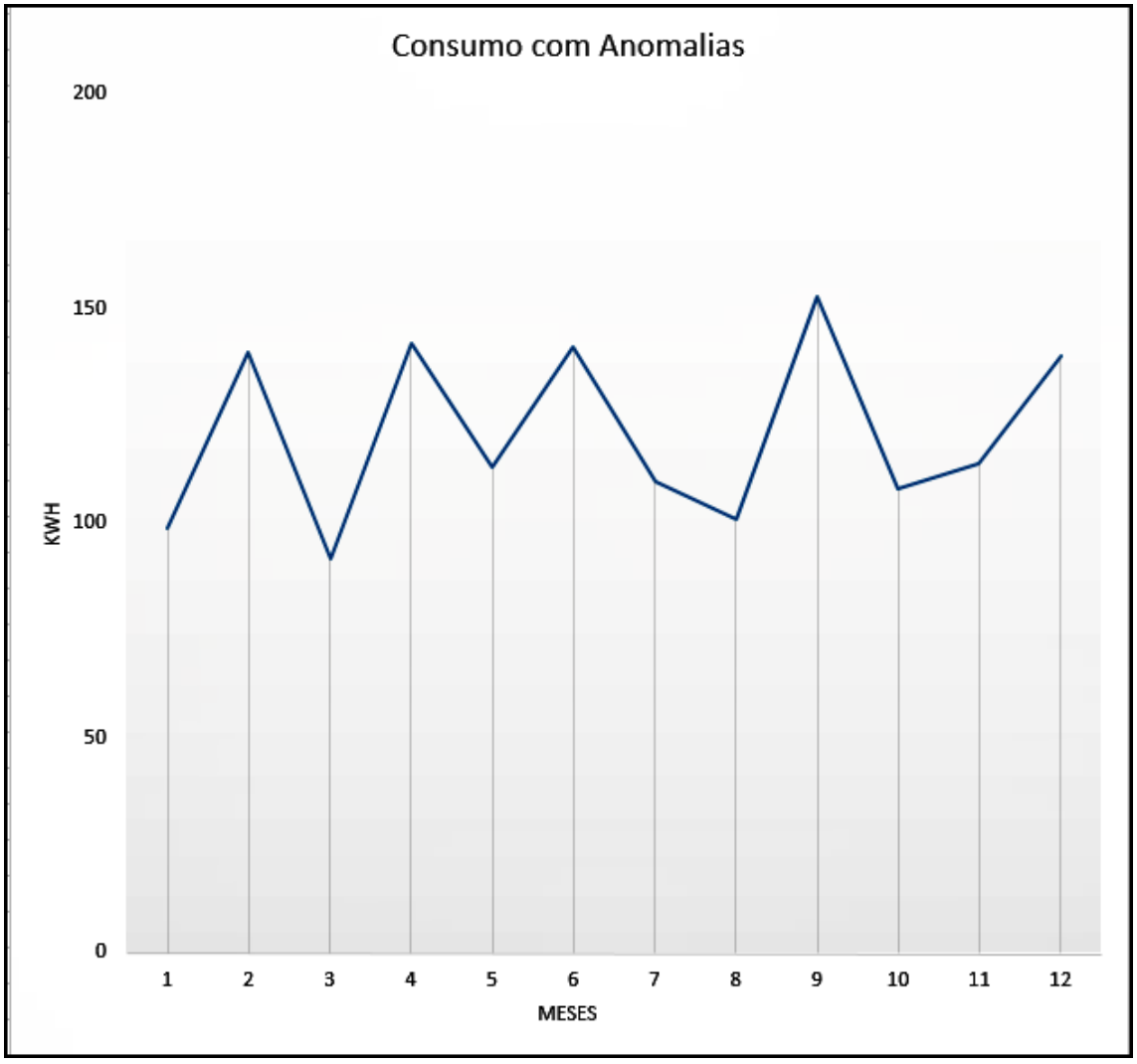

Figura 2.7: Gráfico das curvas de consumo de energia elétrica de usuários com meses sem a leitura no medidor

Nota-se que o padrão das curvas quando existe algum tipo de fraude nas redes de distribuição de energia elétrica é notável se comparado com curvas de consumo normais. Vale lembrar que os tipos de padrões de consumo, com anomalias ou não, variam de consumidor para consumidor, mas mantém sempre um padrão de consumo ao longo do ano.

Neste conceito pode-se afirmar que com ferramentas adequadas de analises de dados é possível detectar padrões de consumo e anomalias nas curvas de consumo as quais podem representar fraudes e furtos nas redes elétricas.

2.3 Detecção e Prevenção de Anomalias nas Redes de Distribuição de Energia Elétrica

Os problemas de anomalias nas redes elétricas, em especial, os casos de fraudes nas redes de distribuição, geram prejuízos tanto para o consumidor quanto para as distribuidoras e ainda colocam em risco a segurança pública de toda a comunidade. De acordo com Coura (2015), no ano de 2014, 299 pessoas morreram em acidentes que envolveram contato com fio de energia elétrica. E em uma pesquisa realizada pela ABRADEE (2015), as ligações clandestinas nas redes de distribuição de energia elétrica ocupam o segundo lugar nas ocorrências com mortes por acidentes na rede elétrica, perdendo apenas para acidentes em construções e manutenções de edifícios.

Com o objetivo de prevenir e detectar fraudes, as empresas concessionárias de energia elétrica utilizam estratégias reativas e proativas para localizar, prevenir e evitar tais perdas (Lópes et al., 2014).

As estratégias reativas mais comuns utilizadas pelas empresas concessionárias de energia elétrica são as fiscalizações de denúncias feitas pelos próprios consumidores e os atendimentos de emergência criados por acidentes nas redes elétricas.

Como estratégias proativas para diminuir as perdas nas redes de distribuição, as concessionárias contam com inspeções manuais nas unidades consumidoras e análises estatísticas das informações obtidas através da leitura dos medidores.

Ainda, para prevenir acidentes com fios de rede de energia elétrica, incluindo os casos de fraudes e furtos, as empresas concessionárias investem em campanhas para alertar e conscientizar a população sobre os métodos de prevenção de acidentes e os riscos de acidentes nas redes elétricas.

2.4 Conclusão

As perdas nas redes de distribuição podem ser classificadas de dois modos, perdas técnicas e perdas não-técnicas, sendo que a maior parte dessas perdas estão classificadas como perdas não-técnicas e são causadas principalmente por furtos e fraudes nas redes de distribuição.

As perdas nas redes de distribuição de energia elétrica causam prejuízos tanto para as concessionárias quanto para os consumidores, pois tais perdas são divididas entre a empresa concessionária e o usuário final, que pagam por uma energia não consumida. E além disso, fraudes e furtos nas redes de distribuição não causam apenas prejuízos financeiros, mas também são um problema de segurança pública.

Levando em consideração que o objetivo principal desse trabalho é encontrar padrões de

consumo com anomalias nas redes de energia elétrica, este capítulo também apresentou padrões de consumo e análises da curva de consumo como exemplo do que será realizado no presente trabalho.

Com o objetivo de encontrar padrões em grandes bases de dados de consumo de energia elétrica, e levando em consideração que técnicas manuais de análises são insuficientes para analisar grandes bases de dados, o presente trabalho utiliza Inteligência Computacional para a detecção de padrões de consumo de energia elétrica.

Capítulo 3

Inteligência Computacional

Neste capítulo serão descritas alguma definições de Inteligência Computacional encontradas na literatura e apresentado também, as principais colaborações para a Inteligência Computacional. Ainda neste capitulo, é descrito um breve histórico da Inteligência Computacional e do Estado da Arte de Inteligência Computacional.

3.1 Inteligência Computacional

Inteligência Computacional é a área da ciência que estuda teoria e aplicações inspiradas na natureza, como Redes Neurais, Lógica Nebulosa e Computação Evolucionária.

Nas palavras de Coppin (2013), Inteligência Computacional é o estudo dos sistemas que agem de um modo inteligente. O autor ainda complementa que a Inteligência Computacional envolve a utilização de métodos baseados em comportamentos inteligentes de humanos e outros animais para solucionar problemas complexos.

De acordo com Sá (2010), a Inteligência Computacional é o estudo, projeto e avaliação de agentes inteligentes, também conhecidos como sistemas inteligentes. Ainda esclarece que sistemas inteligentes são sistemas que imitam o comportamento humano, tais como: aprendizado, percepção, raciocínio, evolução e adaptação.

A utilização do comportamento natural aliado ao poder computacional dos computadores atuais resultam em ferramentas importantes em diversas áreas, como no reconhecimento de padrões, classificação, previsão de séries temporais e aproximação de funções (Silva, 2012).

Russell & Norvig (2013) refletem que as inúmeras definições de Inteligência Computacional podem ser dividas de acordo com o que elas se relacionam, podendo ser com o "processo de pensamento e raciocínio" ou/e com o "processo do comportamento e atitudes".

O objetivo científico da Inteligência Computacional, de acordo com Silva (2012), é o entendimento dos princípios que induzem o comportamento inteligente em sistemas naturais e artificiais.

3.2 Colaborações para a Inteligência Computacional

Filosoficamente falando, Aristóteles foi o primeiro estudioso a criar um conjunto rigoroso de leis que governam a parte racional da mente humana. O sistema desenvolvido por ele permitia gerar conclusões computadorizadas a partir de premissas iniciais (Russell & Norvig, 2013). Ainda dentro da filosofia, um elemento vital para a inteligência computacional é a conexão entre conhecimento e ação, pois apenas quando é compreendido a justificativa de uma ação, pode-se construir um agente cujas ações sejam justificáveis.

Na matemática, a Inteligência Computacional é formalizada junto de três áreas fundamentais: a lógica, a computação e a probabilidade. Acredita-se que o primeiro algoritmo inteligente e não trivial foi desenvolvido por Euclides e tinha como objetivo calcular o maior divisor comum (Russell & Norvig, 2013). Resumidamente falando, a matemática contribuiu com a Inteligência Computacional através de diversas teorias estudadas por pesquisadores como James Bernoulli, Thomas Bayes, Blaise Pascal, entre outros.

Na economia, a "teoria da decisão", a qual combina a teoria da probabilidade com a teoria da utilidade a fim de oferecer estruturas para as tomadas de decisões sobre as incertezas, teve grande auxílio a inteligência computacional levando em consideração que um de seus principais objetivos é auxiliar as organizações nas tomadas de decisões (Luce & Raiffa, 1957).

O mapeamento e estudos desenvolvidos do sistema nervoso, em especial do cérebro, tornaram possíveis medições que correspondem a aspectos de processos cognitivos em ação, e são utilizados principalmente nas técnicas de Redes Neurais Artificiais (Russell & Norvig, 2013).

Uma das principais colaborações da psicologia à Inteligência Computacional é o trabalho de William sobre a psicologia cognitiva, que tem uma visão do cérebro como um dispositivo de processamento de informações e especifica três passos fundamentais de um agente inteligente baseado no conhecimento (Craik, 1943):

- o estimulo deve ser traduzido em uma representação interna;
- a representação é manipulada por processos cognitivos para derivar novas representações internas;
- as representações são novamente traduzidas em ações.

A engenharia de computadores é uma área de grande importância para a Inteligência Computacional levando em consideração que para a Inteligência Computacional ter sucesso, é necessário a utilização de inteligência e artefatos (computadores) de alta qualidade. E desde a época da criação do primeiro computador, a criação de hardwares com maiores potências e capacidades de processamento vêm evoluindo constantemente, permitindo assim o desenvolvimento de softwares de Inteligência Computacional com processamento em tempo real (Russell & Norvig, 2013).

3.2.1 A História da Inteligência Computacional

A história da Inteligência computacional teve inicio na década de 1940 e a partir de então vem sofrendo avanços que colaboram em diversos sentidos para a evolução da Inteligência Computacional. Vale apontar alguns acontecimentos que marcaram a história da Inteligência Computacional, entre eles:

- Na década de 1940, McCulloch & Pitts (1943) propôs um modelo de neurônio artificial onde cada neurônio se caracterizava por estar ligado ou desligado (Norvig, 1992);
- Ainda na mesma década, Hebb (1949) descreveu uma regra de atualização para alterar as intensidades entre as conexões de um neurônio artificial. O "Modelo de Hebb"se tornou uma referencia influente em estudos até os dias de hoje;
- Em 1956, John McCarthy reuniu pesquisadores interessados nas teorias dos autômatos, redes neurais e estudos da inteligência em um seminário com o objetivo de prosseguir com o processo de aprendizado e inteligência. Após esse seminário, ficou definido o termo "Inteligência Artificial/ Computacional" como uma nova área de pesquisa, surgindo assim, oficialmente, a Inteligência Artificial como novo campo de pesquisa (**?**);
- No ano de 1958, John McCarthy contribuiu com três realizações cruciais na história da Inteligência Computacional: o desenvolvimento da linguagem de alto nível **Lisp**, a qual se tornou a linguagem de programação predominante em Inteligência Computacional pelos 30 anos seguintes, o desenvolvimento do compartilhamento de tempo(*time sharing*), desenvolvido para solucionar o problema de computadores com recursos escassos e dispendiosos e a publicação do artigo *Programs With Common Sense*, o qual descrevia o funcionamento do primeiro sistema de Inteligência Computacional completo, conhecido como *Advice Taker* (Russell & Norvig, 2013);
- Em 1972 foi criado o *General Problem Solver - GPS*, o primeiro programa a incorporar técnicas de Inteligência Artificial, o qual serviu de base para desenvolvimento de diversas novas tecnologias (Newell & Simon, 1972);
- Em 1970, foram desenvolvidos os sistemas especialistas, os quais possuíam conhecimento específico e detalhado de determinado assunto e poderia auxiliar nas tomadas de decisões (Shortliffe, 1972);
- Em 1980, a utilização de sistemas especialistas e inteligência computacional passaram a ser comercializados em grande escala. Ainda nesta mesma década, a Inteligência Computacional passou a ser considerada um método cientifico com grande crescimento em pesquisas, seminários e congressos de pesquisadores de áreas relacionadas (**?**);
- A partir do século XXI, passaram a ser disponibilizadas grandes bases de dados para estudos e treinamentos de agentes inteligentes, resolvendo o problema da falta de dados para testes (Russell & Norvig, 2013).

3.3 Estado da Arte

Hoje, a Inteligência Computacional está espalhada por todas as áreas e faz parte do dia a dia de todos, utilizando diversas técnicas e algoritmos para ser implementada (Coppin, 2013).

Alguns exemplos podem ser citados sobre o estado da arte da Inteligência Computacional (Russell & Norvig, 2013):

- **Veículos robóticos:** carros sem a necessidade de um motorista já estão sendo desenvolvidos em diversos países e são capazes de detectar o ambiente através de radares, câmeras, telêmetros a *laser* e computador de bordo, que comanda a pilotagem do veículo. Tais carros são capazes de se locomoverem nas ruas com velocidade controlada, respeitando as regras de trânsito, reconhecendo pedestres e outros veículos;
- **Reconhecimento de voz e transcrição da fala em textos:** sistemas com reconhecimentos de voz são utilizados em atendimentos de *call centers*, seleção de opções em compras via telefone, execução de tarefas, entre diversas outras funções no nosso dia a dia. A transcrição de fala em textos também é uma tecnologia de Inteligência Computacional muito utilizada em celulares, e atualmente em computadores para transcrever a fala do usuário;
- **Jogos:** jogos de tabuleiros foram os pioneiros em utilizar Inteligência Computacional para simular o jogador adversário. Mas a utilização de tecnologias inteligentes em jogos não se limita apenas em tabuleiros, sendo usada atualmente em jogos online, relação homem-máquina, simulações, entre outras tecnologias de jogos;
- **Combate ao spam e vírus:** através de algoritmos de aprendizagem, hoje é possível classificar mensagens de vírus e spam em correios eletrônicos com até 90% de acertos;
- **Robótica:** robôs treinados para executar tarefas do dia a dia, assim como também trabalhos em grandes industrias vem surgindo cada dia mais, e com um nível de inteligência crescente. O treinamento de tais robôs para adquirirem inteligência e executarem determinada atividade vem sendo possível devido a vasta quantidade de dados de diferentes domínios disponível;
- **Identificação de Padrões**: através da mineração de dados, a inteligência computacional tem avançado de forma notável no processo de identificação de padrões. Por exemplo, padrões de imagens e padrões de comportamento.

3.4 Conclusão

Com o estudo realizado através de pesquisa bibliográficas sobre Inteligência Computacional, observa-se que a utilização dessa tecnologia é ampla, podendo ser utilizada em diversas

áreas. Nota-se também que a Inteligência Computacional vem sendo usada desde muitos anos e evoluído de maneira notável, se tornando uma tecnologia utilizada com grande frequência no dia a dia das pessoas.

Neste sentido e levando em consideração a abrangência da Inteligência Computacional, o presente trabalho faz a utilização de inteligência computacional para detectar padrões de consumo em grandes bases de dados através de técnicas de Mineração de dados. No próximo capítulo, será apresentado um processo específico dentro da Inteligência Computacional, a Descoberta do Conhecimento em Bases de Dados, que será utilizada para detectar padrões de consumo em bases de dados de consumidores de energia elétrica.

Capítulo 4

Descoberta do Conhecimento em Bases de Dados

Neste capítulo são fornecidos conceitos do processo de descoberta do conhecimento em bases de dados. Ainda neste capítulo, é realizado um estudo conceitual de cada etapa desse processo, enfatizando as principais funções de cada etapa citando a opiniões de diferentes autores.

4.1 Descoberta do Conhecimento em Bases de Dados

A Descoberta do Conhecimento em Base de Dados (*Knowledge Discovery in Databases - KDD*) é um processo de Inteligência Computacional capaz de extrair o conhecimento de bases de dados através de estudos criteriosos e apresentá-los de maneira que possa ser útil nas tomadas das decisões das organizações (Coppin, 2013).

Fayyad et al. (1996) compreende que o objetivo do processo de KDD é desenvolver métodos e técnicas de extração do conhecimento a partir de dados armazenados em grandes bases de dados das corporações. Com esse objetivo, o processo de KDD envolve diversas áreas do conhecimento, tais como: estatística, matemática, banco de dados, inteligência artificial, reconhecimento de padrões, entre outras (Castanheira, 2008).

O processo de KDD é composto por um conjunto de atividades sequenciais e permite identificar padrões e/ou modelos novos, com alto potencial para serem utilizados nas tomadas de decisões (Fayyad et al., 1996).

4.2 Fases da Descoberta do Conhecimento em Base de Dados

O processo de KDD é interativo e iterativo, e na literatura é possível encontrar diversas formas de dividir as etapas do processo de KDD. De acordo com Adriaans & Zantinge (1996), a extração do conhecimento em base de dados é constituída pelas seguintes etapas:

- Definição do problema;
- Enriquecimento dos dados;
- Codificação dos dados;
- Mineração dos dados (MD);
- Relatório e divulgação do conhecimento adquirido.

A definição do conjunto de atividades que compõe o processo de KDD mais utilizada é a de Fayyad et al. (1996), a qual defini que o processo de KDD inicia-se com a seleção dos dados, em sequência executa-se o preprocessamento. Logo é realizado o processo de transformação dos dados, e em continuação, executa-se o processo de mineração de dados, e por fim, o resultado do processo de MD é então interpretado e avaliado de forma que o conhecimento adquirido possa ser compreendido, conforme mostra a Figura 4.1:

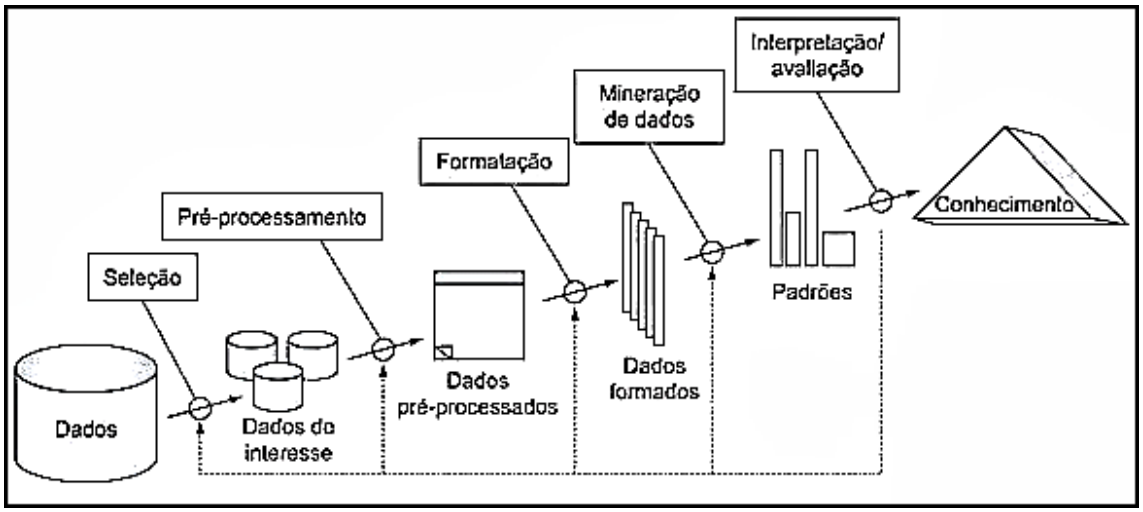

Figura 4.1: Fases de um processo de KDD
Fonte: Fayyad et al. (1996)

4.2.1 Seleção de Dados

Sendo a primeira etapa do processo de KDD, a seleção dos dados inicia-se com o agrupamento dos dados, de forma organizada, a fim de selecionar os dados que serão úteis para a extração do conhecimento em determinado problema (Castanheira, 2008).

Para que a atividade de seleção dos dados obtenha resultados adequados no processo de extração do conhecimento, é necessário que os objetivos da execução do processo de KDD estejam bem definidos, e assim, que os dados escolhidos levem a resultados úteis para se obter os resultados esperados (Cruz, 2008).

De acordo com Almeida & Dumontier (1996), a atividade de seleção dos dados é criteriosa, pois os dados úteis podem estar disponíveis em formatos diferentes, sem rótulos e desorganizados.

Um dos principais problemas da atividade de seleção dos dados é descobrir onde encontrá-los. Castanheira (2008) afirma que a maioria dos sistemas de gerenciamento de dados dos dias de hoje são particulares, mas existe uma grande tendência que as empresas implementem novos bancos de dados que sejam direcionados à dar suporte as pessoas responsáveis por extrair o conhecimento da base de dados e utilizá-los para auxiliar nas tomadas de decisões das empresas.

4.2.2 Pré-processamento dos Dados

O pré-processamento dos dados, também conhecido como processo de limpeza dos dados no processo de KDD é a etapa responsável pelo tratamento de dados errados e omissos, organização dos dados não uniformes e verificação de dados faltantes e/ou incompletos.

Segundo Mannila (1996), a fase de pré-processamento é a mais complexa do processo de KDD, podendo utilizar 80% de todo o tempo utilizado no processo completo.

Dentro da atividade de pré-processamento, durante a integração dos dados heterogêneos, eliminação da incompletude e verificação da consistência dos dados, podem aparecer problemas que são específicos de cada aplicação, exigindo soluções de um especialista com domínio da aplicação e dos dados (Castanheira, 2008).

É nesta fase que os dados duplicados ou corrompidos são identificados e eliminados, podendo assim, diminuir o tempo de processamento na etapa de MD.

A etapa de pré-processamento dos dados pode ser subdividida em duas etapas:

- atividades que só devem ser executadas por especialistas do domínio dos dados, tais como: rotulação de dados e verificação de consistência dos dados;
- atividades que são independentes do domínio dos dados, tais como: verificação de dados redundantes e incompletos.

4.2.3 Formatação dos Dados

A formatação dos dados, também conhecida como transformação, é a etapa do processo de KDD que armazena os dados de forma adequada para serem lidos pelos algoritmos de MD (Cruz, 2008).

Essa fase é realizada levando em consideração o algoritmo escolhido para ser aplicado na etapa de MD, pois dependendo do algoritmo, algumas limitações podem ser impostas à

base de dados, tais como a normalização de dados quantitativos e transformação de atributos qualitativos em quantitativos (Castanheira, 2008).

Batista (2003) defende que se uma base de dados for composta por dados quantitativos e qualitativos, para utilizá-la em redes neurais é ideal que se crie um nó para cada atributo qualitativo, sendo estes dados desmembrados em *n* atributos binários, para *n* valores diferentes na base qualitativa.

De forma geral e resumida, essa etapa do processo de KDD transforma os dados para a forma mais adequada a ser utilizada pelo algoritmo escolhido para utilização no processo de MD.

4.2.4 Mineração de Dados

A mineração de dados inicia-se com a escolha do algoritmo a ser utilizado. Essa escolha é feita levando em consideração o objetivo da execução do processo de KDD sobre uma determinada base de dados.

Esta etapa é considerada a mais importante no processo de KDD, pois, de acordo com Possa et al. (1998), o processo de MD é capaz de ampliar de 8, que é o número de comparações simultâneas que o cérebro humano consegue fazer, para "infinito" e tornar os resultados visíveis ao olho humano.

Zaki et al. (2007) considera que a MD é um processo de descoberta automática de novos e compreensíveis modelos e padrões em grandes quantidades de dados.

O resultado da etapa de MD é um arquivo que deve ser avaliado e interpretado, gerando assim, um novo conhecimento.

4.2.5 Avaliação e Interpretação dos Resultados

Os resultados obtidos através da etapa de MD podem ser apresentados em relatórios, gráficos e tabelas. O arquivo gerado, precisa então, ser analisado e interpretado.

Nesta fase, todos os participantes do processo de KDD avaliam de forma criteriosa os resultados com o objetivo de interpretar o modelo gerado na etapa de MD. É dessa criteriosa avaliação que o novo conhecimento é extraído e poderá posteriormente ser utilizado nas tomadas de decisões (Castanheira, 2008).

4.3 Conclusão

Com o estudo apresentado neste capítulo pode-se concluir que a descoberta do conhecimento em bases de dados é um processo complexo que implica iterações entre várias etapas, bem como a intervenção de um especialista do domínio durante partes do processo.

É importante ressaltar que caso não seja possível a participação de um especialista do domínio, as transformações dos dados que exigem a presença de tal especialista devem ficar intactas, evitando assim a distorção dos dados e a modificação do arquivo de resultado final gerado pelo algoritmo de MD.

Conclui-se também que para a execução de cada etapa do processo de KDD, é necessário que seja seguido um critério, considerando os dados disponíveis e o que se espera do processo de KDD.

O presente trabalho faz a descoberta do conhecimento em uma base de dados composta por consumo mensal de usuários de energia elétrica através da detecção de padrões de consumo com o objetivo de identificar padrões com anomalias que represente fraudes nas redes de distribuição de energia elétrica.

Levando em consideração que o processo de MD é o mais importante dentro do processo de KDD, o capítulo seguinte apresentará uma visão mais aprofundada desta etapa do processo de KDD.

Capítulo 5

Mineração de Dados

Neste capítulo é estudado um dos processos de KDD, a mineração de dados. São apresentadas as principais tarefas do processo de MD, citando aplicações e características de cada uma delas. Ainda neste capítulo, são descritas as principais técnicas de MD, por exemplo: Redes Neurais Artificiais, Árvores de Decisão, entre outras.

Para complementar, este capítulo ainda apresenta um resumo bibliográfico a respeito das técnicas de MD escolhidas para serem utilizadas neste trabalho, e por fim, apresenta uma conclusão do capítulo.

5.1 Mineração de Dados

Embora seja confundida com o processo completo de KDD, MD é considerada a mais importante de todas as etapas do processo de KDD (Cruz, 2008).

Na literatura é possível encontrar diversas definições para MD, mas pode-se resumir que a MD é um processo não trivial de identificar em bases de dados, novos padrões válidos e potencialmente úteis no processo de tomada de decisão das corporações (Fayyad et al., 1996).

De acordo com Nimer & Spandri (1998), MD é uma ferramenta que através de análises de grandes quantidades de dados armazenados em um *Data Warehouse*, utiliza técnicas de estatística, matemática e reconhecimento de padrões para descobrir novos padrões e tendências nos dados de uma corporação.

Possa et al. (1998) salienta que MD é um conjunto de técnicas que, para descobrir padrões e regularidades em grandes conjuntos de dados, envolvem métodos matemáticos e heurísticos.

Para Silva (2000), MD pode ser conceituada como uma técnica utilizada para determinar padrões em grandes bases de dados, auxiliando nas tomadas das decisões.

Nas palavras de Guizzo (2000), MD é a extração de informações potencialmente úteis e previamente desconhecidas de grandes bancos de dados, com o objetivo de descobrir perfis de

consumidores e outros comportamentos que não seriam identificados nem mesmo por especialistas, sem a utilização de MD.

É possível resumir que MD se caracteriza pela existência de algoritmos que executam uma tarefa proposta, com o objetivo de extrair conhecimento implícito e útil de uma base de dados (Castanheira, 2008).

Entretanto, é importante ressaltar que para definir o algoritmo à ser utilizado na etapa de MD, os objetivos da execução do processo de KDD devem estar bem definidos.

A mineração de dados apesar de ter grande semelhança com processos estatísticos, se diferencia deles porque ao invés de verificar padrões hipotéticos, a MD utilizada os próprios dados para descobrir tais padrões (Castanheira, 2008). Conforme Thearling et al. (1999), aproximadamente 5% de todas as relações podem ser encontradas através de métodos estatísticos, e utilizando MD, pode-se descobrir outras 95% de relações que anteriormente estavam desconhecidas.

Existem diversas ferramentas e tarefas de MD que podem ser utilizadas separadamente ou em combinação, a fim de obter os resultados esperados de um processo de KDD.

Nas palavras de Batista (2003), as etapas do processo de MD podem ser dívidas em:

- Escolha da tarefa de MD;
- Escolha da técnica de MD;
- Escolha do algoritmo a ser utilizado;
- Aplicação do processo de MD.

5.1.1 Principais Tarefas de Mineração de Dados

As tarefas de extrair o conhecimento vêm crescendo de forma gradiente sendo necessário que antes de iniciar o processo de MD, seja decidido qual tipo de conhecimento o algoritmo escolhido deve extrair da base de dados.

As tarefas de MD são descritivas ou preditivas. As descritivas têm o objetivo de encontrar padrões para descrever os dados, enquanto as preditivas usam algumas variáveis para prever valores desconhecidos ou futuros de outras variáveis (Ales, 2008).

Classificação

Com o objetivo de prever novos padrões, a tarefa de classificação é preditiva e baseia-se em uma função de aprendizado que encontra um número finito de classes através do mapeamento dos dados (Tan et al., 2005).

O principal objetivo da tarefa de classificação é encontrar alguma correlação entre os atributos de entrada e uma das classes criadas inicialmente, de modo que, o processo de classificação possa usá-la para predizer a classe de um exemplo novo e desconhecido (Castanheira, 2008).

De acordo com Batista (2003), o princípio da tarefa de classificação é descobrir algum tipo de relacionamento entre os atributos preditivos e o atributo objetivo, de modo à descobrir um conhecimento que possa ser utilizado para prever a classe desconhecida.

Pode-se citar como exemplos de tarefas de classificação:

- Separar pedidos de créditos em baixo, médio e alto risco a partir de históricos de consumo e débitos;
- Identificar o diagnóstico mais correto à uma dada doença, levando em consideração um conjunto de sintomas;
- Constatar quais produtos são mais vendidos em determinado período do ano observando o consumo de anos anteriores.

Regressão ou Estimativa

A tarefa de regressão tem o objetivo de prever uma função desconhecida cuja saída tem um domínio de valores reais (Cruz, 2008). Essa tarefa é conceitualmente similar à tarefa de classificação, mas tem como principal diferença que o atributo a ser descoberto é contínuo ao contrário da classificação, na qual o atributo é discreto.

A tarefa de regressão também é conhecida por predição funcional, predição de valor real, função de aproximação, ou ainda, aprendizado de classes contínuas (Uysal & Guvenir, 2004). Um exemplo de regressão é a previsão de um determinado valor de uma ação da bolsa de valores com base em indicativos financeiros.

Os métodos de regressão vêm sendo estudados há bastante tempo, no entanto, segundo Indurkhya & Weiss (n.d.), em aprendizado de máquinas e MD, a maioria das pesquisas é voltada para problemas de classificação, que são mais comumente encontrados na vida real do que os problemas de regressão (Castanheira, 2008).

Alguns autores descrevem a tarefa de regressão como um caso particular de classificação, já que seu objetivo é encontrar um modelo para predição de um atributo contínuo como função dos outros atributos (Ales, 2008).

A tarefa de regressão também é chamada como tarefa de estimativa. De acordo com Fayyad et al. (1996), a tarefa de estimativa é aprender uma função que mapeie um item dado para uma variável de predição real estimada.

Pode-se citar como exemplos de tarefa de estimativa:

- Estimar números de filhos em uma família;
- Estimar a probabilidade de chuvas em determinado mês baseado em anos anteriores;
- Estimar a faixa etária de vida de uma pessoa levando em consideração os demais membros da família;
- Entre outros.

Agrupamento ou Segmentação

A tarefa de agrupamento, também conhecida como segmentação é um processo de divisão de um grupo heterogêneo em vários subgrupos mais homogêneos. Neste processo não existem classes pré-definidas e os dados são agrupados de acordo com suas características próprias, sendo essa a principal diferença deste processo com a tarefa de classificação.

Tan et al. (2005) conceitua que o agrupamento procura grupos de padrões tal que padrões em um grupo são mais similares uns aos outros e dissimilares a padrões em outros grupos.

Na maioria das vezes, a tarefa de agrupamento é feita como um preprocessamento de dados para outra forma de MD. Um exemplo que pode ser citado da utilização de agrupamento como tarefa de preprocessamento é a utilização em uma aplicação de segmentação de mercado, que pode-se primeiro dividir os clientes em grupos de acordo seus comportamentos de compra (agrupamento), para posteriormente aplicar um processo de classificação (Castanheira, 2008).

Regras de Associação

A tarefa de regra de associação tem o objetivo de encontrar tendências que possam ser utilizadas para entender e explorar padrões de comportamento dos dados. Nas palavras de Tan et al. (2005), o objetivo das regras de associação é produzir regras de dependência que irão predizer a ocorrência de um atributo baseado na ocorrência de outros atributos.

De acordo com Agrawal & Srikant (2004), as regras de associação caracterizam o quanto a presença de um determinado conjunto de itens nos dados da base implica na presença de algum outro conjunto distinto de itens nos mesmos registros.

Um exemplo claro da utilização de regras de associação pode ser visto observando os dados de vendas de uma concessionária de veículos. Sabe-se que 70% dos clientes que compram o modelo de carro *"W"* também adquirem o seguro do tipo *"B"*. Nessa regra, 70% corresponde a confiabilidade da associação.

Um exemplo de utilização de regras de associação no dia a dia é para planejar a disposição de produtos nas prateleiras das lojas, de modo que itens geralmente adquiridos na mesma compra sejam vistos próximos entre si, chamando atenção dos clientes que ainda não se encaixam

a essa regra de associação (Castanheira, 2008).

Desvio

A tarefa de desvio tem o objetivo de descobrir um grupo de valores que não seguem um padrão definido, ou seja, descobrir um grupo de valores que tem padrões diferentes do que está definido. É importante ressaltar que, nesta tarefa, os padrões devem ser definidos com antecedência.

Pode-se usar esta tarefa para identificar fraudes baseadas em elementos que estão fora do padrão (Castanheira, 2008).

5.1.2 Principais Técnicas de Mineração de Dados

As técnicas de MD são executadas utilizando algoritmos criados para atingirem determinados objetivos, de acordo com a tarefa escolhida para ser realizada.

Uma única técnica pode ser implementada por inúmeros algoritmos diferentes, podendo também resultar, de acordo com cada algoritmo, resultados diversos. Existe também, a possibilidade de utilizar técnicas e algoritmos diferentes para executar uma determinada tarefa e se obter resultados similares.

É importante ressaltar que os resultados variam de acordo as técnicas e algoritmos utilizados, sendo assim, é importante que a escolha das técnicas e algoritmos sejam feitos de forma criteriosa, levando em consideração os resultados que se pretendem obter. Cortez (2002) complementa que as técnicas se distinguem pela forma de representação do conhecimento (modelo) e pelo algoritmo de procura dos parâmetros internos do modelo.

As principais técnicas de MD são (Cruz, 2008):

- Árvores de Decisão: são representações gráficas de classificação. Os algoritmos de árvores de decisão constroem árvores a partir de dados de treino, dividindo em subconjuntos até que esses representem uma só classe (nó folha);
- Redes Neurais Artificiais: são modelos simplificados do sistema nervoso central compostos por neurônios conectados entre si e tem capacidade de aprendizado, generalização, associação e abstração;
- Máquinas de Vetores de Suporte: são algoritmos de classificação que provê o estado da arte para um amplo domínio de aplicações, como o reconhecimento da escrita, detecção de fraudes, previsão de séries, entre outras;
- Entre outros.

5.2 Redes Neurais Artificiais

Assim como o cérebro humano, as Redes Neurais Artificiais (RNAs) apresentam uma estrutura altamente paralelizada, composta por neurônios conectados entre si.

De acordo com Schuch et al. (2010), uma RNA consiste de unidades computacionais interconectadas por canais unidirecionais chamados de conexões e essa estrutura tem a função de processar informações de forma paralela e distribuída.

As RNAs são inspiradas, precisamente, no funcionamento do cérebro humano, capaz de realizar tarefas cognitivas que antes eram somente realizadas pelo cérebro humano, como o aprendizado, a generalização, associação e abstração (Castanheira, 2008) (Cortez, 2002).

Haykin (2001) menciona que uma propriedade importante das RNAs é a habilidade de aprender a partir do ambiente na qual estão inseridas e melhorar seu desempenho através da aprendizagem. Os processos de aprendizagem das RNAs buscam adquirir experiência através de um processo de repetidas apresentações dos dados à rede (Castanheira, 2008).

Conforme Coppin (2013), a palavra artificial é utilizada na descrição de redes neurais para diferenciá-las das redes neurais biológicas do cérebro humano, e consistem em nós organizados em camadas.

A tecnologia de RNA procura imitar o processo de resolver problemas do cérebro humano que aplica o conhecimento adquirido ao longo das experiências da vida para resolver novos problemas e/ou situações (Almeida, 2009).

Segundo Ferreira (2008), “Rede Neural Artificial” era um termo raro de se encontrar na literatura cientifica há cerca de duas décadas atrás, mas hoje, representa uma vigorosa área de aplicação multidisciplinar. Para concluir o pensamento, Ferreira (2008) acrescenta que a RNA tem sido uma das ferramentas mais exploradas e que apresentam bons resultados nas mais diferentes áreas do conhecimento em resolução de problemas complexos.

Uma das definições mais utilizadas de RNA é dada por Costa (2010), que diz que as RNAs são generalizações de modelos matemáticos do sistema biológico nervoso do cérebro humano e uma das principais vantagens da utilização de RNA é a capacidade de construir um modelo do problema utilizando os dados a partir de medições experimentais do domínio do problema.

5.2.1 Evolução das Redes Neurais

Um dos primeiros modelos artificiais de neurônio biológico foi resultado do trabalho pioneiro de McCulloch & Pitts (1943). Tal trabalho concentrou-se em descrever um modelo artificial de um neurônio e apresentar suas capacidades computacionais e pouco foi falado sobre as técnicas de aprendizado.

A partir de 1943 têm sido desenvolvidos modelos muito detalhados e realistas de neurônios e também de sistemas maiores do cérebro humano de forma matemática (Russell & Norvig, 2013).

Hebb (1949) apresentou em sua pesquisa o primeiro trabalho a ter ligação direta com o aprendizado de RNA, o qual mostrou como a praticidade da aprendizagem de RNA é obtida através da variação dos pesos de entrada dos neurônios.

Em 1958 foi demostrado um novo modelo de RNAs, o *perceptron* (Rosenblatt, 1958), o qual defendia que se o modelo criado inicialmente fosse acrescido de sinapses ajustáveis, poderiam ser treinadas para classificar certos tipos de padrões (Castanheira, 2008).

Em Rumelhart & McClelland (1986), foi desenvolvido o primeiro algoritmo de retro propagação para treinamento de redes *Multi Layer Perceptron* (MLP), que são redes *perceptron* de múltiplas camadas.

Uma RNA com todas as entradas conectadas diretamente com as saídas é conhecida como rede *perceptron* ou rede neural de camada única (Russell & Norvig, 2013). O *perceptron* pode ter inúmeras entradas e pode ser utilizado para realizar processo de classificação de imagens ou tarefas de reconhecimento simples (Coppin, 2013).

Os testes de McCulloch & Pitts (1943) provam que uma única camada não resolveria todos os problemas que poderiam ser resolvidos utilizando RNAs. Diante disso, argumentaram que qualquer funcionalidade desejada poderia ser obtida através da ligação de grande número de unidades em rede de profundidades aleatórias (Russell & Norvig, 2013).

Perceptron multicamadas são capazes de modelar funções complexas incluindo funções que não são linearmente separáveis. A arquitetura típica de uma RNA multicamadas é composta por uma primeira camada que é a de entrada e cada neurônio recebe um sinal único de entrada. Na maioria das vezes, os neurônios da camada de entrada atuam somente na passagem dos sinais de entrada para os neurônios da camada seguinte, que são camadas ocultas (Coppin, 2013).

Uma rede multicamadas pode ter diversas camadas ocultas que contenham neurônios que trabalhem no aprendizado das redes. O sinal de entrada é repassado para todos os neurônios da camada oculta seguinte até que o sinal chegue na camada de saída. A camada de saída é a etapa de processamento e envia sinais de saída (Coppin, 2013).

5.2.2 Paradigmas de Aprendizagem

Existem dois conceitos de aprendizado das RNAs que são estudados, suas utilizações variam de acordo com o objetivo que se espera de uma RNA e quais dados se tem disponível. Estes conceitos são de aprendizado supervisionado e aprendizado não supervisionado.

Atualmente o paradigma de aprendizagem mais comum é o de aprendizagem supervisio-

nada. Para tal, são utilizados algoritmos de treino dos quais o mais conhecido é o *Backpropagation*.

As RNAs com aprendizado supervisionado aprendem ao serem apresentadas a dados de treinamento pré-classificados, ou seja, as entradas e saídas desejadas para a rede são fornecidas por um supervisor externo (Castanheira, 2008).

Neste processo, em muitas situações, as RNAs são capazes de generalizar, com grande grau de precisão, dados a partir de um conjunto de treinamento, pois tais redes aprendem ao modificarem os pesos das conexões de suas redes, para classificar mais precisamente os dados de treinamento (Coppin, 2013).

O método de aprendizado não supervisionado consiste nas redes aprenderem sem nenhuma intervenção humana (Coppin, 2013). Hebb (1949) propôs um método de aprendizado não supervisionado em RNAs conhecido como aprendizado de Hebb. Tal método tem o princípio de que, se dois neurônios, em uma RNA, estiverem conectados e forem ativados ao mesmo tempo, quando ocorrer uma entrada na rede, então deve-se reforçar a conexão entre esses dois neurônios (Coppin, 2013). Neste sentido, Edelman (1990) afirma que o que acontece no aprendizado de Hebb provavelmente é o mesmo que ocorre no cérebro humano quando há um novo aprendizado.

Algoritmo *Backpropagation*

O Algoritmo *Backpropagation* é baseado em ciclos de aprendizagem onde, os erros obtidos através das diferenças entre as saídas da RNA e os valores de treino são propagados para trás, desde os neurônios de saída, até a entrada, e posteriormente ocorre um reajuste de pesos das conexões, e o treino só é finalizado quando se obtém o mínimo de erro de saída, ou o mínimo de classificações erradas (Cruz, 2008).

Computacionalmente falando, o algoritmo *Backpropagation* é pesado e de execução lenta. Levando em consideração esses fatores, Fahlman (1998) desenvolveu o *QuickPropagation* e Riedmiller (1994) criou o algoritmo *Resilient Backpropagation*, que são melhorias do algoritmo *backpropagation* inicial.

5.3 Máquinas de Vetores de Suporte

Desenvolvidas inicialmente por Vapnik (1979), as máquinas de vetores de suporte (*Support Vector Machines* - MVS) são, atualmente, uma abordagem nova e popular para aprendizagem supervisionada. Inicialmente sua utilização foi introduzida nas engenharias com o objetivo de classificação de padrões e calibragens (Marreto, 2011)

As MVS têm a capacidade de resolver problemas de classificação e regressão, adquirindo na etapa de treinamento, através do aprendizado, a capacidade de generalização durante execução de novas entradas (Cruz, 2008).

A ideia principal de uma MVS é construir um hiperplano como superfície de decisão de uma maneira que a margem de separação entre exemplos positivos e negativos seja máxima (Marreto, 2011).

As definições de MVS na literatura podem ser resumidas em que MVS é um algoritmo de classificação que provê o estado da arte para um amplo domínio de aplicações, como reconhecimento de escrita, detecção de faces, previsão de series, entre outros.

Existem três propriedades que tornam a MVS atraente e muito utilizada nos últimos anos (Russell & Norvig, 2013):

- As MVSs constroem um separador máximo de margens, ou seja, um limite com a maior distância possível entre as classes distintas;
- As MVSs criam um hiperplano com uma separação linear, mas possuem também, a capacidade de incorporar os dados em um espaço de dimensão superior, usando uma metodologia conhecida como **Kernel**;
- As MVSs são métodos não paramétricos, ou seja, eles mantêm exemplos de treinamento e podem precisar armazenar todos eles.

De acordo com Dosciatti et al. (2013), a boa capacidade de generalização e robustez em grandes dimensões de dados são algumas das principais características das MVSs e que torna sua utilização mais interessante perante outras técnicas.

O nome “vetores de suporte” deriva do fato do algoritmo “suportar-se” em alguns dados para definir a distância entre as classes (Cruz, 2008). Sua implementação é feita utilizando um método de minimização estrutural de riscos. Este princípio de implementação é baseado no fato de que a taxa de erro de uma máquina de aprendizagem sobre dados de teste é limitada pela soma da taxa de erro de treinamento e por um termo que depende da dimensão de *Vapnik-Chervonenkis* (V-C); no caso dos padrões separáveis, a MVS produz um valor de zero para o primeiro termo e minimiza o segundo termo (Marreto, 2011).

Os algoritmos que implementam MVSs são constituídos por basicamente três etapas, que podem ser observadas na Figura 5.1 (Russell & Norvig, 2013):

- *mapping*: consistem em transformar o espaço dos vários atributos dos dados em um espaço multidimensional (hiperespaço), com o objetivo de permitir a separação linear dos dados;

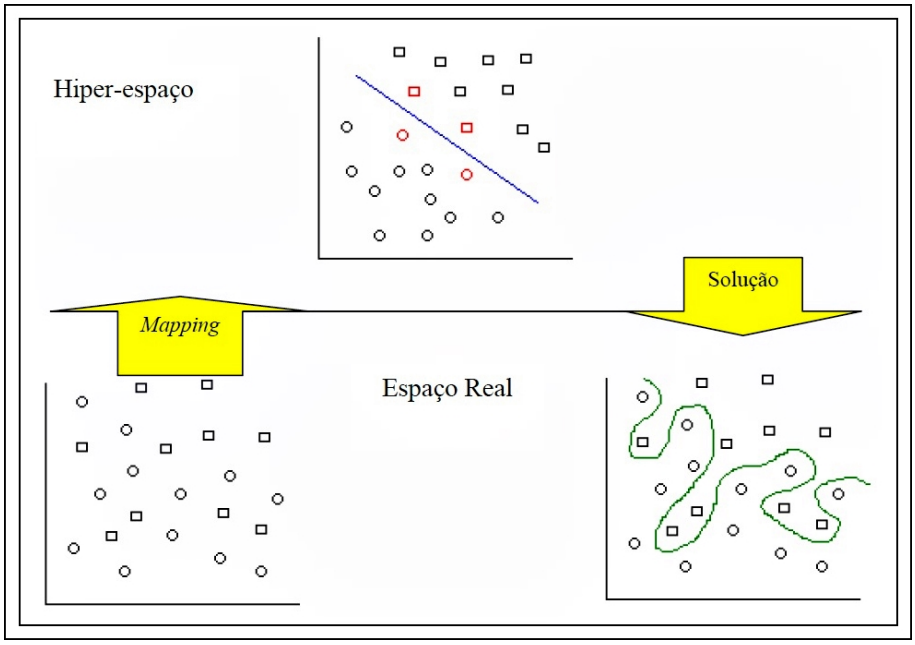

Figura 5.1: As etapas de um algoritmo de MVS.
Fonte: Cruz (2008)

- definição do hiperespaço: nesta etapa é feita a definição do hiperespaço com recurso de vetores. Estes por sua vez, são construídos a partir de alguns atributos dos dados após a etapa de *mapping*;
- apresentação dos resultados: esta etapa tem o objetivo de apresentar os resultados para avaliação e interpretações futuras.

5.3.1 Funções e Métodos de Kernel

O *Kernel* tem a finalidade de projetar vetores com características de entrada em um espaço de características de alta dimensão para a classificação de problemas que se encontram em espaços não linearmente separáveis (Mendoza, 2009).

Nas palavras de Cruz (2008), os métodos de *Kernel* dão aos dados uma dimensionalidade tal que sejam linearmente separáveis, produzindo assim, uma transformação nos dados.

O processo de transformação de um caso onde o problema não é linearmente separável, em um domínio linearmente separável através do aumento da dimensão pode ser observado na Figura 5.2. Nesse caso o mapeamento é feito por uma função de *Kernel* $F(x)$ (Rebelo, 2008).

A escolha do *Kernel* que fará o *mapping* é muito importante para que o resultado da técnica de MVS seja satisfatório.

Vapnik (1979) propõe um método para que o *Kernel* tenha a capacidade de transformar o espaço dos dados num espaço de dimensionalidade suficiente para que a separação seja linear,

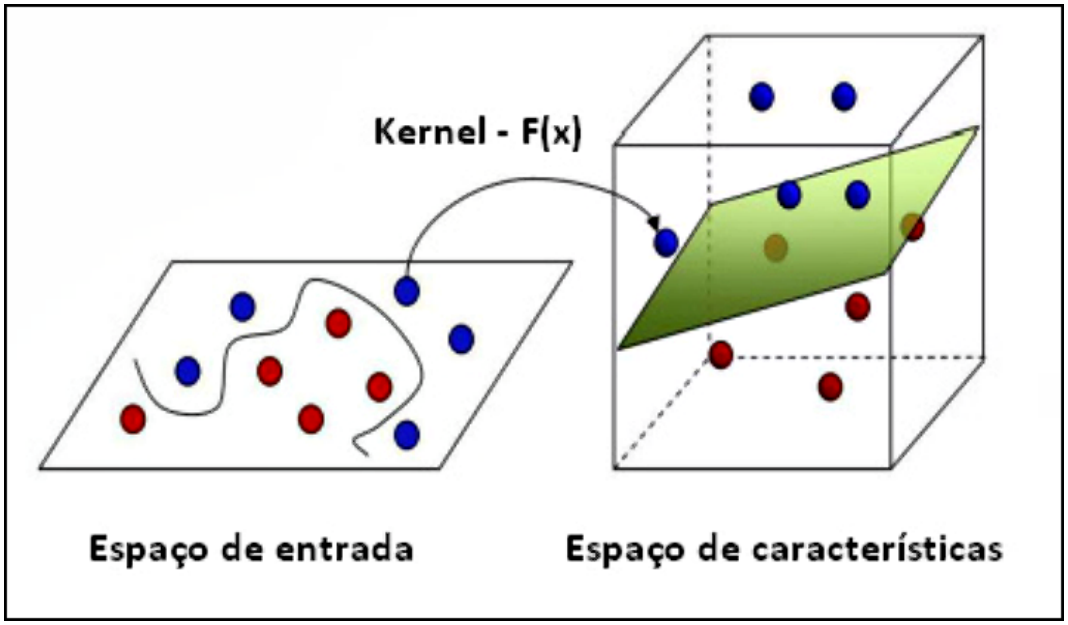

Figura 5.2: Transformação utilizando mapeamento por uma função *Kernel*.
Fonte: Rebelo (2008)

e que por outro lado, a capacidade de transformação não seja tal que leve ao sobre ajustamento. Este método é conhecido como "Método de Minimização do Risco" (Cruz, 2008).

É importante ressaltar que existem casos em que o *Kernel* não consegue transformar o espaço dos dados em um espaço de dimensionalidade suficiente para que haja uma separação linear, pois caso seja forçado uma separação linear, os dados podem sofrer alguns defeitos (Cruz, 2008).

5.4 Conclusão

O objetivo da execução da MD deve estar bem definido para escolher qual tarefa melhor atenderá as necessidades do usuário.

Como o objetivo deste trabalho é classificar se o consumidor de energia é fraudulento ou não em uma rede de distribuição de energia elétrica, a tarefa mais adequada dentro do processo de MD é a de classificação. As demais tarefas do processo de MD não se encaixaram na execução do presente trabalho pois o objetivo do mesmo é classificar os usuários e as demais tarefas tem objetivos diferentes que incluem previsões, agrupamentos, entre outros.

A técnica de MD escolhida para ser utilizada neste trabalho foi a de Redes Neurais Artificiais e Máquinas de Vetores de Suporte pois para a tarefa de classificação essas são as principais técnicas utilizadas. A Arvore de Decisão, mesmo sendo uma técnica de classificação, não foi utilizada porque neste trabalho busca-se utilizar ferramentas novas, e existem trabalhos correlatos que já utilizaram Arvores de Decisão. As demais técnicas são especificas para outros tipos

de tarefas, que não estão ligadas a natureza deste trabalho.

Ainda neste capítulo, pode-se concluir que uma RNA consiste em unidades computacionais que imitam o funcionamento do cérebro humano, capaz de realizar processos de aprendizagem, classificação, generalização, associação e abstração. Conclui-se também que a utilização de RNAs vem crescendo e sofrendo evoluções em seus algoritmos de aprendizado e treinamento levando em consideração os paradigmas de aprendizagem, que podem ser supervisionados, ou não supervisionados.

É importante ressaltar que não existe uma configuração exata de uma RNA para ser usada em cada problema a ser resolvido, tal configuração deve ser obtida através de experiências e pode variar de acordo com os dados que serão utilizados.

A respeito das Máquinas de Vetores de Suporte pode-se afirmar que apesar de ter tido início na década de 1970, ela só passou a ser conhecida nos últimos anos e é considerada uma abordagem nova e popular de aprendizagem supervisionada. Também é possível observar que as MVS constroem um hiperplano através dos espaços dos valores dos atributos e posteriormente permite a separação linear dos dados através desse hiperplano.

Capítulo 6

Metodologia

Neste capítulo apresenta-se informações sobre a base de dados utilizada no presente trabalho, com demostrações dos principais fraudes encontrados nas revisões manuais feitas pela empresa concessionária. É realizado uma breve descrição da ferramenta WEKA. Em sequência é descrito o processo de modelagem e de pré-processamento manual dos dados a fim de que se enquadrem aos algoritmos aplicados.

Ainda nesse capítulo, é apresentado o processo de extração do conhecimento, com apresentação da criação dos modelos utilizando algoritmos de RNA e MVS. Dando continuidade, são relatados os testes realizados com os modelos criados e os resultados obtidos na classificação utilizando tais modelos.

Para finalizar, é realizado uma análise dos resultados obtidos durante os testes.

6.1 Base de Dados

A base de dados utilizada no presente trabalho foi obtida através de uma concessionária Colombiana de energia elétrica, que por motivos de sigilo de informações, não pode ser citada.

A base de dados em seu estado inicial possui extensão *xls*, uma extensão comum para tabelas de dados.

A base de dados é composta por dados reais em kWh, do consumo de energia elétrica de 4.996.333 usuários, dos anos de 2011, 2012 e 2013. E além dessas informações, a base de dados é complementada com informações do tipo de consumidor, código de usuário, fases de conexão, tipo de alimentador e de transformador utilizados, entre outras informações irrelevantes para o presente trabalho. A Figura 6.1 mostra uma parte da base de dados utilizada no trabalho, lembrando que os códigos dos usuários foram borrados por motivo de sigilo de informações. Para complementar as informações, a empresa concessionária também forneceu uma base de dados com as informações de revisões manuais realizadas em todos os clientes cadastrados e as anomalias encontradas durante tais revisões. Essas informações são de grande importância pois

podem ser utilizadas para validar o processo de detecção de padrões de anomalias nas redes de distribuição de energia elétrica utilizando MD.

					CONSUMOS kWh																							
USUARIO	ALIMENTADOR	TRANSFORMADOR	CONEXIÓN	CLIENTE	AÑO 2011												AÑO 2012											
U_CODIGO	A_CODIGO	T_CODIGO	FASES	TIPO	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez
[illegible]	AGU23L12	N40003	ABN	RS3	254	266	229	191	222	224	264	299	241	241	232	235	241	247	212	200	205	215	242	240	267	254	268	240
[illegible]	AGU23L12	N40003	BN	RS3	46	51	54	55	51	63	65	58	66	67	64	60	66	72	86	64	39	63	55	54	56	71	65	68
[illegible]	AGU23L12	N40003	AN	RS2	82	93	93	73	88	73	67	72	85	66	70	59	90	93	81	74	57	67	69	66	77	65	33	55
[illegible]	AGU23L12	N40003	AB	RS3	56	13	91	73	88	76	84	83	99	92	91	100	88	88	100	95	102	86	87	87	99	92	117	108
[illegible]	AGU23L12	N40003	AB	RS3	48	42	126	135	131	99	184	202	218	158	176	177	192	180	142	126	182	52	79	96	89	90	24	87
[illegible]	AGU23L12	N40003	AB	RS3	42	40	93	105	133	126	118	103	94	98	76	104	119	92	72	86	64	42	45	38	98	112	106	108
[illegible]	AGU23L12	N40003	AB	RS3	111	78	70	66	63	67	77	65	66	64	64	72	86	69	65	62	62	56	63	65	65	70	63	74
[illegible]	AGU23L12	N40003	BN	RS3	126	62	98	83	0	8	25	19	0	0	35	43	20	31	41	72	64	66	51	15	18	12	8	13
[illegible]	AGU23L12	N40003	AN	CR6	0	70	80	36	18	11	0	5	12	46	32	25	26	27	25	31	34	34	28	36	34	32	23	22
[illegible]	AGU23L12	N40003	AN	RS3	29	37	31	41	36	52	15	27	82	92	26	37	61	67	38	29	57	35	29	45	31	27	30	43
[illegible]	AGU23L12	N40003	AN	RS3	174	173	163	138	149	148	155	144	157	162	165	154	141	136	150	143	140	136	155	163	173	178	152	157
[illegible]	AGU23L12	N40003	AN	RS2	219	203	221	191	131	272	227	200	218	207	214	220	228	200	220	194	213	205	187	205	197	184	198	212
[illegible]	AGU23L12	N40003	ABN	RS2	0	133	150	111	123	131	149	116	140	136	141	156	143	142	153	162	173	155	118	48	44	62	71	73
[illegible]	AGU23L12	N40003	ABN	CR6	2	37	65	49	54	56	57	53	55	52	59	31	79	3	1	15	16	19	22	20	17	19	18	19
[illegible]	AGU23L12	N40003	ABN	RS3	200	209	178	174	197	176	211	192	183	162	152	155	179	168	190	157	180	177	196	196	209	219	184	190
[illegible]	AGU23L12	N40003	BN	RS3	75	58	78	61	24	15	23	20	22	18	17	19	22	19	19	50	77	98	65	34	20	21	33	50
[illegible]	AGU23L12	N40003	AN	RS3	127	108	109	97	87	73	83	94	109	95	95	104	112	106	110	104	126	118	120	119	110	120	122	111
[illegible]	AGU23L12	N40003	BN	RS3	21	3	11	1	1	1	11	70	67	23	32	36	86	88	76	69	66	66	49	1	4	4	6	12
[illegible]	AGU23L12	N40003	ABN	CR6	0	45	53	44	44	44	50	47	48	49	54	52	56	52	57	61	57	57	73	61	55	63	58	43
[illegible]	AGU23L12	N40003	AN	RS3	65	41	42	45	51	56	61	54	55	49	50	93	107	100	100	98	106	94	95	102	98	108	101	106
[illegible]	AGU23L12	N40003	ABN	CR6	0	52	41	30	39	41	47	62	49	44	43	44	61	49	50	65	56	40	51	38	45	34	44	46
[illegible]	AGU23L12	N40003	BN	RS2	15	63	53	46	53	53	65	87	77	53	53	60	62	48	27	29	39	40	47	40	55	52	40	66
[illegible]	AGU23L12	N40003	ABN	CR6	0	54	72	65	68	68	53	57	48	32	33	46	49	31	30	23	33	34	26	23	28	29	47	34
[illegible]	AGU23L12	N40003	AN	RS1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[illegible]	AGU23L12	N40003	BN	RS2	6	7	7	6	3	2	2	2	4	2	5	5	10	4	3	2	4	6	3	3	3	3	3	4
[illegible]	AGU23L12	N40003	ABN	CR6	0	63	66	57	69	52	55	52	70	69	68	71	84	77	72	71	79	76	69	77	87	72	66	68

Figura 6.1: Base de dados recebida de uma empresa concessionária de energia elétrica Colombiana

6.1.1 Principais Padrões de Consumo com Anomalias Repassados pela Concessionária

Durante a vistoria manual, a empresa concessionária de energia elétrica que forneceu a base de dados, encontrou diversos padrões de consumo com anomalias, sendo que as principais causas dessas anomalias são fios emendados de forma clandestina, medidor furado, ligação anterior à caixa de medição e caixa de medição sem selo de segurança. A Figura 6.2 mostra algumas imagens ilustrativas de alguns fraudes apresentados pela concessionária de energia elétrica. Vale lembrar que diversas outras anomalias foram encontradas durante as verificações, mas com uma quantidade menor do que os apresentados neste trabalho.

A Figura 6.3 mostra graficamente outro tipo de padrão de fraude comum encontrado pelas concessionárias. É o caso dos fraudes nos medidores. Neste caso, a energia é medida em menor quantidade do que é inserida na rede, ou seja, através de furos no disco do medidor, ou/e através da criação de dentes no disco, o medidor passa a errar na medida, sempre marcando menos do que é realmente passado pelo medidor. Nota-se que, analisando o ano de 2011 e 2012, a curva de consumo mantém o mesmo padrão, apenas com uma redução de consumo a partir do mês de junho do ano de 2012.

Figura 6.2: (A)Ligação clandestina, (B) Medidor furado, (C) Ligação anterior a caixa de medição e (D) Caixa de medição sem selo

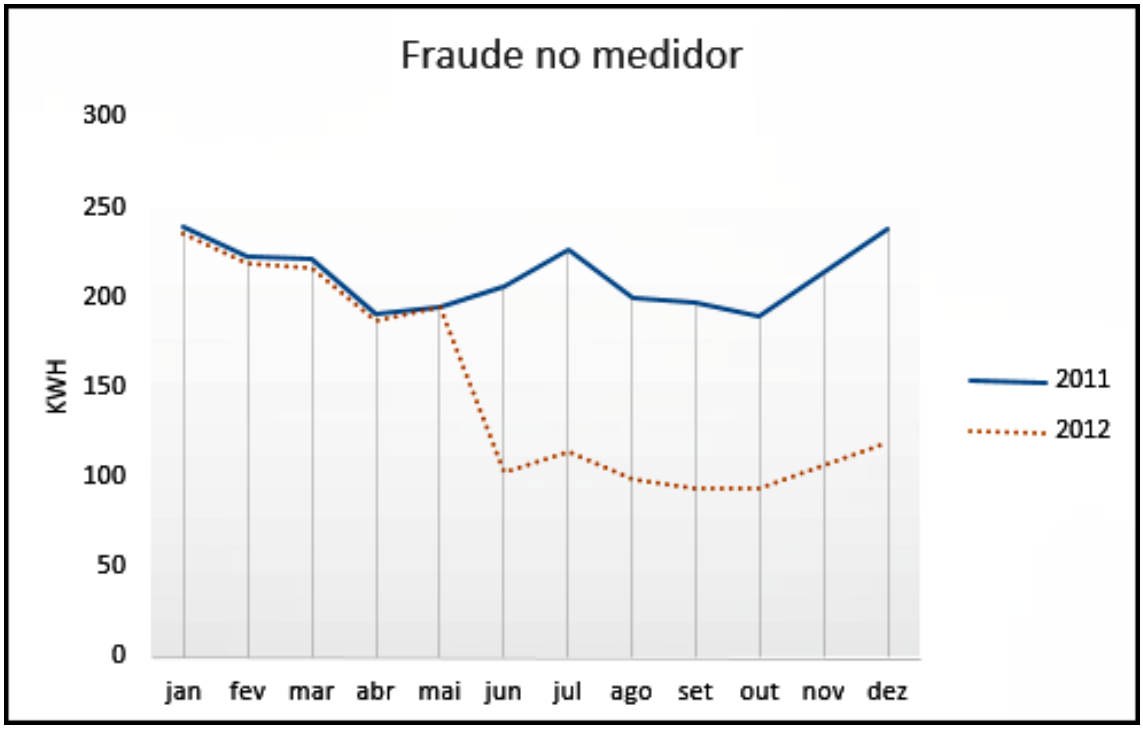

Figura 6.3: Padrão de consumo com medidor fraudado

As curvas de consumo podem variar mesmo quando a anomalia encontrada é a mesma em ambos consumidores. A Figura 6.4 mostra dois tipos de padrões de consumo que representam a mesma fraude, a ligação clandestina. Nota-se que no gráfico A, o consumidor fez uma ligação clandestina de uma maneira em que seu consumo de energia era feito através dos fios emendados de forma irregular, sendo assim, seu consumo passou a ser o mínimo permitido pela concessionária durante todos os meses. No gráfico B, a ligação clandestina foi feita de tal maneira que, determinados consumos são ligados na rede que passa pelo medidor de energia, e os de maiores consumos são ligados nos fios emendados direto na rede de distribuição, sem passar pelo medidor de energia elétrica.

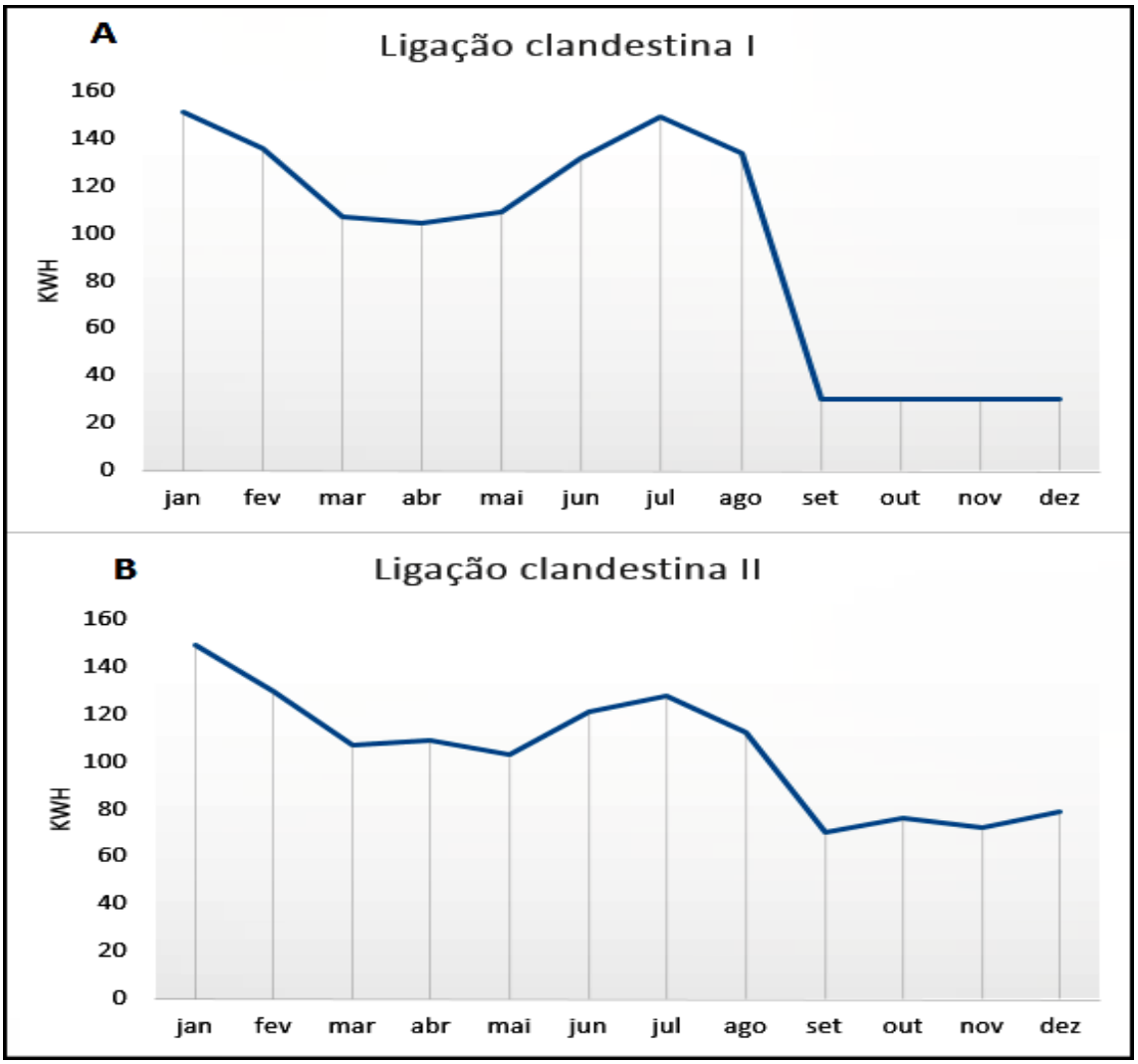

Figura 6.4: Padrões de consumo com ligações clandestinas

Pode-se concluir que, para o mesmo tipo de anomalia existem diversos padrões de consumo, e que quanto mais casos houver de determinada fraude, maior o número de padrões de consumos diferentes encontrados com tais fraudes. Casos de padrões similares para tipos de fraudes diferentes também são comuns, por isso, o processo de MD indica padrões de consumo que representem possíveis fraudes, mas é necessário sempre que, após a detecção da anomalia

no consumo do cliente, a concessionária vá até o determinado usuário para verificar manualmente.

6.2 Ferramenta WEKA

Waikato Environment for Knowledge Analysis é uma ferramenta de KDD que possui uma série de algoritmos de pré-processamento, MD e validação de resultados (Waikato, 2015).

Criada pela Universidade de Waikato na Nova Zelândia, Weka foi desenvolvido na linguagem de programação Java, possui distribuição livre e permite a implementação e inserção de novos algoritmos para solucionar problemas de classificação, regressão, associação e segmentação (Cruz, 2008).

A ferramenta Weka da preferência ao reconhecimento de ficheiros em um formato próprio do Weka, o *Attribute Relation File Format* (*arff*), mas também reconhece arquivos com extensão *csv* (Alcantara, 2012).

Como pode ser visualizado na Figura 6.5, a interface gráfica da ferramenta Weka possui diversas opções de trabalho, sendo que no presente trabalho será utilizado somente a opção *Explorer*.

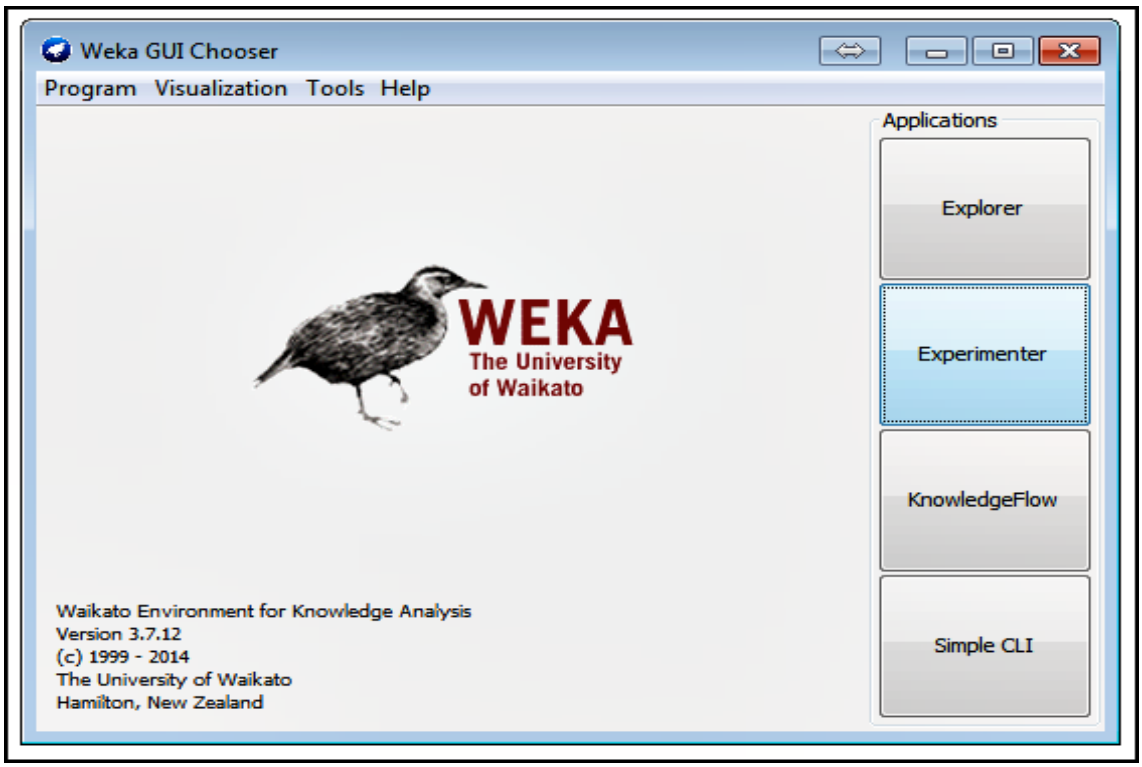

Figura 6.5: Interface da ferramenta WEKA

6.3 Modelagem e Pré-processamento dos Dados

Sendo que os dados recebidos não estavam em um modelo compatível com a ferramenta WEKA, primeiramente foi realizado o modelamento dos dados para arquivos com extensão *csv*.

Após essa transformação, foi necessária uma limpeza nos dados com o objetivo de eliminar os registros com dados incompletos e retirar da base de dados as informações irrelevantes para a verificação de anomalias na curva de consumo dos usuários.

Na Figura 6.6 é possível observar que após o processo de pré-processamento, houve uma redução dos dados, diminuindo de 462.433 para 314.023.

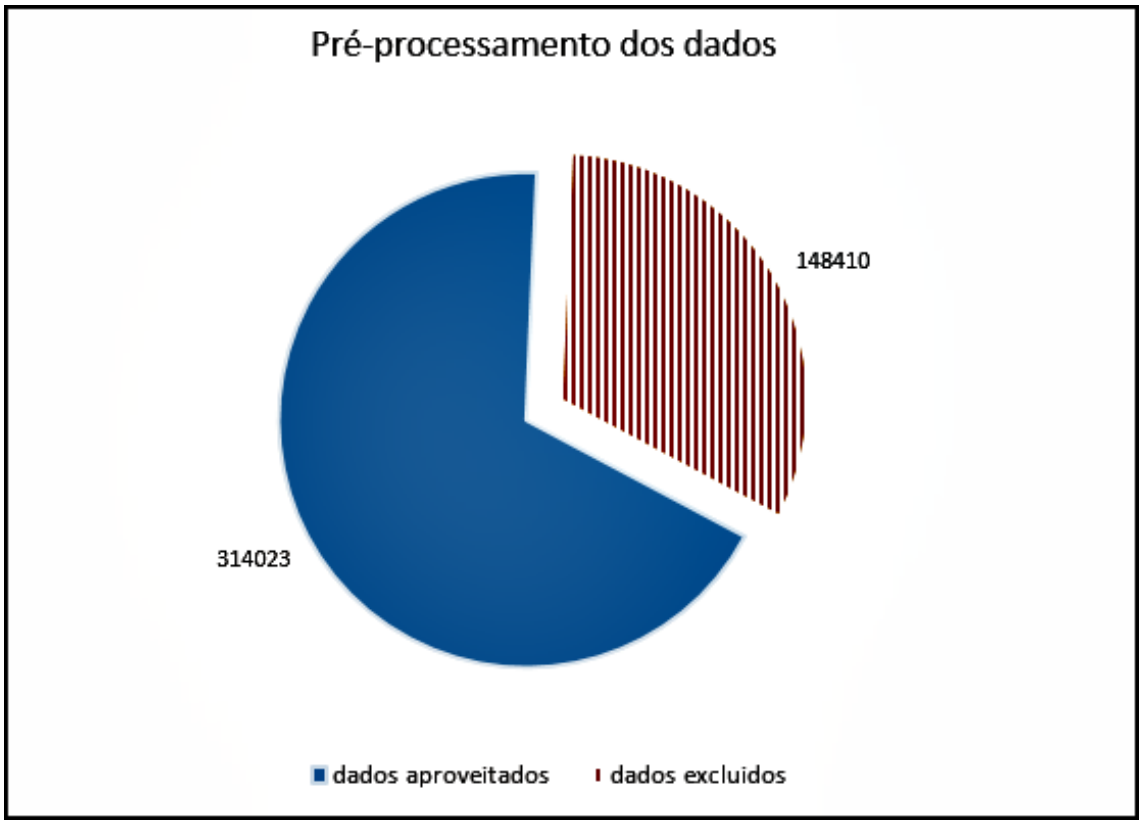

Figura 6.6: Dados pré-processados

É importante ressaltar que, como foi feito um pré processamento dos dados onde foram retirados registros incompletos, os modelos criados com RNA e MVS devem ser utilizados em bases de dados com dados pré processados, ou seja, sem dados incompletos. Caso esse pré-processamento não tivesse sido realizado, os modelos poderiam ser testados também, com dados sem pré-processamento, mas levando em consideração o tamanho da base de dados utilizada, a limpeza dos dados auxiliou na velocidade de criação dos modelos e execução dos mesmos após serem criados.

Ainda na etapa do pré-processamento dos dados, foi inserido na tabela de dados uma coluna responsável para acrescentar a conclusão das revisões manuais realizadas pela concessi-

onária.

6.4 Extração do Conhecimento e Criação dos Modelos com RNA e MVS

Com o objetivo de extrair o conhecimento da base de dados e encontrar padrões de consumos com anomalias nas curvas de consumo dos usuários de energia elétrica e levando em consideração que ambos os algoritmos criam modelos que são utilizados para classificar novos dados, foi realizado o processo de MD utilizando RNA e MVS, utilizando os dados de 2010 para criar modelos, e os demais dados foram armazenados para testes de verificação.

Os padrões de consumo de usuários com anomalias, gerados a partir do processo de MD, podem ser utilizados de maneira mais inteligível comparado a base de dados antes de todo o processo da extração do conhecimento. E tais dados, transformados em conhecimento, podem ser utilizados nas tomadas de decisões das empresas de forma sucinta.

6.4.1 Modelos Criados com a Mineração de Dados

A criação dos modelos é realizada com a inserção da base de dados, contendo uma coluna com os resultados se os usuários são ou não fraudulentos na ferramenta WEKA e aplicado os algoritmos selecionados para o presente trabalho. Cada algoritmo cria um modelo para que, então, novos dados possam ser inseridos para serem classificados.

Modelo MVS

As simulações realizadas utilizando o algoritmo de MVS permitiram alterações de diversos parâmetros, como os valores do coeficiente, degrau, entre outros, mas nenhuma alteração nas configurações do algoritmo alterou a porcentagem de acerto (acurácia) na classificação dos usuários com a base de dados, ou seja, nenhuma das alterações mudaram a quantidade de padrões identificados durante a criação do modelo.

A acurácia na criação do modelo com a aplicação desse algoritmo foi de 76,6740%, não encontrando novos padrões em nenhuma alteração nas configurações. Na Figura 6.7 é possível observar o modelo criado utilizando algoritmo de MVS e suas acurácias e erros na classificação.

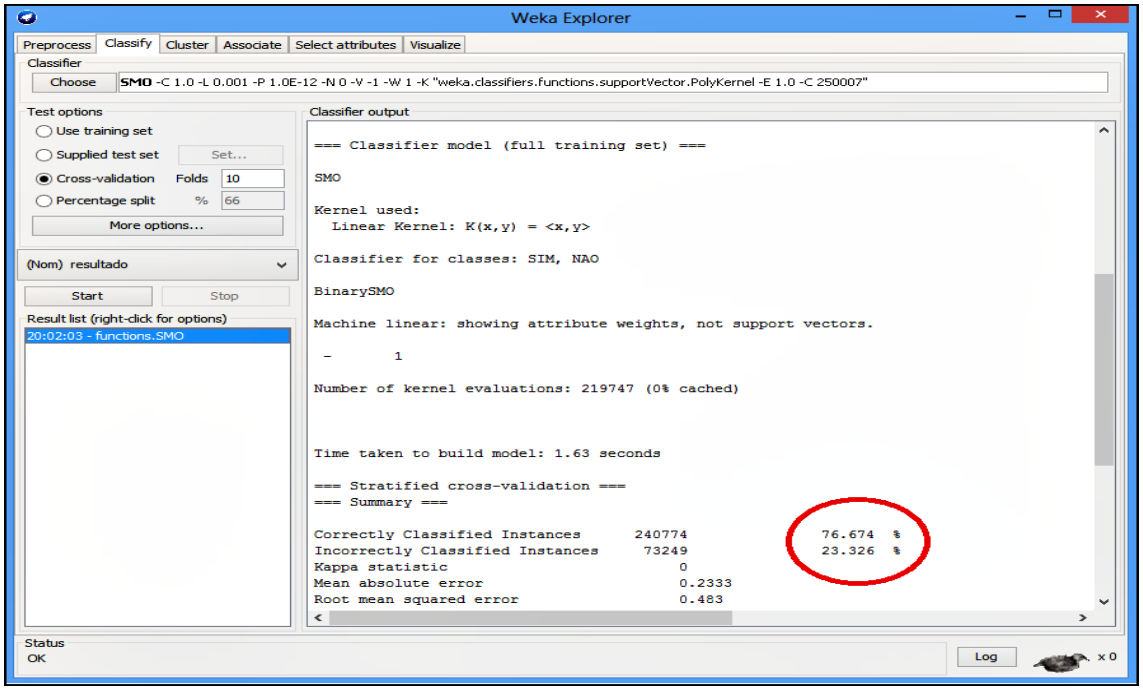

Figura 6.7: Modelo MVS com 76,6740% de acerto

Modelo RNA

Aplicando o algoritmo de RNA e alterando as configurações do algoritmo em busca de um melhor resultado no treinamento, foi possível observar que a alteração nas configurações que modificou a acurácia foi somente o *TrainingTime*. Essa configuração representa o número de épocas para treinar completamente a rede, e a acurácia de maior valor alcançou 92,7324%, como pode ser observado na Tabela 6.1:

Tabela 6.1: Alterações nas configurações do algoritmo de RNA

Teste	*TrainingTime*	Acerto
01	2.000	80,5422%
02	5.000	84,3011%
03	10.000	86,4733%
04	50.000	91,0249%
05	100.000	92,7324%
06	200.000	92,6829%

É importante ressaltar que na Tabela 6.1 só foram apresentados os dados, os quais, após as alterações na configuração, obtiveram alterações notáveis nos resultados durante o treinamento da rede.

Vale ressaltar que as modificações de aumento no número de épocas, quando chega a um determinado valor, começa a ter queda no resultado, como pode ser observado no teste de

número 06 apresentado na Tabela 6.1.

6.4.2 Testes com os Modelos Criados

Após a criação dos modelos utilizando os dados do ano de 2010, os quais já possuíam classificação de fraudulentos ou não, os modelos de melhores acurácia na classificação são, então, testados com os dados dos anos de 2011 e 2012, sem a classificação, para que os modelos criados classifiquem os novos dados inseridos.

MVS

O modelo MVS que alcançou 76,6740% durante a criação do modelo obteve um resultado com uma pequena variação entre a acurácia nos anos de 2011 e 2012. A Tabela 6.2 mostra a acurácia durante o treinamento com os dados de 2010 e os testes de classificação com os dados de 2011 e 2012.

Tabela 6.2: Testes com o modelo MVS

	acerto	erro
Criação do Modelo	76,674%	23,326%
Classificação de 2011	76,739%	23,261%
Classificação de 2012	81,239	17,761%

O tempo de execução do modelo com os 314023 usuários de 2011 foi de 2 segundos e para o ano de 2012, com a quantidade de 313972 usuários, também foi de 2 segundos para classificar se os usuários possuem ou não anomalias no consumo. Ainda, utilizando a ferramenta WEKA, é possível verificar a segurança de acerto na classificação, que foi de 81,9345% em ambos os testes.

RNA

O modelo RNA que alcançou 92,7324% de acerto durante a criação do modelo classificou os 314023 usuários de 2011 e os 313972 usuários de 2012 com um tempo de execução para classificação de 4 segundos, e a acurácia na classificação foi de 92,7324% em ambas as classificações, ou seja, dos usuários inseridos, 92,7324 % foram classificados entre usuários que possuem anomalias ou não no consumo de energia elétrica. A Tabela 6.3 mostra a acurácia na classificação durante o modelo e os dois anos classificados. Vale lembrar que a porcentagem na segurança de acerto se manteve em ambos os anos com 98,2909%.

Tabela 6.3: Testes com o modelo RNA

	acerto	erro
Criação do Modelo	92,7324%	7,2676%
Classificação de 2011	92,7324%	7,2676%
Classificação de 2012	92,7324%	7,2676%

6.5 Análises dos Resultados

Com as simulações realizadas durante o desenvolvimento do presente trabalho, tendo em mãos uma base de dados de consumo mensal de consumidores de energia elétrica é possível comparar e analisar os resultados obtidos durante o treinamento e também durante os testes com os modelos criados, e afirmar qual dos algoritmos teve um melhor resultado e poderá contribuir de forma mais eficiente nas tomadas de decisões de empresas que possuem base de dados semelhantes.

Como os resultados obtidos durante o treinamento para a criação dos modelos usando o algoritmo de MVS não levaram a alteração na acurácia, não foi possível fazer uma análise do motivo da alteração dos resultados a cada alteração.

No caso da aplicação do algoritmo de RNA, nota-se que o aumento do número de épocas para treinamento obteve o aumento da acurácia até um certo tempo, quando o valor do *Training-Time* foi aumentado para 200.000, a acurácia caiu. Isso acontece devido o super treinamento dos dados, que pode causar uma má classificação dos dados.

Mesmo nas configurações iniciais de ambos os algoritmos, os resultados do treinamento para criação do modelo apresentaram uma acurácia positiva, com resultados entre 75% e 85% de acerto na classificação. Essa positiva acurácia mostra que, mesmo em configurações padrões dos algoritmos, para a base de dados utilizada e bases semelhantes, a MD pode ser utilizada com um grau de confiança notável na colaboração das tomadas de decisões.

Na Figura 6.8 é possível observar a melhor acurácia de ambos os algoritmos durante o treinamento para criação dos modelos. É possível observar que, para esse tipo de dados, o algoritmo de RNA foi notavelmente melhor no processo de treinamento para classificação dos dados.

É importante ressaltar que o processo de treinamento é, em ambos os algoritmos, lento. Alguns treinamentos demoram cerca de 15 dias para se obter um modelo com boa quantidade de acerto. Mas o tempo de treinamento, como é realizado apenas para a criação do modelo, não prejudica na utilização de MD para classificações, pois depois do modelo pronto, a execução dos modelos para classificar novos dados, variaram de 3 a 10 segundos com os dados e modelos utilizados no presente trabalho.

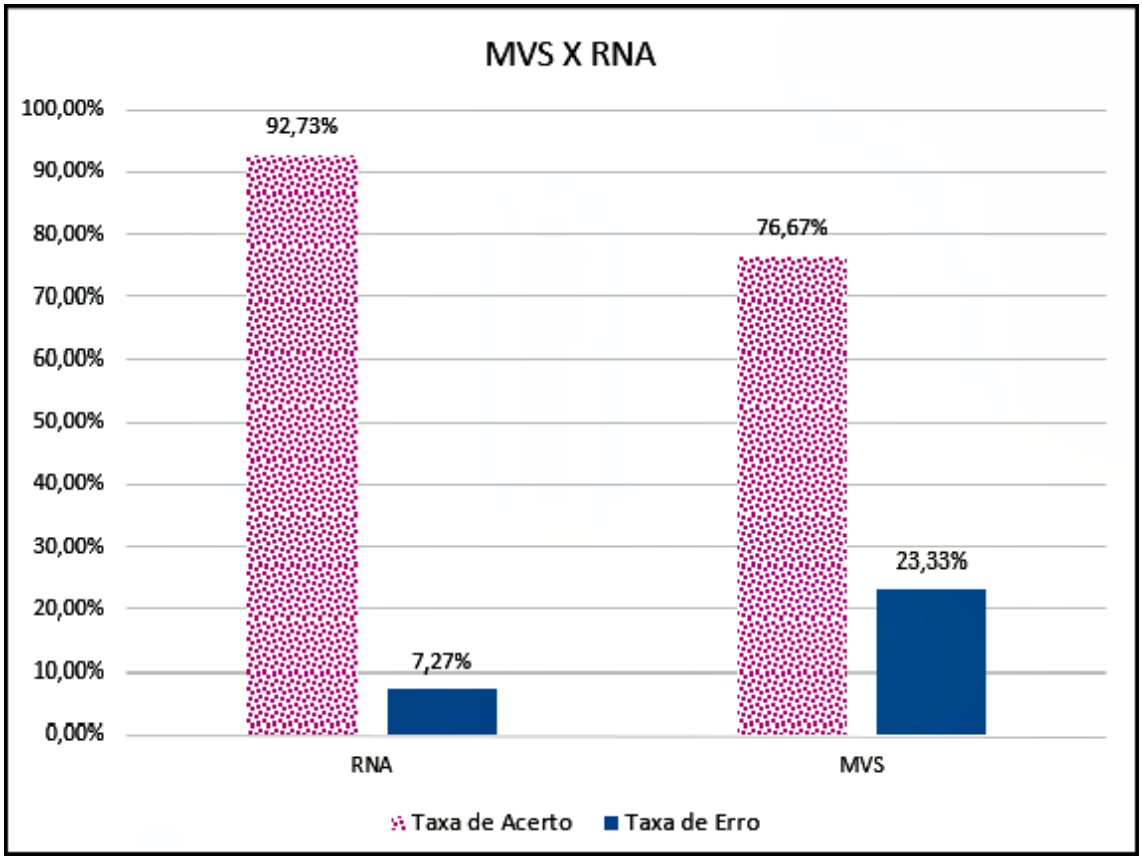

Figura 6.8: Comparação dos resultados obtidos no treinamento para criação dos modelos

Na execução dos modelos para classificar novos dados, utilizando a base de dados dos anos de 2011 e 2012, o algoritmo de RNA também obteve resultado mais positivo do que o algoritmo de MVS, como pode ser observado na Figura 6.9, onde é realizado a comparação entre os dois algoritmos.

Neste contexto, e analisando os gráficos, é possível afirmar que, apesar de ambos os algoritmos apresentarem um resultado positivo na classificação dos dados, o algoritmo RNA alcançou melhor acurácia durante o processo de classificação. Além disso, durante a utilização do modelo de RNA foi possível detectar anomalias no consumo de usuários que na tabela de dados oferecida pela concessionária, constava como usuários sem anomalias.

A Figura 6.10 apresenta parte da tabela de classificação fornecida pela concessionária, e a mesma parte da classificação realizada utilizando o modelo de RNA. Nota-se que em diversos casos, a classificação feita pela concessionária não detectou fraude e o processo de MD os classificou como usuários com possíveis fraudes.

Em alguns casos, onde a empresa concessionária classificou como não fraudulento e o processo de MD notou uma anomalia no padrão de consumo, é visível na curva de consumo o comportamento com possível anomalia. Analisando o gráfico da Figura 6.11, o qual a curva de consumo do usuário na base de dados oferecida pela concessionária não constava como usuário com anomalia, e no processo de RNA detectou-se possível anomalia, pode-se observar que

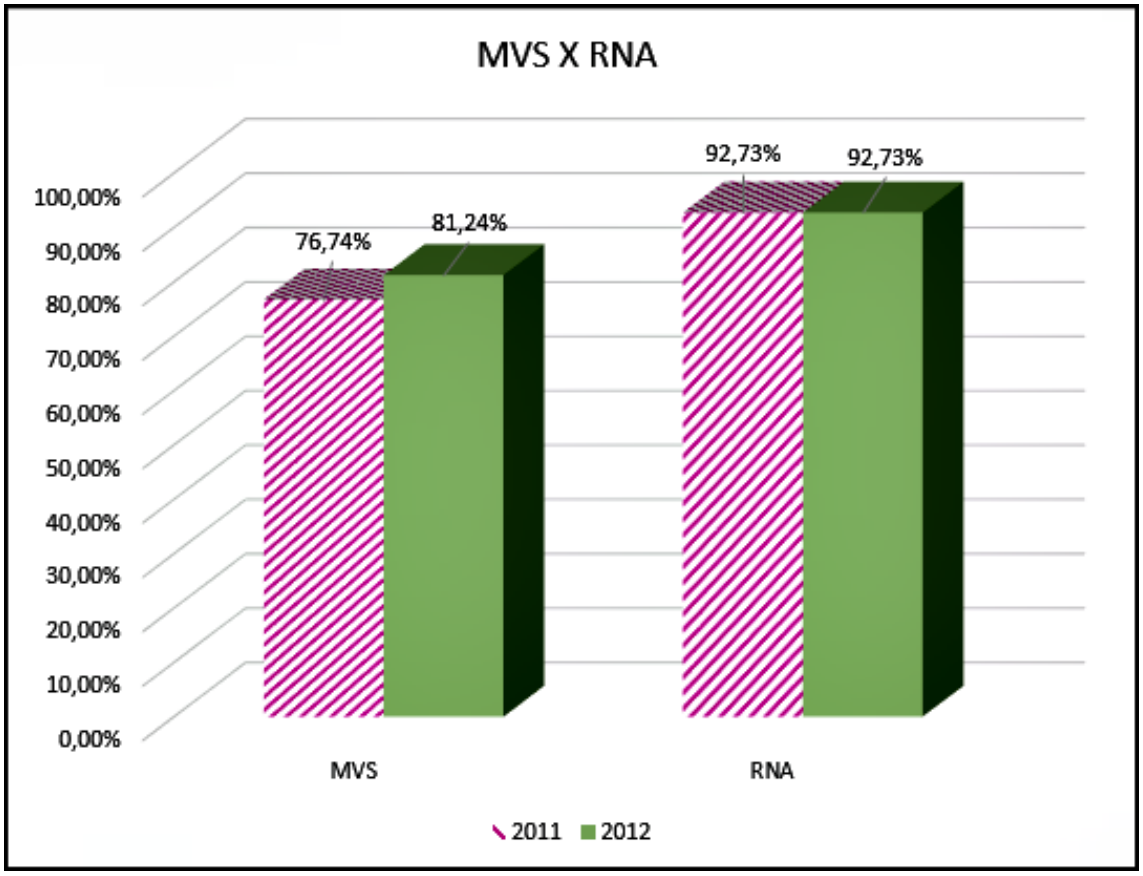

Figura 6.9: Comparação dos resultados obtidos nas classificações utilizando os modelos criados

Classificação repassada pela concessionária			Classificação obtida através de MD		
101719359	09/02/2011	selo quebrado	101719359	SIM	
236108766	23/03/2011	sem anomalias	236108766	SIM	⇦
101720195	11/03/2011	sem anomalias	101720195	NÃO	
101779793	28/04/2011	blocos de terminais queimados	101779793	SIM	
239337997	28/03/2011	sem anomalias	239337997	SIM	⇦
101792062	29/01/2011	sem anomalias	101792062	NÃO	
101793849	29/01/2011	sem anomalias	101793849	NÃO	
101796190	28/01/2011	selo quebrado	101796190	NÃO	
347674401	28/01/2011	sem anomalias	347674401	SIM	⇦
298088661	17/02/2011	ligação direta	298088661	SIM	
168096606	29/03/2011	sem anomalias	168096606	NÃO	
168102284	01/02/2011	sem anomalias	168102284	SIM	⇦

Figura 6.10: Comparação de usuários com anomalias na tabela oferecida pela concessionária e a tabela criada com o modelo de RNA

o consumo reduziu cerca de 50% se comparado os quatro primeiros meses do ano com os seis meses seguinte. Este é um possível caso de freio no medidor, onde o usuário, através de técnicas de fraudes, freia o medidor para marcar menor quantidade que a energia realmente consumida.

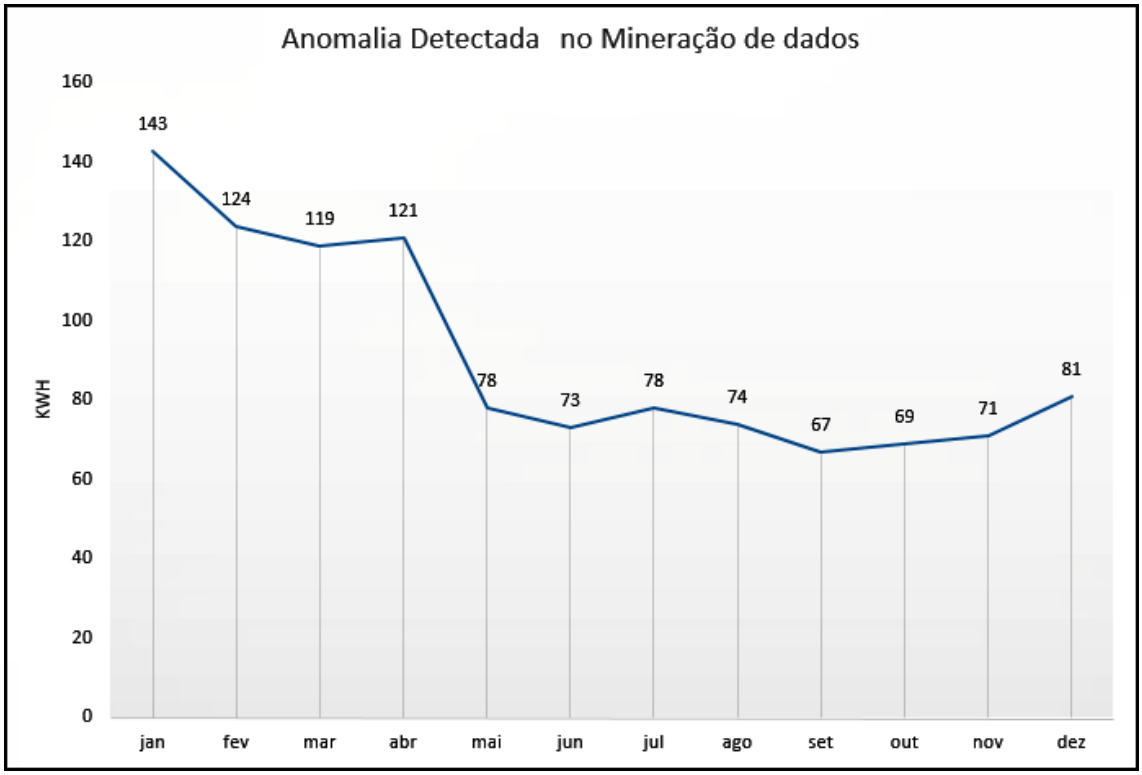

Figura 6.11: Usuário com anomalias a partir do mês de Maio

O gráfico da Figura 6.12 apresenta um segundo caso em que a classificação feita pela concessionária não detectou anomalia, mas o processo de MD detectou na curva de consumo, um padrão que indica ser fraude. Neste caso, a concessionária não detectou fraude pois a energia consumida estava sendo cobrada normalmente, e a fraude possivelmente está localizada após o medidor e feito por um usuário vizinho.

Após verificar esse tipo de fraude, que pode ser detectado pelo próprio consumidor através da observação da sua fatura mensal de energia elétrica, o usuário prejudicado com o aumento do consumo pode solicitar uma verificação realizada por técnicos da própria concessionária, ou por técnicos particulares.

Ligações diretas feitas nas redes de distribuição são detectadas com facilidade quando todo o consumo do usuário fraudador é feito através dessa ligação. No caso de usuários que fazem ligação direta somente para utilizar determinados aparelhos elétricos, e parte do consumo continua sendo através da energia passada pelo medidor são mais difíceis de serem percebidos pelas concessionárias. Na Figura 6.13 pode-se observar um possível caso de ligação direta na rede de distribuição, o qual não foi detectado pelo processo de classificação da concessionária, e no processo de MD, foi classificado como usuário com possíveis anomalias. É possível observar que, o consumo de energia foi reduzindo mensalmente até chegar no mínimo permitido pela concessionária.

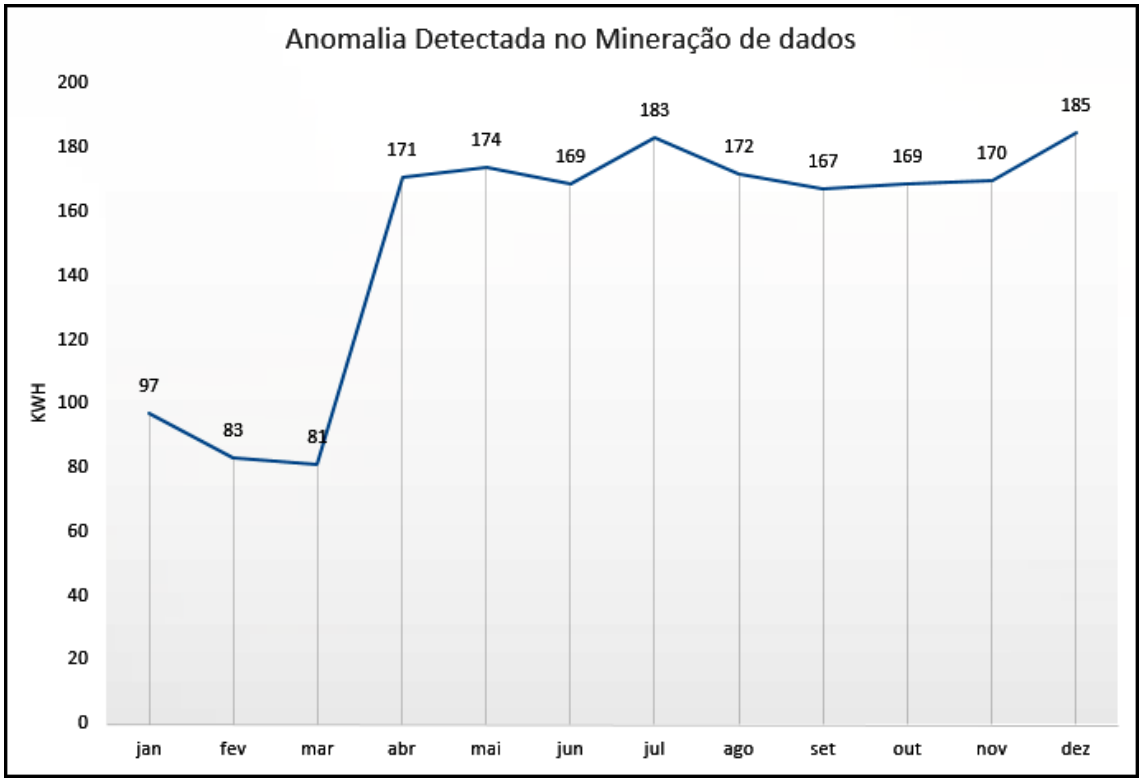

Figura 6.12: Caso de possível fraude não detectada no processo de classificação feito pela concessionária

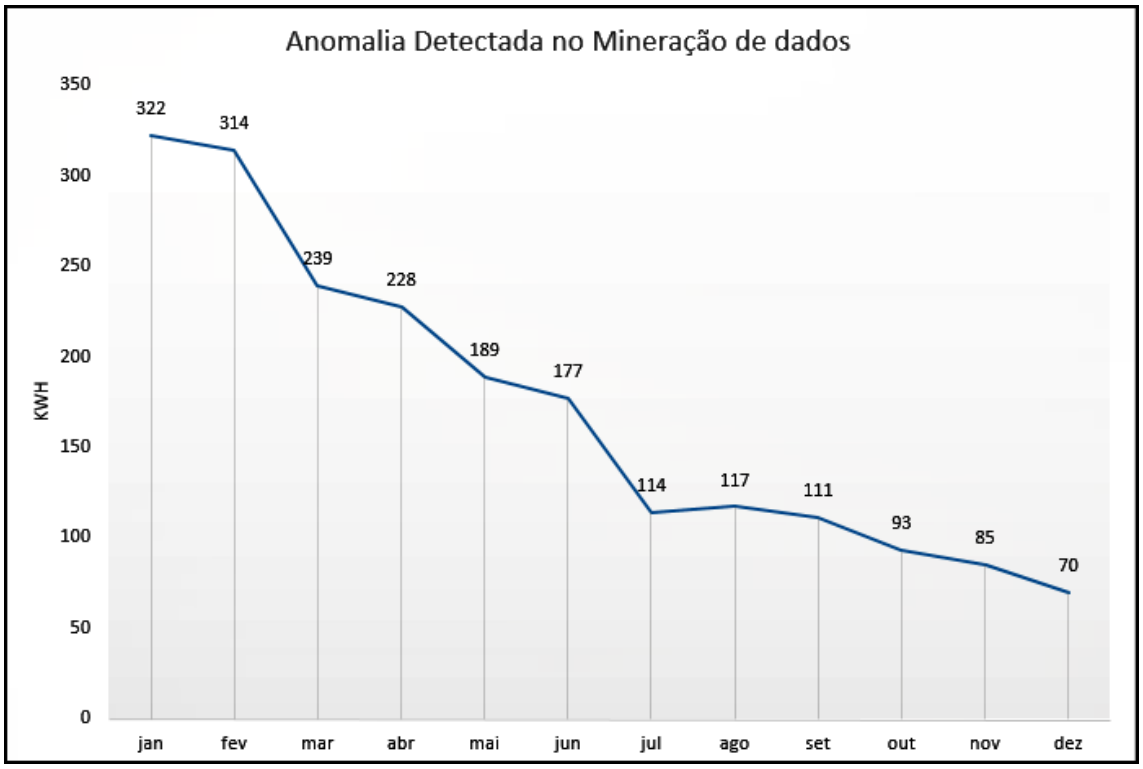

Figura 6.13: Caso de possível fraude não detectado no processo de classificação feito pela concessionária

6.6 Conclusão

Neste capítulo foram apresentados informações sobre a base de dados utilizada no processo de MD, mostrando as principais fraudes encontrados na base de dados estudada. Foi feito um breve relato da ferramenta WEKA, utilizada para extrair o conhecimento da base de dados.

Ainda neste capítulo, foi apresentado o processo de modelagem e pré-processamento dos dados para transformar a base de dados em um arquivo que pudesse ser reconhecido pela ferramenta WEKA e também excluir da base de dados, informações irrelevantes para o presente trabalho.

Observando a análise dos resultados apresentada neste capítulo, é possível concluir que a utilização da Inteligência Computacional, com o processo de MD, pode, com uma acurácia alta, auxiliar no processo de decisão das empresas.

Ainda é importante lembrar que uma base de dados é de difícil compreensão se observada manualmente os dados brutos, não passando por um processo de MD.

Comparando os dois algoritmos utilizados, é possível concluir que o algoritmo de RNA obteve resultados mais positivos, com uma acurácia mais alta que o algoritmo de RNA.

Capítulo 7

Conclusão

Neste capítulo tecem-se as informações obtidas durante todo o desenvolvimento do presente trabalho. Inicia-se com uma síntese da dissertação, seguindo de uma discussão sobre a mesma.

Ainda, para concluir, aponta-se algumas possibilidades de trabalhos futuros.

7.1 Síntese

Nos dias atuais, a competitividade no setor elétrico brasileiro e a busca por novas tecnologias no setor elétrico vêm crescendo de forma eminente. O aumento do crescimento dos dados armazenados diariamente com a leitura nos medidores vêm gerando a necessidade de transformar tais dados em conhecimento que possa ser utilizado para combater as perdas nas redes de energia. As perdas nas redes de distribuição de energia crescem na mesma velocidade que os avanços tecnológicos na engenharia elétrica.

Levando em consideração que os seres humanos possuem capacidade limitada de raciocínio, as técnicas estatísticas convencionais podem ser falhas e possuem uma elevada complexidade dependendo do tipo e quantidade de dados, tornando difícil distinguir qual informação é útil para colaborar na tomada de decisão em uma determinada situação.

Caracterizada pelas várias etapas das quais se destaca o processo de MD, o processo de KDD utiliza algoritmos de extração de conhecimento para detectar padrões em uma base de dados e transformar dados em conhecimento.

Nessa dissertação, tendo em mãos uma base de dados composta pelo consumo de energia elétrica de usuários de energia durante três anos, foram utilizados dois algoritmos de MD utilizados para classificação de dados. Para testar os algoritmos foram realizados diversos testes alterando as configurações disponíveis nos dois diferentes algoritmos, analisando as vantagens e desvantagens de ambos, buscando o melhor resultado em porcentagem de acerto da classificação dos dados.

É importante ressaltar que os algoritmos utilizados, sendo um de RNA e MVS, se distinguem dos demais algoritmos de MD por produzirem modelos que tendem a obter resultados melhores quando os dados apresentam relações não lineares.

Para execução dos testes com os dois algoritmos, a ferramenta escolhida para ser utilizada foi a ferramenta WEKA, tal ferramenta possui código aberto, é de fácil utilização, possui uma vasta quantidade de algoritmos inclusos na ferramenta e com configurações modificáveis.

Durante a fase experimental e levando em consideração que os algoritmos permitem a alteração de diversos parâmetros de configurações, optou-se por iniciar os testes com os valores padrões da ferramenta e então alterá-los individualmente em busca de uma melhor porcentagem de acerto em ambos os algoritmos para criação dos modelos.

Os resultados foram inseridos em tabelas e gráficos para facilitar a visualização e compreensão dos mesmos.

7.2 Peroração

Sendo que o objetivo principal do presente trabalho é detectar possíveis padrões de fraudes através das análises da curva de consumo dos consumidores de energia elétrica, pode-se afirmar que, com a utilização de técnicas de MD, foi possível encontrar padrões de consumo que possuíam anomalias e que com as técnicas de busca de manuais, utilizadas pelas concessionárias, não se conseguem detectar.

Sendo que este trabalho também pretendia analisar e comparar algoritmos de MD para classificação, pode-se afirmar que os resultados obtidos utilizando algoritmo de RNA obtiveram um resultado mais positivo, com uma porcentagem de acerto maior na classificação dos usuários fraudulentos quando comparado com o algoritmo de MVS.

É importante ressaltar que o tempo de treinamento e criação dos modelos com padrões de consumo através da análise de consumo de energia elétrica dos usuários disponíveis na base de dados é muito lento utilizando o algoritmo de RNA, podendo levar até semanas para o treinamento. Isso acontece devido o uso de ferramentas como o WEKA, ao invés de se desenvolver diretamente a RNA. Mas a partir dos dados treinados e os modelos criados, o processo de classificação com novos dados é muito rápido, chegando a classificar mais de 100.000 dados por segundo.

No caso do algoritmo de MVS, o qual obteve um resultado de acerto da classificação menor do algoritmo de RNA, possui um processo de treinamento e criação de modelos mais rápida, e o processo de classificação após os dados treinados é aparentemente igual ao algoritmo de RNA.

Os resultados obtidos com ambos os algoritmos encorajam suas utilizações como ferra-

mentas de auxilo nas tomadas de decisões. Alcançando mais de 90% de acerto, a MD se torna uma técnica viável, e essa porcentagem pode ser melhorada com novos treinamentos.

Essa alta porcentagem de acerto na classificação dos dados reduz uma grande quantidade de fiscalizações manuais que antes eram realizadas pelas concessionárias a fim de encontrar fraudes nas redes de distribuição de energia elétrica.

O fato de que a utilização de técnicas de MD permitiu encontrar novos casos de fraudes não detectados anteriormente, mostra que tais técnicas são mais eficientes que as técnicas usadas anteriormente de forma manual, e faz com que fique comprovado que, para bases de dados similares, a utilização de técnicas de Inteligência Computacional é positiva e permite a detecção de padrões em velocidade alta.

A utilização da ferramenta WEKA no presente trabalho foi positiva, e permitiu a criação do modelo e testes com facilidade de manuseio. A ferramenta utilizada apresentou ótima velocidade nos testes após a criação dos modelos com ambos algoritmos. Para a criação dos modelos, a ferramenta apresentou uma velocidade lenta quando aplicada a bases de dados de grande porte, o que poderia ser resolvido, no caso de uma aplicação empresarial, com o desenvolvimento da RNA e MVS para serem utilizados sem a ferramenta WEKA. Para testes e pesquisas, a ferramenta pode ser utilizada com facilidades por possuir algoritmos prontos, não necessitando o desenvolvimento dos mesmos. Todavia, em casos de aplicações empresariais, indica-se a o desenvolvimento dos algoritmos pois a ferramenta WEKA não permite que os algoritmos reaprendam com novos dados durante os testes, e possui baixa velocidade na criação de novos modelos com grandes bases de dados.

7.3 Trabalhos Futuros

Essa dissertação proporciona diversas perspectivas de trabalhos futuros que abordam o assunto de perdas de energia e a utilização de IC para diminuir tais perdas, entre essas possibilidades é possível citar:

- Levando em consideração o número elevado de algoritmos de MD disponíveis na ferramenta WEKA e também em outras ferramentas de mineração, é possível novas comparações entre outros algoritmos a fim de se obter melhor porcentagem de acerto na classificação dos dados;
- Os mesmos algoritmos de MD podem ser implementados em diversas ferramentas, essa possibilidade proporciona a possibilidade de realizar a comparação entre os resultados obtidos nesse trabalho com resultados obtidos através das mesmas técnicas em ferramentas diferentes ao WEKA;
- A utilização de filtros disponíveis na ferramenta WEKA ou filtros manuais podem alterar

os resultados, neste contexto, é possível fazer uma comparação de qual o melhor filtro para a base de dados utilizadas neste trabalho e base de dados similares;

- Desenvolvimento de RNA e MVS para serem utilizadas sem a ferramenta WEKA e comparar o resultado dos mesmos, analisando tempo e porcentagem de acerto.

Os trabalhos futuros não se resumem as sugestões citadas nesse trabalho, deixando aberto a possibilidade de diversas aplicações utilizando a mesma base de dados e buscando a melhor classificação de consumos anômalos na base de dados.

Referências Bibliográficas

ABRADEE (2015). Associação Brasileira de Distribuição de Energia Elétrica, Furto e Fraude de Energia - Associação Brasileira de Distribuidores de Energia Elétrica. Acesso em: maio/2015.
Disponível em: *http://www.abradee.com.br/setor-de-distribuicao/perdas/furto-e-fraude-de-energia*

Adriaans, P. & Zantinge, D. (1996). *Data Mining, 1ª ed*, Editora Addison-Wesley, Harlow - Inglaterra.

Agrawal, R. & Srikant, R. (2004). Fast Algorithms for Mining Association Rules, *Proceeding of the 20th VLDB Conference* pp. 487–499. Chile.

Alcantara, A. R. (2012). *Estudo de Ferramentas Baseadas no WEKA para Mineração de Dados Distribuídas em Ambientes de Grid*, Dissertação de mestrado, Universidade Federal de Lavras.

Ales, V. T. (2008). *O Algoritmo Sequential Minimal Optimization para Resolução do Problema de Support Vector Machine: Uma Técnica de Reconhecimento de Padrões*, Dissertação de mestrado, Universidade Federal do Paraná - UFPR.

Almeida, F. C. & Dumontier, P. (1996). O Uso de Redes Neurais em Avaliação de Riscos de Inadimplência, *Revista de Administração FEA/USP* **31**(1): 52–63. São Paulo.

Almeida, M. V. (2009). *Aplicação de Técnicas de Redes Neurais Artificiais na Previsão de Curtissimo Prazo da Visibilidade e Teto para Aeroporto de Guarulhos - SP*, Tese de doutorado, Universidade Federal do Rio de Janeiro - UFRJ.

ANEEL (2000). Agencia Nacional de Energia Elétrica, Condições Gerais de Fornecimento de Energia Elétrica - Resolução 456. Acesso em: junho/2015.
Disponível em: *http://www.aneel.com.br*

ANEEL (2015). Agencia Nacional de Energia Elétrica, Espaço do Consumidor - Perdas de Energia. Acesso em: setembro/2015.
Disponível em: *http://www.aneel.gov.br/area.cfm?idArea=801*

Aráujo, A. C. (2006). Considerações Sobre Perdas da Distribuição de Energia Elétrica no Brasil.

Batista, G. A. (2003). *Pré-processamento de Dados em Aprendizado de Máquina Supervisionado*, Tese de doutorado, Instituto de Ciências Matemáticas e de Computação, USP, São Carlos - SP.

Castanheira, L. G. (2008). *Aplicação de Técnicas de Mineração de Dados e Problemas de Classificação de Padrões*, Dissertação de mestrado, Universidade Federal de Minas Gerais - UFMG.

Coppin, B. (2013). *Inteligencia Artificial*, Editora LTC, Rio de Janeiro.

Cortez, P. (2002). *Modelos Inspirados na Natureza para a Previsão de Séries Temporais*, Tese de doutorado, Universidade do Minho - UM.

Costa, C. H. V. (2010). *Posicionamento Geográfico de Dispositivos Móveis em Ambientes Externos Utilizando a Tecnologia W-Fi e Redes Neurais Artificiais*, Dissertação de mestrado, Universidade Federal do Rio de Janeiro.

Coura, B. (2015). Acidentes com Fios da Rede Elétrica Causaram 299 Mortes em 2014. Noticias - Agência Brasil.
Disponível em: *http://www.ebc.com.br/print/noticias/2015/08/acidentes-com-fios-da-rede-eletrica-causaram-299-mortes-em-2014*

Craik, K. J. (1943). *The Nature of Explanation*, Editora Cambridge University Press.

Cruz, A. J. R. (2008). *Data Mining via Redes Neurais Artificiais e Máquinas de Vetores de Suporte*, Dissertação de mestrado, Universidade do Minho - UM.

Delaiba, A. S., Filho, J. R., Gontijo, E. M., Mazina, E., Cabral, J. E. & Pinto, J. O. P. (2004). Fraud Identification in Eletricity Company Costumers Using Decision Tree, *International Conference on System, Man and Cibernetics* **4**: 3730 – 3734. Netherlands.

Dosciatti, M. M., Ferreira, L. P. C. & Paraiso, E. C. (2013). Identificando Emoções em Texto em Português Brasileiro Usando Máquinas de Vetores de Suporte em Solução Multiclasse, *Encontro Nacional de Inteligência Artificial e Computacional* pp. 1–12. Fortaleza - CE.

Edelman, G. M. (1990). *Neural Darwinism: The Theory of Neuronal Group Selection*, Editora Oxford University Press.

Eller, N. A. (2003). *Arquitetura de Informação para Gestão de Perdas Comerciais de Energia Elétrica*, Dissertação de mestrado, Universidade Federal de Sâo Carlos, São Carlos - SP.

Fahlman, S. (1998). Faster Learning Variations on BackPropagation: An Empirical Study, *Proceedings od Connectionist Models Summer School*, pp. 38–51.

Fayyad, U., Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases, *American Association for Artifical Intelligence* **17**(3): 1–18. Press/AAAI Press - Cambridge.

Ferreira, J. G. H. M. (2008). *Tratamento de Dados Geotécnicos para Predição de Módulos de Resiliência de Solos e Britas Utilizando Ferramentas de Data Mining*, Tese de doutorado, Universidade Federal do Rio de Janeiro - UFRJ.

Guizzo, E. (2000). Quem Procura Acha, Revista Negócios Exame. Acesso em: maio/2014.
Disponível em: *http://portalexame.abril.uol.com.br/complementos/revista0002.html*

Hansen, A. M. D. (2000). *Padrões de Consumo de Energia Elétrica em Diferentes Tipologias de Edificações Residenciais, em Porto Alegre*, Dissertação de mestrado, Universidade Federal do Rio Grande do Sul - UFRS.

Haykin, S. (2001). *Redes Neurais - Princípios e Prática*, Editora Bookman. 1º ed.

Hebb, D. O. (1949). *The Organization of Behavior*, Editora Wiley.

Hernandes, L. J., Duarte, L. F. C., MORAIS, F. O., FERREIRA, E. C. & DIAS, J. A. S. (2013). Processo não invasivo de baixo custo para otimização da rotina de inspeção na detecção de furto de energia elétrica, *Revista Pesquisa e Desenvolvimento da Aneel* **5**(1): 47–51.

Indurkhya, N. & Weiss, S. M. (n.d.). Estimating performance Gains for Voted Decision Trees, IBM Research Division Technical Report in Intelligent Data Analysis (IDA). 1999.

Lópes, G. P., Rider, M. J. & Zapata-Grisales, C. J. (2014). Detecção de Fraudes nas Redes de Distribuição de Energia Elétrica Utilizando Árvore de Decisão, *Simpósio Brasileiro de Sistemas Elétricos* pp. 1–6. Anais 2014 - Foz do Iguaçu - PR.

Luce, D. R. & Raiffa, H. (1957). *Games and Decisions*, Editora Wiley.

Mannila, H. (1996). Data Mining: Machine Learning, Statistics and Databases, *In Proceeding of the eight IEEE International Conference* pp. 2–9. ISBN: 0-8186-7264-1.

Marreto, D. A. (2011). *Aplicação de Máquinas de Vetores de Suporte para Desenvolvimento de Modelos de Classificação e Calibração Multivariada em Espectroscopia no Infravermelho*, Tese de doutorado, Universidade Estadual de Campinas - UNICAMP.

McCulloch, W. S. & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity, *Bulletin od Mathematical Biophysics* **5**(1): 115–137.

Mendoza, L. A. F. (2009). Redes Neurais e Máquinas de Vetores de Suporte no reconhecimento de locutor usando coeficiente MFC e características do sinal glotal.

Neves, L. A. P., Neto, H. V. & Gonzaga, A. (2012). *Avanços em Visão Computacional*, Editora Ominipax, Curitiba - PR.

Newell, A. & Simon, H. A. (1972). *Human Problem Solving*, Editora Prentice-Hall.

Nimer, F. & Spandri, L. (1998). Data Mining, *Revista Developers*. pag.32.

Norvig, P. (1992). *Paradigms of Artificial Intelligence Programming: Case Studies in Common Lisp*, Editora Morgan Kaufmann.

Passos, J. C. (2009). Os experimentos de Joule e a primeira lei da termodinâmica, *Revista Brasileira de Ensino a Física* **31**(3): 1–18.

Penin, C. A. S. (2008). *Combate, Prevenção e Otimização das Perdas Comerciais de Energia Elétrica*, Tese de doutorado, Universidade de São Paulo.

Possa, B. A. V., Carvalho, M. L. B., Rezende, R. S. F. & Meira, J. (1998). *Data Mining: Técnicas para Exploração de Dados*, Dissertação de mestrado, Universidade Federal de Minas Gerais - UFMG.

PRODIST (2015). Procedimento de Distribuição de Energia Elétrica no Sistema Elétrico - PRODIST, Espaço do Consumidor - Perdas de Energia. Acesso em: setembro/2015. Disponível em: *http://www.aneel.gov.br/area.cfm?idArea=801*

Rebelo, L. D. T. (2008). *Avaliação Automática dos resultados estéticos do Tratamento Conservador do Cancro da Mama*, Dissertação de mestrado, Faculdade de Engenharia - Universidade do Porto.

Riedmiller, M. (1994). Supervised Learning in Multilayer Perceptrons - from backpropagation to Adaptive Learning Techniques, *Computer Standards an Interfaces* **16**(3): 265–275.

Rosenblatt, F. (1958). The Perceptron: a probabilistic model for information storage and organization in the brain, *Psychological Review* **65**(6): 386–408.

Rumelhart, D. E. & McClelland, J. L. (1986). *Parallel Distributed Processing*, Editora MIT Press.

Russell, S. & Norvig, P. (2013). *Inteligencia Artificial*, Editora Campus - Elsevier, São Paulo.

Sá, H. E. G. (2010). *Um Método Baseado em Inteligência Computacional para Geração Automática de Casos de Testes de Caixa Preta*, Dissertação de mestrado, Escola Politécnica da Universidade de São Paulo.

Schuch, R., Dill, S. L., Suasen, P. S., Padoin, E. L. & Campos, M. (2010). Mineração de Dados em uma Subestação de Energia Elétrica, *9° Brazilian Conference on Dynamics: Control and their Applications - DINCON10* pp. 804–810. Anais 2010, Serra Negra, SP.

Shortliffe, E. H. (1972). *Computer-Based Medical Consultations: MYCIN*, Editora Elsevier/North-Holland.

Silva, E. M. (2000). *Avaliação do Estado da Arte e produtos. Data Mining*, Dissertação de mestrado, Universidade Católica de Brasília - UCB.

Silva, L. R. M. (2012). *Inteligência Computacional aplicada à Modelagem de Cargas Não-Lineares e Estimação de Contribuição Harmônica*, Dissertação de mestrado, Universidade Federal de Juiz de Fora.

Tan, P. N., Steinbach, M. & Kumar, V. (2005). *Introduction to Data Mining*, Addison - Wesley Longman Publishing Co. Boston, MA, USA.

Thearling, K., Berson, A. & Smith, S. (1999). *Building Data Mining Application for CRM*, Editora McGrawHill.

Uysal, I. & Guvenir, H. A. (2004). Instance-based Regression by Partitioning Feature Projections, *Applied Intelligence*, pp. 57–59.

Vapnik, V. (1979). Estimation of Dependences Based on Empirical Data. Nauka - Moscow.

Waikato, U. (2015). Weka - Machine learning Software in java, The University of Waikato. Acesso em: agosto/2015.
Disponível em: *http://www.cs.waikato.ac.nz.ml.weka*

Zaki, M., Karypis, G. & Yang, J. (2007). Data Mining in Bioinformatics (BIOKDD), *Algorithms for Molecular Biology*, Vol. 2.