PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE ENGENHARIA ELÉTRICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

MARLON MENDES MINUSSI

METODOLOGIA DE MINERAÇÃO DE DADOS PARA DETECÇÃO DE DESVIO DE COMPORTAMENTO DO USO DE ENERGIA EM CONCESSIONÁRIA DE ENERGIA ELÉTRICA

Porto Alegre
2008

MARLON MENDES MINUSSI

METODOLOGIA DE MINERAÇÃO DE DADOS PARA DETECÇÃO DE DESVIO DE COMPORTAMENTO DO USO DE ENERGIA EM CONCESSIONÁRIA DE ENERGIA ELÉTRICA

Dissertação apresentada para obtenção do grau de Mestre, pelo Programa de Pós-graduação em Engenharia Elétrica da Faculdade de Engenharia Elétrica da Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Dr. Rubem Dutra Ribeiro Fagundes

Porto Alegre
2008

MARLON MENDES MINUSSI

METODOLOGIA DE MINERAÇÃO DE DADOS PARA DETECÇÃO DE DESVIO DE COMPORTAMENTO DO USO DE ENERGIA EM CONCESSIONÁRIA DE ENERGIA ELÉTRICA

Dissertação apresentada para obtenção do grau de Mestre, pelo Programa de Pós-graduação em Engenharia Elétrica da Faculdade de Engenharia Elétrica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovada em ___ de _____ de 2008 .

BANCA EXAMINADORA:

_____
Dr. Rubem Dutra Ribeiro Fagundes

_____
Dr. José Vagner Maciel Kaehler

_____
Dr.Maurício Amaral de Almeida

_____
Dra. Maria Cristina Felippetto de Castro

Esta dissertação é dedicada a minha namorada Gabriela, meus pais Cleone e Elaine, minhas irmãs Márcia e Manuela.

AGRADECIMENTOS

Agradeço a minha família por me dar a educação e o apoio necessário.

A minha namorada e sua família pelo incentivo incessante.

A meus amigos assisenses de longa data.

Aos colegas e amigos do GPGE, pela ajuda, companheirismo e gargalhadas.

Ao Prof. Dr. José Wagner Maciel Kaehler e o Prof. Dr. Rubem Dutra Ribeiro Fagundes pela orientação ao longo desta dissertação.

A Fabiano Caetano Etchichury, pela amizade e pela grande ajuda nesta conquista.

Ao Prof. Dr. Jairo Bevegnu Bordinhão pela inspiração.

A Luis Carlos Lange colega de mestrado e um grande minerador de dados.

As gurias da Secretaria do Programa de Pós-graduação de Engenharia Elétrica da Pontifícia Universidade Católica do Rio Grande do Sul, pelo grande ajuda com os assuntos burocrático.

A Deus pela força.

A educação é um processo social, é desenvolvimento. Não é a preparação para a vida, é a própria vida.

John Dewey

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

METODOLOGIA DE MINERAÇÃO DE DADOS PARA DETECÇÃO DE DESVIO DE COMPORTAMENTO DO USO DE ENERGIA EM CONCESSIONÁRIA DE ENERGIA ELÉTRICA

MARLON MENDES MINUSSI

Fevereiro 2008.

Orientador: Dr. Rubem Dutra Ribeiro Fagundes.
Área de Concentração: Sistemas de Energia.
Linha de Pesquisa: Planejamento e Gestão de Sistemas de Energia.
Projeto de Pesquisa Vinculado: Gestão de energia em Programas Anuais de Eficiência Energética e Promoção do Uso Racional de Energia.
Palavras-chave: Mineração de Dados, *Data Warehouse*, Desvio de comportamento de uso de energia elétrica, Banco de Dados.

Com abertura do mercado de energia elétrica e o aumento da competitividade no setor elétrico brasileiro, as concessionárias de energia buscam ferramentas para minimizar as perdas comerciais e maximizar seus lucros. Visando solucionar este problema foi desenvolvido um método de mineração de dados para detecção de desvio de comportamento no uso de energia em concessionária de energia elétrica. Pois quanto menos perde-se, menos precisa ser gerado, e menos se desperdiça recursos naturais.

Na elaboração do método compreendeu etapas de análise e avaliação dos dados, assim como construção de um *Data Warehouse* mais adequado para o desenvolvimento deste trabalho. Foram analisadas curvas de cargas dos clientes e através dessa análise observou-se o perfil de consumo dos mesmos, embasados na análise foram aplicados os algoritmos de mineração de dados, como o algoritmo de associação Apriori para fornecer padrões de indicadores de perfil dos consumidores bem como os algoritmos de Árvore de Decisão e Classificadores Bayesianos.

Os resultados validam o método desenvolvido e implementado permitindo sua utilização em uma concessionária de energia elétrica sendo utilizado como mais um ferramenta de GLD para auxiliar e somar-se as a ações já existentes na concessionária.

Abstract of Dissertation presented to PUCRS as one of the requirements to obtain Masters Degree in Electrical Engineering.

METHODOLOGY OF DATA MINING FOR DETECTION OF FRAUD OR THEFT OF ENERGY IN CONCESSIONAIRE OF ELECTRICAL ENERGY

MARLON MENDES MINUSSI

February 2008

Advisor: José Wagner Maciel Kaehler
Concentration Field: Energy Systems
Line of Research: Planning and Energy Management Systems
Link Research Project: Management of energy in Programs Annual Energy Efficiency and Promotion of Rational Use of Energy
Keywords: Data Mining, Data Warehouse, Fraud or theft of electricity, Database, Artificial Intelligence, Statistics.

With opening of the market of electric energy and increased competitiveness in the Brazilian electric sector, the energy companies seek tools to minimize losses and maximize their commercial profits. In order to have a balance it was developed a data mining method to discover the bad user behavior of the use of energy in electrical energy company. The less is lost, less needs to be generated and less natural resources are wasted.

Stages of review and evaluation of data, as well as construction of a Data Warehouse more appropriate for the development of this work were accomplished. Customers Load curves were analyzed and through this analysis there was the profile of consumption of these customers and through this analysis the data mining algorithms are applied. The association algorithm provides indicators patterns of consumers profile besides a decision tree and Bayesianos Classifyings.

The results validate the developed and implemented method allowing their use in an electric energy company being used as another tool for GLD to help and add to the existing actions in the company.

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE EQUAÇÕES

SUMÁRIO

Capítulo 1

1 Introdução

1.1 Apresentação

Tendo em vista a abertura do mercado de energia elétrica e o aumento da competitividade no setor elétrico brasileiro, as concessionárias de energia buscam ferramentas que vão ao encontro dos objetivos de minimização de perdas comerciais e maximização dos recursos financeiros disponíveis para investimentos. Dentre as alternativas disponíveis hoje temos a política de Gerenciamento pelo Lado da Demanda (GLD), que é constituído por ações concebidas, implementadas e fundamentadas no contexto de companhias de eletricidade.

As perdas em um sistema elétrico podem ser divididas em perdas técnicas e perdas comerciais. As perdas técnicas representam as perdas inerentes ao processo de distribuição, ou seja, aquelas que ocorrem durante o transporte da energia, enquanto que as perdas não técnicas ou comerciais (ligações clandestinas, fraudes, auto-religação e erros não intencionais), são aquelas que ocorrem quando a energia é entregue ao consumidor, sendo computadas pela concessionária, porém não sendo as mesmas faturadas adequadamente. Isso acontece não somente em regiões socialmente desfavorecidas, onde elas são presentes através do roubo de energia, mas também em empresas e instituições cujas atividades dependam fundamentalmente do fornecimento de energia para suas operações. Nestes casos encontram-se a fraude e o roubo de energia. Em muitas concessionárias brasileiras, estas perdas ultrapassam 30% da comercialização de energia, acarretando uma deterioração na qualidade de serviço prestado, assim como um prejuízo na receita para as empresas distribuidoras de energia. No Brasil, segundo a ANEEL (Agência Nacional de Energia Elétrica), em média 15% da energia elétrica gerada no país é furtada.

Durante os anos 90, com o processo de privatização, iniciou-se a desverticalização do setor elétrico. Com isso, as tarifas de energia elétrica passaram a ser estabelecidas nos Contratos de Concessão.

No modelo verticalizado, os gastos envolvidos em perdas tanto comerciais como técnicas eram repassadas para as contas dos clientes ou o Estado assumia este prejuízo. Diferentemente, no modelo desverticalizado este tipo de custo é inteiramente repassado aos clientes da área de concessão [PATRICIO, 2005].

Dado este cenário, para que haja maior controle dos mercados, surgiram as agências reguladoras. Assim, em 1997 foi criada a ANEEL, autarquia com vínculo ao Ministério de Minas e Energia (MME). Tendo a ANEEL a atribuição de fiscalização sobre a geração, transmissão e comercialização de energia elétrica, a referida agência atende o consumidor e tenta promover um equilíbrio entre as partes, procurando deste modo beneficiar a sociedade.

A partir desse novo modelo e em função do crescimento constante das Tecnologias de Informação, as concessionárias passaram a buscar novas maneiras de armazenamento de dados referentes a seus clientes, com o intuito de ter um melhor controle das informações de uso da energia por parte dos mesmos, bem como do tratamento dos dados da própria empresa. Por exemplo, fazer a análise de dados e verificar o perfil de consumo dos clientes da classe Horo-sazonal, onde cada cliente gera noventa e seis registros diários. Com esse avanço, levando em consideração a quantidade de clientes e o número de dias, atualmente os dados obtidos são de centenas de milhares de registros diários. Para a concessionária ter um bom gerenciamento destes dados e um melhor aproveitamento dessas informações, faz-se necessário o uso de um SGBD's (Sistema de Gerenciamento de Banco de Dados) aliado a várias técnicas e métodos referentes à manipulação e interpretação dos dados.

Para uma adequada exploração desses bancos de dados e aprofundamento do conhecimento dessa rica massa de informações, encontraram-se disponíveis na literatura aplicações utilizando KDD (*Knowledge Discovery in Database* - Descoberta de conhecimento em Banco de Dados) e Mineração e Dados (*Data Mining* - MD).

Aplicações práticas encontram-se principalmente nas áreas:

- financeira, a fim de obter a redução de fraudes, previsão de falências, perfil de clientes usuários de crédito;

- no Mercado de Negócios, tendo em vista a prospecção de clientes, ou estudo de lançamento de produto no mercado, através da análise sobre quais são os produtos com maior consumo;

- na Medicina com o objetivo de detectar ou prevenir doenças, detecção de sintomas em procedimentos cirúrgicos, ou grupos potenciais à contaminação por uma infecção [BOSE e MAHAPATRA, 2001].

No setor de energia elétrica e de telecomunicação existem alguns trabalhos publicados, tais como Aplicação de Técnicas de Mineração de Dados em Gestão de Sistemas de Energia Elétrica [SOARES, 2005], Aplicação e Modelo de Mineração de Dados para Classificação de Clientes e, Telecomunicações [PETERMANN, 2006]. Neste último foram utilizados três métodos visando à classificação e predição de clientes com previsão de cancelamento de serviço através do consumo e faturamento. Os resultados obtidos por estes métodos validaram o modelo, permitindo sua utilização e aperfeiçoamento no ambiente de uma operadora de telecomunicações.

No setor de distribuição de energia ainda não existem muitas referências a respeito do uso das descobertas que estão sendo feitas e conseqüentemente do conhecimento que se tem na área. Este vem crescendo cada vez mais com as transformações ocorridas no setor elétrico nas últimas duas décadas, em função da desverticalização, que ocasionou o aumento da competitividade nos segmentos de geração e de comercialização da energia elétrica. O processo de MD vem aumentando e está se tornando cada vez mais popular nas mais diversas áreas, como já citado, com resultados bastante positivos e proveitosos, principalmente onde se tem uma grande quantidade de informações, como é o caso das concessionárias de energia elétrica.

1.2 Objetivos da Dissertação

Diante da contextualização estabelecida e frente ao rápido crescimento da tecnologia de informação e das redes de comunicação, busca-se através dessa dissertação desenvolver uma metodologia com base no *KDD,* envolvendo a Mineração de Dados como sua etapa principal. Fundamenta-se no desenvolvimento de um *Data Warehouse* que permita a segregação e a detecção das fraudes comerciais nas diversas tipologias de clientes, através de análise de padrão de consumo.

Partindo de uma revisão bibliográfica e de uma avaliação de aplicações práticas em certos segmentos, será feita uma análise visando fornecer informações consistentes para que

as concessionárias de energia possam efetuar ações que permitam combater as fraudes comerciais.

A aplicação da metodologia desenvolvida nesta dissertação poderá indicar o local onde possivelmente esteja ocorrendo uma fraude. Utilizando os bancos de faturamento e de consumo, os dados serão analisados e submetidos aos processos de Mineração.

A metodologia desenvolvida será testada junto aos clientes Horo-sazonais da região de fronteira do Estado do Rio Grande do Sul, onde serão levados em consideração os arquivos de memória de massa dos medidores desses clientes. Será ponderado e avaliado o comportamento do consumo diário desses clientes, sendo observada desta maneira a ocorrência de desvio no padrão de consumo dos mesmos, assinalando todo o comportamento anormal.

O software de Mineração de Dados a ser utilizado será o Weka, por ser um software de distribuição gratuita, desenvolvido pela Universidade de Waikato, Nova Zelândia, muito utilizado para a Mineração de Dados.

Os resultados possibilitam ainda auxiliar na elaboração de ações de GLD da concessionária buscando melhorar a qualidade e a eficácia nas tomadas de decisões. Torna-se, assim, uma ferramenta de auxílio no combate ao roubo de energia ou fraude, completando as demais ações existentes para este fim.

1.3 Justificativa

As perdas comerciais de energia elétrica representam um enorme prejuízo, tanto para a sociedade quanto para a concessionária. Contribuem fortemente para a deterioração da qualidade do serviço prestado pela concessionária, bem como podem gerar uma série de riscos, inclusive à saúde humana, assim como a perda de receita, decorrendo pagamento de multas ou indenização.

O uso indevido ou ilegal da energia elétrica não isenta a concessionária de parte da responsabilidade quanto aos riscos e danos que podem incorrer clientes ilegais. Juridicamente a responsabilidade fica imputada à concessionária.

Partindo dessas premissas, as concessionárias buscam a redução destas ações indevidas, o que serviu de motivação a realização desta dissertação na busca de uma solução que venha a subordinar o combate ou a prevenção de atos ilícitos.

A classe horo-sazonal de clientes das distribuidoras de energia gera diariamente uma massa de dados significativamente grande, advinda de várias fontes. A classe de clientes que dispõem da estrutura tarifária horo-sazonal contam com medidores eletrônicos, os quais registram a cada quinze minutos suas demandas ativas e reativas, integralizando-as a cada hora em termos de energia ativa e reativa e seu correspondente fator de potência. Estes dados são carregados na memória de massa do medidor, sendo mensalmente descarregados junto aos servidores da concessionária.

A informação assim obtida é um recurso muito valioso para as organizações empresariais. A sua valorização torna-se um instrumento capaz de reduzir os riscos e as falhas e aumentar a competitividade das empresas.

Faz-se necessário diferenciar a análise dos dados vindos dos bancos de consumo, faturamento daqueles advindos das memórias de massa coletadas nos aparelhos de medição de grandes consumidores. A Mineração de Dados de grandes consumidores tem um processo diferente da Mineração de clientes convencionais em baixa tensão. Nestes o faturamento é feito mensalmente pelo registro da energia elétrica utilizada no e pela demanda verificada ou contratada no caso dos clientes empresariais ligados em baixa tensão.

Os dados de consumo e faturamento passarão por uma transformação, tornando-os mais acessíveis para o entendimento, facilitando a aplicação dos métodos e passos necessários para um bom resultado da Mineração. O controle da fraude e do roubo de energia elétrica não deixa de ser uma medida de Gerenciamento pelo Lado da Demanda (GLD), tornando-se uma ferramenta de auxílio, completando as demais ferramentas existentes nas tomadas de decisão.

Como as concessionárias dispõem de uma infinidade de dados sobre seus clientes, é necessário, em um primeiro momento, analisar a qualidade dos mesmos, para um processo correto, desde a captação até o uso final destes dados. Desta maneira, os dados captados devem passar por uma transformação para que sejam armazenados adequadamente, facilitando assim o uso das técnicas de Mineração de Dados.

O processo de Mineração de Dados pode ser utilizado como ferramenta na sinalização de fraudes, pois através da análise de dados pode-se apresentar uma predição ou um desvio do comportamento do padrão de consumo de determinada classe, indicando os clientes com potenciais de fraudar, assim como outra série de fatores que justifiquem a possibilidade de fraudes.

As concessionárias geralmente apresentam uma boa estrutura de aquisição de dados de seus clientes, só que as mesmas pouco exploram os bancos de dados. Por exemplo, são emitidos apenas relatórios de tabelas com dados referentes a seus consumidores por classe de consumo. A utilização de um cadastro para fins de fiscalização necessita de um entendimento mais detalhado e aprofundado dos campos da tabela de cadastro existente. Logo, faz-se necessário o preenchimento correto destes para contemplar as exigências do fiscalizador (consultas e geração de relatórios). Para isto, é necessário um conhecimento maior do uso de energia por parte do cliente. A informação existente no cadastro da concessionária depende diretamente da legislação (especificamente Portaria 466 do DNAEE de 12 de novembro de 1997) [SAUER, 2000].

1.4 Metodologia da Dissertação

Inicialmente, é preciso levar em consideração que os bancos de dados das concessionárias encontram-se estruturados de forma que inviabiliza uma análise mais criteriosa relacionada aos objetivos da metodologia proposta por esta dissertação, principalmente por apresentarem muitos dados duplicados, nulos, campos vazios e dados inconsistentes. Além disto, também se encontram tabelas relacionadas que possuem campos com nomes diferentes, sendo que as informações contidas são as mesmas, prejudicando a análise dos dados. Outro problema é uso de tabelas com um número muito grande de campos, o que aumenta em muito o tempo de retorno das consultas. Por este motivo tornou-se necessário um processo de normalização, visando assim à criação de um banco mais adequado aos procedimentos de Mineração de Dados, tornando as consultas mais eficientes para esse fim.

As informações provindas dos Bancos de Dados da concessionária estão divididas em dois grupos: (a) Informações Comerciais, que se referem à identificação dos consumidores, local de cobrança, tipo de atividade, etc.; e (b) Informações técnicas dos alimentadores, transformadores, redes, postes, medidores.

Através do conhecimento e da estratificação das curvas típicas dos clientes e com os dados armazenados no Banco de Dados (BD) da concessionária que registram mensalmente ou diariamente os contratos de suprimento de consumo dos diferentes consumidores, foi criada uma metodologia com base no KDD, tendo como objetivo auxiliar a concessionária na identificação de fraude ou roubo de energia através da MD.

A mineração buscará os dados registrados nos cadastros comerciais da concessionária, juntamente com as dados adquiridos das memórias de massa. Utilizando-se de algoritmos baseados em Redes Neurais, Árvore de Decisão e Classificadores Bayesianos, algoritmos estes abordados no Capítulo 4, será possível detectar as fraudes.

Serão utilizados métodos para limpeza (remoção de dados irrelevantes) e validação, tratamento e redução dos dados. Estes dados resultarão do confronto das informações de consumo e faturamento obtidos nos diversos Bancos de Dados da concessionária para que possa ser feito um levantamento prévio das informações que sejam relevantes para o projeto. Após esta etapa, será desenvolvido um banco de dados respeitando as regras de construção de um *Data Warehouse*, a fim de melhor estruturar os referidos dados. Esta etapa pode ser auxiliada por alguma etapa da Mineração de Dados, já assim fazendo uma seleção dos dados. Com o Banco de Dados consolidado, será permitida a aplicação de regras de Gerenciamento de Informações utilizando Mineração de Dados, a fim de extrair resultados relevantes para esta dissertação.

Os resultados obtidos, tais como detectar regiões e classes de consumo onde possam estar ocorrendo fraudes, serão adquiridos através de um software de Mineração de Dados. Este apresentará algumas saídas, resultados, conforme os algoritmos genéticos.

Os passos seguidos para a execução da dissertação e estudo de caso são os seguintes:

1º. Seleção dos dados vindos dos diversos bancos da concessionária, como o banco de consumo e de faturamento, os quais passam por vários processos, dentre eles o tratamento, a limpeza e a redução. Esta fase está no nível de dados brutos, os elementos puros inseridos pelo usuário no sistema da empresa ou dados advindos de aparelhos de medição e incluídos na base de dados da empresa.

2º. Criação do *Data Warehouse*, que é resultado das informações relevantes procedentes da etapa anterior. Ainda que não seja obrigatória sua construção, esta pode reduzir

significativamente a complexidade e o tempo do processo de Mineração de Dados. Essa fase está no nível de informação[1], que nada mais é do que os dados analisados, isto é, o fruto dos dados de entrada após os processos descritos anteriormente.

3º. Após, os dados poderão ou não passar por uma Mineração; a seguir, os mesmos serão inseridos no *Data Warehouse*, passando por uma Mineração para a detecção de fraude e submetidos à técnicas e algoritmos anteriormente citados, visando a obtenção dos resultados, os quais podem ser apresentados em formas de planilhas, tabelas ou gráficos. Esta etapa está no nível de conhecimento, onde são obtidos os resultados, para tomada de decisões por parte da concessionária.

A Figura 1 ilustra os passos descritos anteriormente:

[1] Os resultados obtidos através de consultas *OLAP* (Online *Analytical Processing*) são diferentes dos obtidos com a MD, as consultas *OLAP* são informações geradas a partir do *Data Warehouse* já informações e conhecimento extraído com a utilização da Mineração de Dados pode ser aplicada nos dados do *Data Warehouse* ou ainda diretamente nos dados dos bancos[REZENDE, 2003].

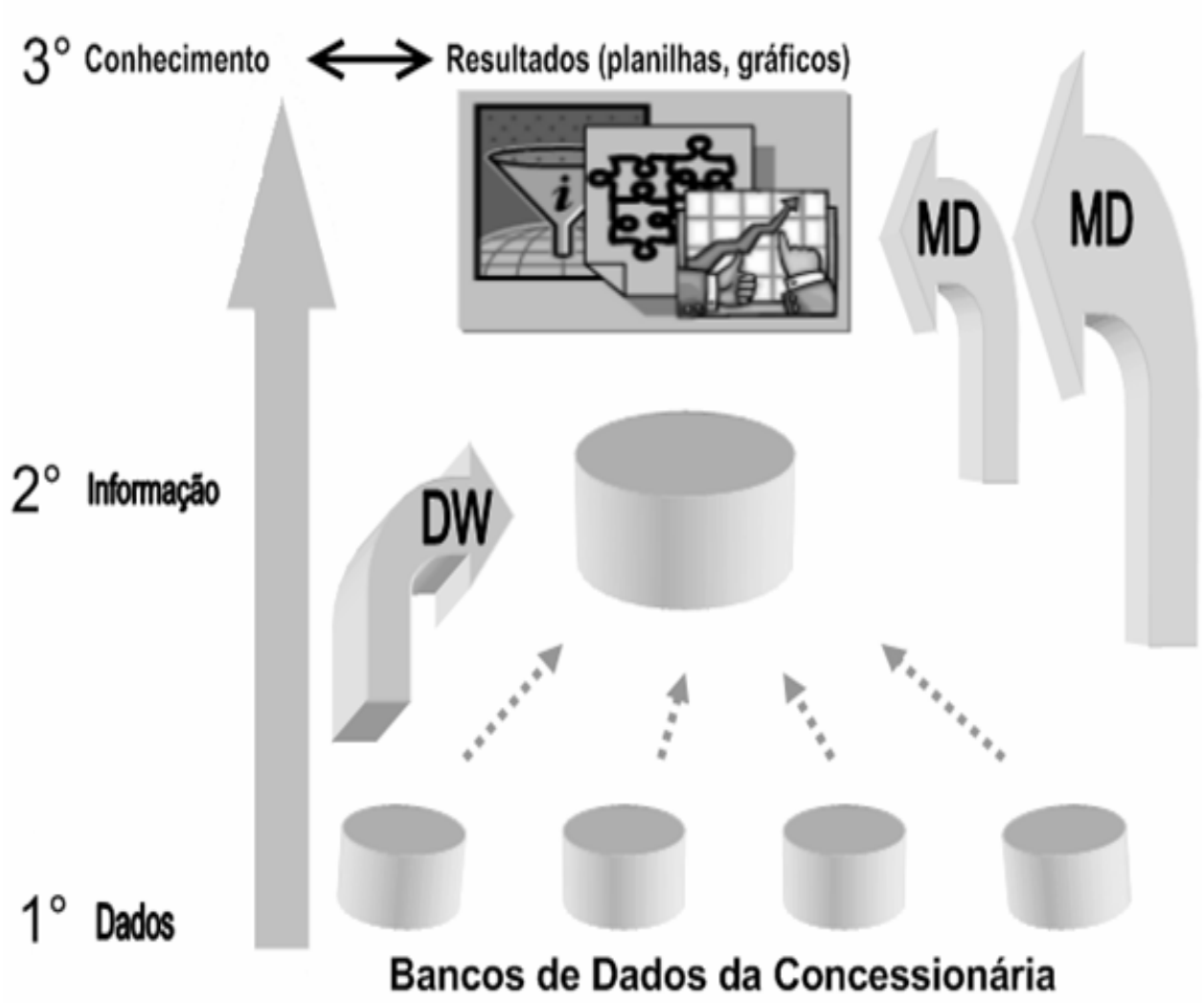


Figura 1 - Etapas que deverão ser executadas neste estudo de caso.

Figura baseada no trabalho de REZENDE et tal (2003).

A Figura 2 apresenta um Diagrama de Entidade de Relacionamentos[2] (DER) do BD, que foi implementado.

O diagrama da Figura 2 corresponde ao BD, após ter passado por alguns processos como a normalização, limpeza, etc., ele é transformado em um *Data Warehouse*, que é a junção de dois bancos da concessionária (o banco de consumo e o de faturamento), só que com as tabelas relevantes ao processo de Mineração. A tabela SGC_RESUMO, onde se

[2] O DER é um diagrama no qual é representado um banco de dados, suas tabelas, entidades, relacionamento, geralmente após passar por um procedimento de análise de sistema.

encontram os dados referentes ao faturamento da concessionária junto aos clientes foi submetida a um processo de normalização (que será explicado no capítulo seguinte). Esta tabela possuía 150 campos com milhões de registros. Assim, as consultas no Bando e Dados que envolviam esta tabela se tornavam demoradas demais. Desta forma, foram retirados os campos irrelevantes ao processo. Outros campos referentes aos dados dos clientes (dados pessoais) acabaram sendo inseridos em uma tabela nova chamada de CLI_CADASTRO.

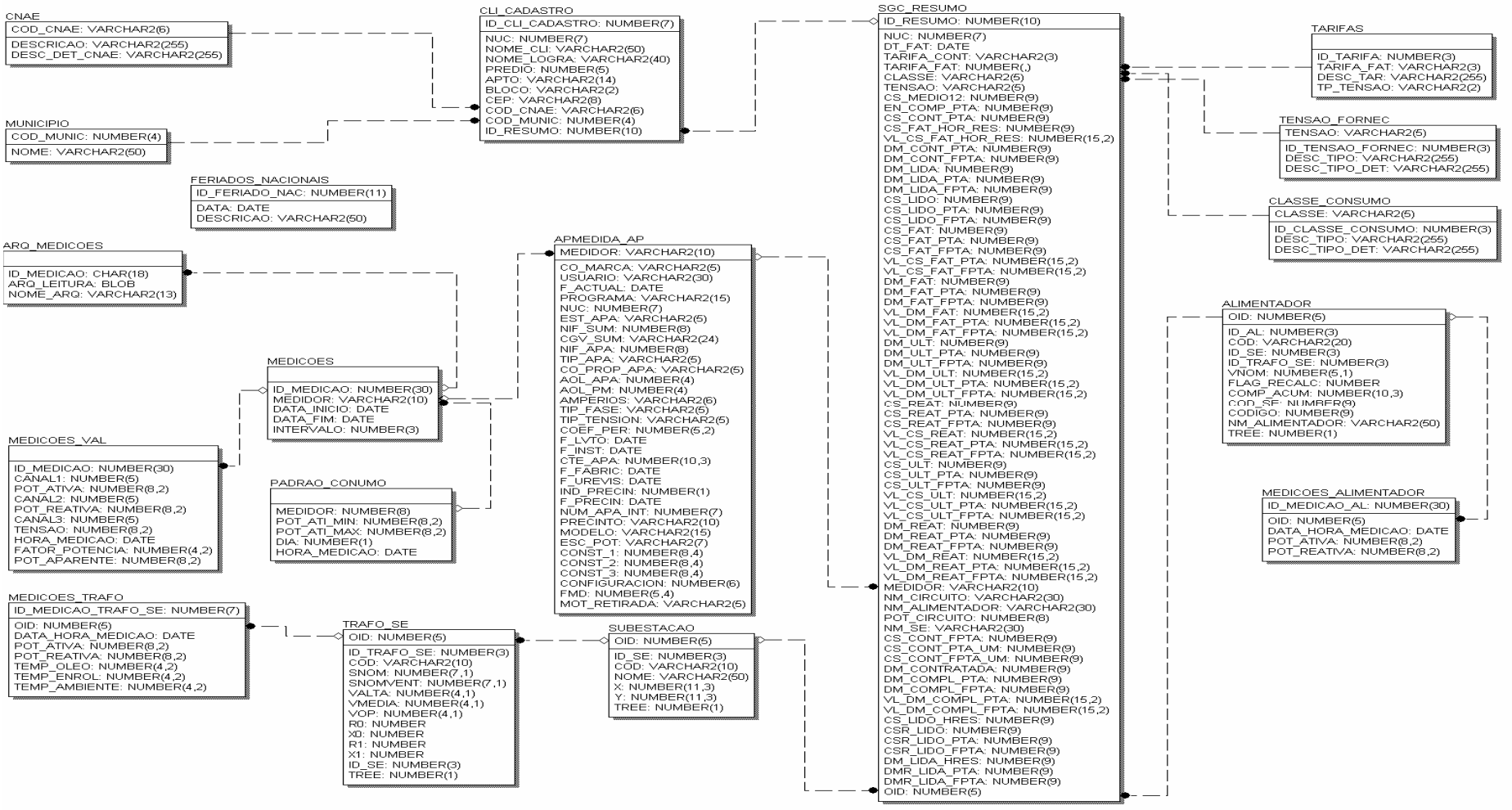


Figura 2 - Diagrama ER

1.5 Estrutura da Dissertação

Cada capítulo descreve em detalhes todos os resultados obtidos na busca dos objetivos da dissertação. Sendo assim, será descrita a seguir a forma de organização em termos dos objetivos propostos:

- O capítulo 1 contém uma introdução, na qual é apresentado o problema e a motivação do trabalho, prospecção de casos de sucesso, de mineração bem sucedida em diversos segmentos. Define-se o objetivo, a justificativa e uma idéia da metodologia desenvolvida, ou seja, uma visão geral do conteúdo dessa dissertação.

- No capítulo 2 é tratada a Descoberta de conhecimento em Base de Dados. Nesse capítulo, encontram-se os conceitos básicos para o melhor entendimento da dissertação, além das definições, da estrutura e das técnicas que compreendem desde a criação de um diagrama DER, a criação de um Data Warehouse e todos os outros métodos envolvidos até chegar à Mineração de Dados.

- No capítulo 3 são apresentadas as principais técnicas de descoberta de padrões, como técnicas de predição, como a definição do Problema, População, Amostragem, Triagem dos Dados e a Transformação dos mesmos.

- No capítulo 4 serão apresentadas as técnicas para classificação e predição para Detecção de Fraude, bem como os tipos de algoritmos utilizados no estudo de caso. Também serão descritas as características de Redes Neurais e Árvore de Decisão.

- No capítulo 5 será mostrado o método desenvolvido, o qual será utilizado no estudo de caso visando à detecção de fraude em clientes Horo-sazonais da região da fronteira do estado do Rio Grande do Sul, método este com o objetivo de procurar e definir um padrão, qual seja um perfil de consumo normal e compará-lo com os perfis dos outros clientes, assim detectando a presença de um desvio muito grande desse padrão, o que pode indicar um perfil fraudulento.

- No capítulo 6 é realizado o estudo de caso propriamente dito, através do qual são apresentados os resultados obtidos por meio da utilização da metodologia proposta pelo uso de um software de Mineração de dados, com a utilização dos algoritmos selecionados para este processo, sendo mostrados os resultados obtidos com cada algoritmo.

- No capítulo 7 serão apresentadas as conclusões do estudo de caso realizado através do método proposto.

Este capítulo apresentou um breve resumo sobre o cenário atual no setor de distribuição de energia elétrica no Brasil, salientando o aspecto de que os dados dessas companhias não têm um aproveitamento adequado para uma detecção de fraude. Também foi apresentado o problema-foco dessa dissertação, que é o roubo de energia elétrica, sendo que em média 15% da energia gerada no país é furtada.

A partir destes fatos, verificou-se a necessidade da criação de uma metodologia com base nos dados da concessionária, aplicando métodos para fazer uma Mineração de Dados e objetivando a detecção de fraudes, com a posterior aplicação destas técnicas no estudo de caso e apresentação dos resultados na conclusão desta dissertação.

Capítulo 2

2 Fundamentação Teórica

2.1 Descoberta de Conhecimento em Base de Dados

A descoberta de conhecimento é um processo, que tem a finalidade de desvendar padrões, a partir de modelos válidos que servirão para a criação de novos arquétipos. Estes terão a utilidade de melhorar a compreensão de um problema para uma tomada de decisão, neste caso por parte da concessionária de energia elétrica [FAYYAD, 1996].

A análise bibliográfica revela que existem muitos artigos, trabalhos e dissertações relacionadas com o tema, especialmente em Inteligência de Negócios (*BI – Business Intelligence*). Este engloba vários conceitos, dentre eles Banco de Dados, *Data Warehouse*, Mineração de Dados, e outros, na aplicação de técnicas baseadas em inteligência artificial (lógica fuzzy, redes neurais) para a tipificação, classificação de clientes e detecção de fraudes, baseados em dados de consumo e faturamento. Também se observa que existem já pesquisas divulgadas em nível internacional, nas quais a técnica de Mineração de Dados é aplicada para descoberta de conhecimento utilizando banco de dados de sistemas de energia.

Em CABRAL (2004) *apud* PATRICIO, foi desenvolvido um método para que sejam identificados os consumidores fraudadores, ou seja, consumidores que roubam energia elétrica, ligados em baixa tensão e pertencentes às diversas classes de energia. As concessionárias têm a necessidade de que sejam feitas inspeções técnicas em campo, uma vez que as mesmas auxiliarão na seleção das unidades consumidoras a serem inspecionadas e, desta forma, se tem uma melhora nos resultados da detecção de fraude nessas classes. A procura de um fraudador é uma busca lenta e de custo elevado para a concessionária. O trabalho mencionado propõe uma metodologia com base em Rough Sets[3] para detecção de

[3] Introduzida em 1982 por Zdzislaw Pawlak, está teoria é conhecida por possuir propriedades permitem que sejam eliminadas variáveis ou atributos que não tem relevância para um determinado propósito, onde são divididos os conjuntos de atributos e os mesmos tem a mesma capacidade de manter as propriedades, a eliminação de atributos irrelevantes é o procedimento que caracteriza essa teoria [CABRAL et al, 2004].

fraudes utilizando dados históricos de consumidores. Assim a metodologia baseada em Rough Sets tem a função de identificar padrões de comportamento fraudulentos no banco de dados das concessionárias. Com a obtenção destes padrões, geram-se regras de classificação que, em novas inspeções, indicarão que clientes potencialmente apresentam perfis fraudulentos. A inspeção conduzida por comportamentos suspeitos, aumenta significativamente a taxa de acerto melhorando também a qualidade de fraudes detectadas, tendo como conseqüência uma diminuição das perdas comerciais.

Pelo que foi pesquisado nenhuma concessionária de energia do Rio Grande do Sul possui alguma forma automatizada para esta tarefa.

No trabalho de REIS (2004) *apud* PATRICIO apresenta-se um sistema (software) que faz uma pré-seleção de consumidores a fim de que se proceda a uma inspeção. O objetivo é detectar desvios e erros de medição utilizando uma árvore de decisão. O ponto de partida são os dados de uma empresa distribuidora de energia, os quais foram selecionados como mostra a Tabela 1:

Tabela 1 – Tabelas de amostras

Amostra	
Cinco atributos	**dentre os 52 disponíveis**
40.000 clientes	**de um total de 600.000**

Foi feito um treinamento de um sistema desenvolvido pelos pesquisadores e se obteve os seguintes resultados, mostrados na Tabela 2:

Tabela 2 – Tabela de resultados

Número de Clientes (selecionados aleatoriamente)	
Resultados obtidos (taxa de acerto de 45% para fraude) pelo programa desenvolvido pelos pesquisadores.	**A taxa alcançada pela concessionária utilizando seus métodos chegou ser 35% a menos que a taxa do sistema.**

Observando estes índices, verifica-se então que os pesquisadores obtiveram um resultado mais eficaz, com maior sucesso do que as medidas utilizadas pela concessionária.

Com o intuito de facilitar o entendimento deste estudo e contribuir para a compreensão desta dissertação e um pouco sobre análise de sistemas, serão apresentados alguns conceitos que são requisitos básicos para o desenvolvimento do trabalho proposto.

2.2 Inteligência de Negócios (*Business Intelligence - BI*)

No cenário atual as empresas, sejam elas de qualquer segmento, estão enfrentado um mercado que a cada ano que passa se torna mais competitivo, exigindo com este ritmo acelerado soluções mais eficientes (sistemas, banco de dados, equipamentos, métodos), que possam oferecer maiores vantagens e auxiliar assim as empresas de grande porte, principalmente no enfrentamento dos desafios que surgem constantemente. Para que exista uma maior cooperação entre os diversos setores de uma empresa surge o *BI*, método e tecnologia que engloba vários outros métodos, esse método busca soluções e estratégias para [REZENDE et tal, 2003]:

a) melhor entendimento dos segmentos de atuação da empresa no mercado;

b) promover uma melhor competência na essência da empresa;

c) poder de identificar oportunidades não destacadas anteriormente ;

d) responder às mudanças bruscas do mercado maneira uma maneira mais adequada e eficiente;

e) promover um bom relacionamento entre clientes e fornecedores;

f) diminuir significativamente os custos operacionais da empresa.

No contexto desta dissertação estão envolvidos conceitos de Inteligência de Negócios, pois a Mineração de Dados é uma das metodologias dentre as várias outras assim como *Data Warehouse*, Banco de Dados, Estatística, que fazem parte deste processo.

Segundo CABENA (1998) *apud* PETERMANN o *BI* varia de uma planilha simples de custos mensais até um complexo e avançado sistema corporativo (exemplo um sistema ERP -

Enterprise Resource Planning ou Sistemas Integrados de Gestão Empresarial). Em um processo de BI a Mineração de dados é um importante, senão o principal, componente deste processo.

A Figura 3 a seguir mostra diversas metodologias de suporte a *BI*.

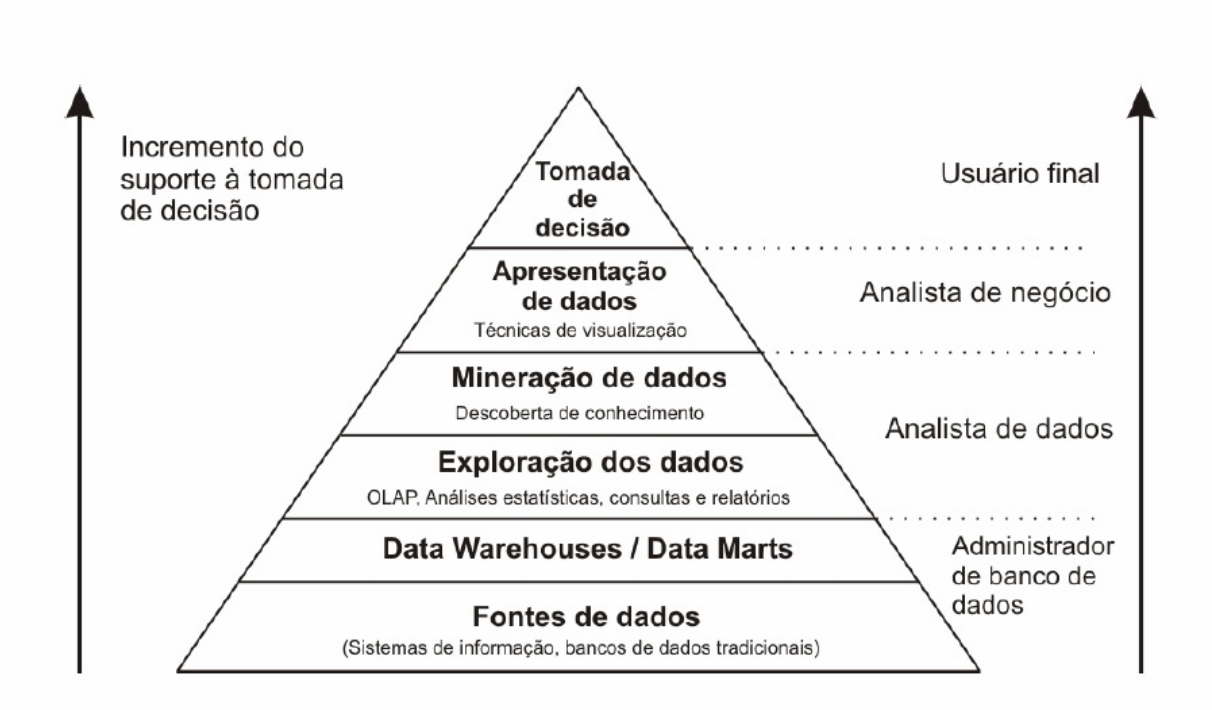


Figura 3 - Mineração e Dados e Inteligência de Negócios, Fonte: CABENA. 1998.

De uma maneira abrangente, o conceito de Inteligência de Negócio pode ser entendido como a utilização de um conjunto de métodos ou informações que auxiliam a definição de estratégias que aumentem a competitividade no mercado e nos negócios de uma empresa. O objetivo maior da Inteligência de Negócio é procurar definir regras e técnicas para estruturação mais adequada dos grandes volumes de dados, tendo como finalidade transformar a informação em um repositório ou depósito mais estruturado, ou seja, um Banco de Dados, independente das informações que o originaram [REZENDE et tal, 2003].

2.3 *Data Marts*

A criação de um *Data Warehouse* não é uma tarefa nada trivial, principalmente se seu objetivo for atender uma organização como um todo. Então as empresas optam pela criação

de *Data Marts*, que comparados ao *Data Warehouse* são significativamente mais compactos e com uma acessibilidade maior de seus dados [HAN, 2001].

Em organizações que já possuem seu *Data Warehouse* consolidado, operando corretamente, os *Data Marts* são ferramentas úteis para resolução de tarefas mais específicas. Um exemplo prático seria uma análise individualizada de dados das curvas típicas dos clientes, que implica em um número menor de tabelas [PETERMANN, 2006].

2.4 *Data Warehousing*

Data Warehousing não é considerado um produto, é mais um processo no qual podemos construir e gerenciar um banco de dados a partir de várias fontes de dados, cujo resultado que teremos a partir desse processo é o *Data Warehouse* [GARDNER, 1998].

Então o *Data Warehousing* nada mais é do que o processo de construção do *Data Warehouse*.

2.5 *Data Warehouse*

O *Data Warehouse* (DW) é um banco de dados de grande porte, resultado da junção de vários sistemas de banco de dados ou técnicas que, aplicadas em conjunto, servirão para geração de um sistema de dados. Tem o objetivo de fornecer o suporte na criação de relatórios, visando gerar informações que auxiliem nas tomadas de decisões de uma empresa. O funcionamento de um *DW* fundamenta-se numa arquitetura cliente/servidor (o banco de dados é instalado e montado em um servidor e nos micros clientes somente instalado o versão cliente do banco, para acesso do mesmo). A princípio, todos os bancos de dados de grande porte são gerenciadores de *Data Warehouse* [COREY, 2001].

Para a construção de um *Data Warehouse* deve ser levado em consideração o projeto da interface com o sistema operacional e o projeto do próprio *DW*. Estes projetos não descrevem exatamente o que acontece na construção do mesmo, pois é construído de modo heurístico (é uma pesquisa realizada por meio da quantificação de proximidade a um determinado objetivo). Primeiramente povoa-se o *Data Warehouse* com alguns dados. Esses dados passam por um analista de sistemas que irá examiná-los para validação. Após, será levado em consideração o *feedback* (resposta por parte do usuário final do *DW*), o qual tem continuidade por toda vida do *Data Warehouse.* Alguns dados serão adicionados; outros, por

sua vez, serão modificados. De acordo com o uso, as necessidades e os problemas vão sendo detectados pelos usuários, que normalmente tem algo a acrescentar ou sugerir, visando melhorar a utilização do mesmo [INMON, 1997].

O *Data Warehouse* possui técnicas para análise de dados. Estas técnicas são consultas *SQL (Structured Query Language* - Linguagem de Consulta Estruturada) ou algum mecanismo de visualização de dados, como ferramentas *OLAP (Online Analytical Processing)*. Desta forma, a análise de dados é importante para tomada de decisão, podendo ser expressas como [REZENDE et tal, 2003]:

a) Quais são os clientes que possuem um potencial para praticar desvios de comportamento?

b) Em que região encontra-se maior ocorrência de desvio?

c) Quais clientes gostariam de pedir um aumento de carga de energia?

A Mineração de Dados auxilia na busca de padrões escondidos nos dados de uma maneira inteligente, pois de outra maneira ficaria impossível que o usuário pudesse encontrar todas as relações e associações existentes em grandes volumes de dados. Por isso, faz-se necessário o uso da Mineração de Dados, que será abordado a seguir, no tópico 2.9 [REZENDE et tal, 2003].

Os passos que foram realizados forma a análise, normalização e seleção de dados e a criação do diagrama de entidade de relacionamentos para a criação do banco de dados para executar as tarefas pertinentes à dissertação, ou seja, um novo *Data Warehouse,* denominado de PROPUSDM (Propus é o nome do servidor de dados do Grupo de Planejamento Integrado de Recursos Energéticos (GPIRE) e DM uma referencia a *Data Mining*), com tabelas, dados e atributos relevantes para aplicação das técnicas de Mineração de Dados.

2.6 *Knowledge Discovery in Database (Descoberta de Conhecimento na base de dados)*

A Mineração de Dados faz parte de um processo maior de tratamento de informações chamado *Knowledge Discovery in Database (KDD* – Descoberta de Conhecimento na base de dados), que consiste primordialmente em estruturar de uma melhor forma um banco de dados.

Atualmente na literatura há duas opiniões divergentes entre a Mineração de Dados e o *KDD*. Autores assim como Fayyad et. (1996) consideram que *KDD* e Mineração são sinônimos. Outros autores levam em consideração que a Mineração de Dados é uma das etapas do processo de *KDD*, sendo que esta é considerada a principal etapa do referido processo [REZENDE et tal, 2003]

O processo de *KDD* possui cinco etapas distintas, ilustradas na Figura 4, quais sejam: a seleção, a preparação ou pré-processamento dos dados, a transformação, a mineração e a interpretação ou análise de dados [CABENA 1998].

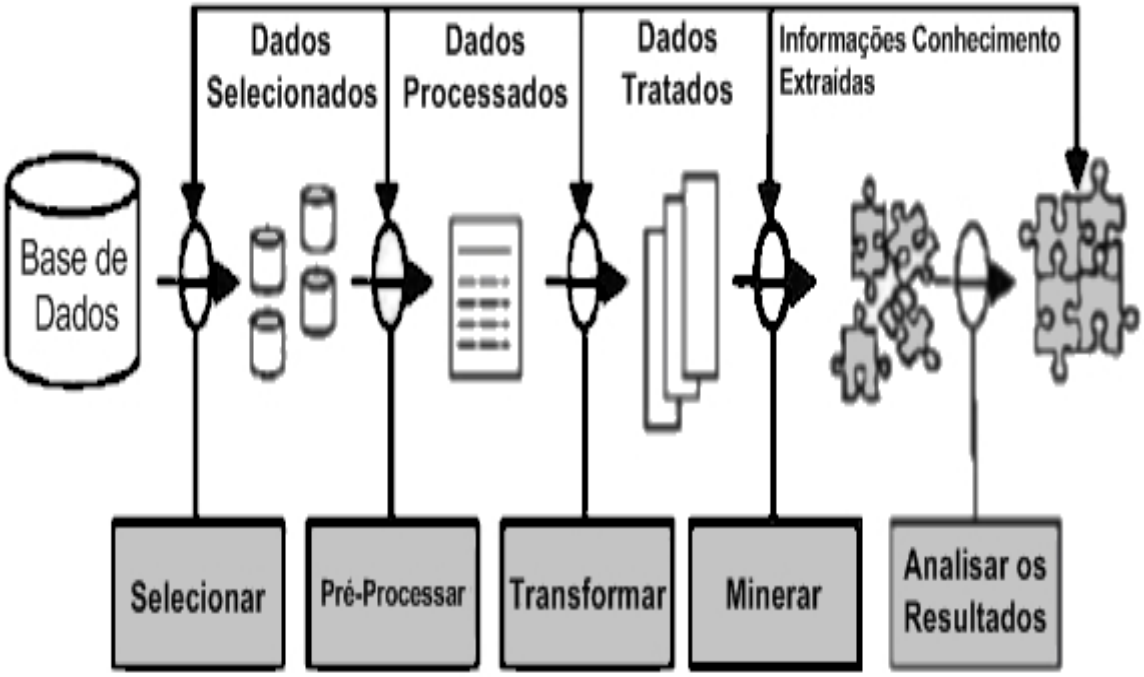


Figura 4 - Etapas do processo KDD [CABENA1998].

- **Seleção / Escolha dos Dados (*Data Selection*):** etapa na qual ocorre a identificação e onde as informações serão escolhidas e depois inseridas no processo de *KDD*. Nesta fase se define quais dados devem entrar, assim como são realizadas as tarefas de limpeza e filtragem dos dados, corrigindo possíveis valores nulos, duplicados ou inconsistentes.

- **Preparação dos Dados (*Data Preprocessing*):** etapa em que os dados passam por uma análise, ou seja, os dados são preparados: Neste momento os dados são selecionados, havendo a correção de dados ruidosos (etapa do processo de limpeza, responsável pela remoção de discrepâncias/impurezas contidas nos dados que serão analisados). Etapa também conhecida como *Data Cleaning* (Limpeza dos Dados), onde são tratados os valores ausentes e, caso necessário, também a remoção de campos irrelevantes para o processo.

- **Transformação dos Dados (*Data Transformation*):** esta etapa tem como objetivo facilitar a análise através da transformação dos dados. Deve ser feito um modelo de dados analíticos que represente a consolidação, integração e reestruturação dos dados selecionados e processados pelas etapas anteriores. Após a definição do modelo, os dados serão submetidos a um refinamento para que então sejam utilizados pelos algoritmos de Mineração de Dados.

- **Mineração dos Dados (*Data Mining*):** como salientado anteriormente, esta é considerada a principal etapa do processo *KDD*. É onde os padrões de comportamento são descobertos. Primeiramente deve-se definir qual o objetivo da Mineração de Dados e quais algoritmos são mais apropriados para a tarefa. Ex: sumarização, classificação, regressão, associação, agrupamento.

- **Interpretação / Análise dos Dados (*Data Interpretation*):** etapa onde os resultados (conhecimentos) são analisados. Engloba a interpretação dos padrões descobertos, podendo ocorrer um possível retorno de alguma etapa anterior.

Tendo início no final dos anos oitenta, a Mineração de Dados tem como objetivo a extração de conhecimento em grandes volumes de dados. Possuindo uma característica multidisciplinar, dada pelas relações existentes com as mais diversas áreas como banco de dados, aprendizado de máquina *(machine learning*), estatística, recuperação de informação, computação paralela e distribuída [CARVALHO, 2001].

A relação entre Mineração de Dados e *KDD* pode ser visualizada através da Figura 5:

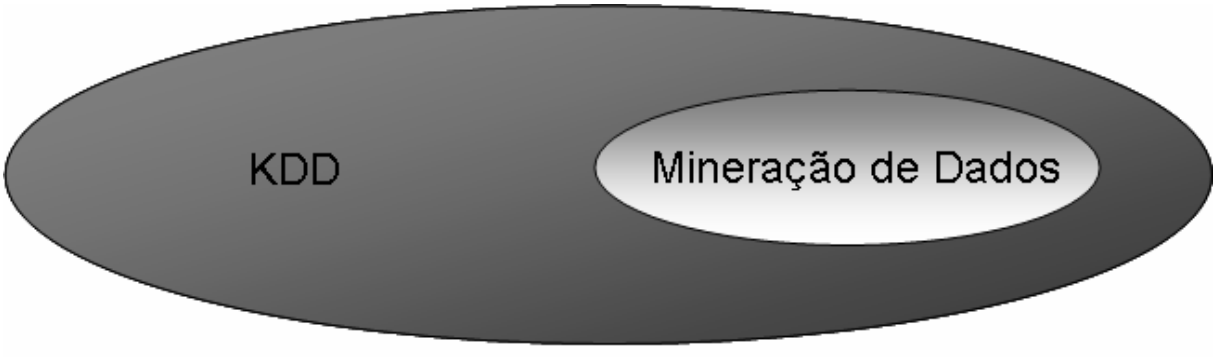


Figura 5 - Relação entre KDD e Mineração de Dados.

Figura 5 baseada no trabalho de SILVA (2005).

2.7 Mineração de Dados

A Mineração de Dados surgiu nos anos 80 e era voltada a resolver unicamente uma tarefa. Era uma ferramenta de análise com um único propósito e não tinha suporte ou auxílio das demais etapas do processo de *KDD* [FAYYAD 1996].

A Mineração de Dados, como foi mencionado anteriormente, é responsável por localizar padrões (sejam eles de comportamento, consumo, faturamento, etc.) nos dados de um banco de dados ou *DW* de uma forma particular. É o processo responsável pela seleção de métodos e ajustes de parâmetros nos algoritmos escolhidos, tendo em vista o melhor resultado na tarefa em questão. Sendo o núcleo ou a principal etapa do processo de *KDD* o uso de técnicas de exploração em grandes bancos de dados, tem-se como objetivo a descoberta de padrões e relações entre os dados, os quais não teriam como ser descobertos a olho nu ou pelas técnicas tradicionais [CARVALHO, 2001].

O processo de Mineração de Dados consiste na aplicação de várias áreas, técnicas e conceitos, assim como sistemas de banco de dados, estatística, inteligência artificial, aprendizado de máquina, todos aplicados em conjunto ou individualmente a um grande volume de dados, conforme a Figura 6. Tal procedimento sempre focado no objetivo, que é o de encontrar padrões e tendências para apoio à tomada de decisão [SILVA, 2005].

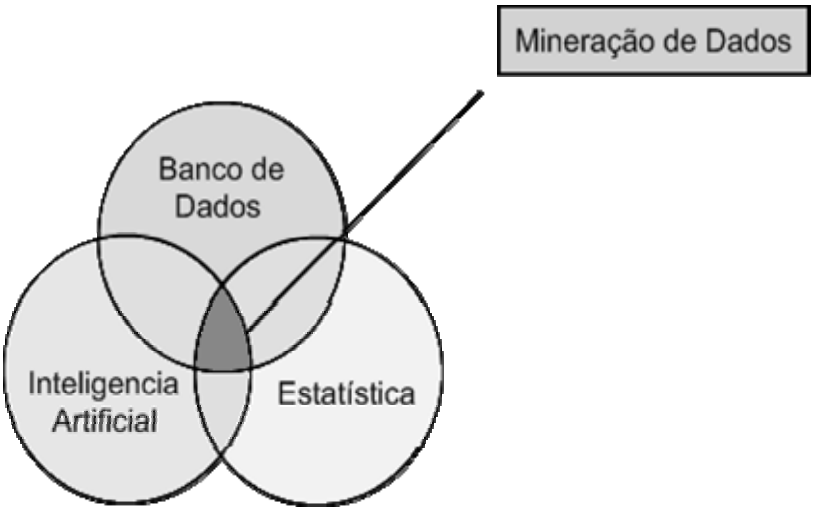


Figura 6 - Mineração utilizando recursos de diferentes áreas

Figura 6 baseada no trabalho de SILVA (2005).

Segundo WILEY (2006) a Mineração dos Dados será um dos desenvolvimentos mais revolucionários da próxima década, noticiado no site ZDNET de tecnologia (fevereiro 8, 2001). No mesmo, o MIT (Massachusetts Institute of Technology) escolheu a Mineração de Dados como uma das 10 tecnologias mais emergentes que deverão mudar o mundo.

2.7.1 Fundamentação da Mineração de Dados

Assim, a Mineração de Dados consiste na intersecção de várias áreas, como a estatística, a inteligência artificial e a aprendizagem de máquina; e banco de dados, que foi descrito anteriormente em uma seção separada. Nesta seção veremos outros métodos, cada um deles individualmente, destacando qual é o seu papel dentro da Mineração de Dados.

2.7.1.1 Estatística

A Mineração de Dados descende de três linhas fundamentais, sendo a mais antiga a Estatística Clássica, que foi a primeira responsável pela existência da MD. A estatística envolve vários conceitos que aplicaremos no trabalho, como cálculos probabilísticos, médias, análise de conjuntos, intervalos aplicados nos dados e seus relacionamentos [PETERMANN, 2006].

Deste modo, a estatística e a probabilidade serão as formas em que os resultados serão apresentados. Como exemplo temos uma estatística de classe de consumo, clientes, etc. que praticam desvio ou têm probabilidade de praticá-los.

2.7.1.2 Inteligência Artificial

A Inteligência Artificial, também conhecida como IA, é considerada a segunda linha ou linhagem da Mineração de Dados. A IA veio em oposição à estatística, pois foi construída a partir da heurística. Ela basicamente surgiu da idéia de imitar o pensamento humano na resolução de problemas do seu cotidiano, reproduzir a resolução de problemas estatísticos, por exemplo. [PETERMANN, 2006].

Segundo CISTER (2005) *apud* PETERMANN, a Inteligência Artificial requer um impressionante poder de processamento computacional, máquinas mais robustas. Esse processamento era inviável até os anos 80 (mesmo que IA seja um conceito muito mais antigo). Nessa época teve início a fabricação de computadores com um melhor processamento e por valores mais acessíveis, tornando este conceito cada vez mais popular nas décadas seguintes.

A Inteligência Artificial faz uso de algoritmos genéricos para predição[4], classificação, entre outras tarefas, o que será visto no Capítulo 4.

2.7.1.3 Aprendizado de Máquina

O Aprendizado de Máquina *(Machine Learning)* é a terceira e última linha da MD. Trata da junção das duas linhas anteriores à estatística e a IA [PETERMANN, 2006]..

Segundo COUTINHO (2003) *apud* PETERMANN, a Mineração de Dados adapta o aprendizado de máquina e suas técnicas para aplicações cientificas e de negócios. A MD utiliza-se dessas técnicas que são usadas em conjunto para estudar, analisar os dados em um banco e achar padrões e tendências, pois seria difícil encontrá-los de outra forma.

---

[4] Segundo SILVA (2005) os algoritmos são considerados preditivos, tem a função na mineração de descobrir, fornecer previsões e tendências de valores futuros ou desconhecidos de um ou mais atributos de um banco de dados a partir de um valor conhecidos ou disponíveis.

A utilização do Aprendizado de Máquina facilita a localização das tendências que apontam para ocorrência de desvio de comportamento. Levando em conta a quantidade considerável de registros armazenados no Banco de Dados do faturamento da concessionária, tendo em vista ainda que a amostra estudada consiste em uma pequena fração da totalidade das informações do Banco, a utilização do Aprendizado de Máquina é indispensável.

2.7.2 Técnicas de Mineração de Dados

Na MD não há uma única técnica que resolva todos os problemas, pois existem cinco diferentes métodos que podem ser utilizados para muitos objetivos. Cada método possui seus algoritmos específicos, com suas peculiaridades, vantagens e desvantagens. Dependendo do objetivo a ser alcançado, a escolha da técnica pode variar entre Classificação, Predição, Regressão, Segmentação *(Clustering),* Associação e Padrões Seqüenciais [GOEBEL e GRUENWALD, 1999]:

A Classificação é a tarefa que consiste em prover ordenação de um grande volume de dados em pequenos grupos, subgrupos ou classes mais compreensíveis. Esta técnica tem como objetivo melhorar o entendimento ou o controle de uma situação [KLÖSGEN e ZYTKOW, 2002]. Como exemplo pode-se citar classificar clientes com mais ou menos chances de inadimplência.

A Predição é a tarefa que tem função de fazer a avaliação do valor futuro ou desconhecido de um atributo, tendo como base os dados conhecidos que descrevem o comportamento passado (dados histórico) deste atributo [SILVA, 2005]. Por exemplo, considerar históricos de consumo de um determinado grupo de clientes versus medições advindas de um transformador ou alimentador, podendo prever ações de furto de energia elétrica.

A tarefa da Regressão consiste na análise da interdependência que existe entre os valores dos atributos de uma mesma tabela, produzindo automaticamente um modelo ou função que, a partir de valores existentes, pode determinar os valores para novos atributos [GOEBEL e GRUENWALD, 1999]. Exemplo: considerando um determinado padrão de consumo de uma classe específica de clientes pode-se prever a probabilidade de um desvio.

A Segmentação é o processo que tem como função dividir uma população ou amostra heterogênea de dados em vários subgrupos ou clusters mais homogêneos [HARRISON, 1998]. Por exemplo, agrupar clientes com comportamento de consumo de energia similar.

Associação é o método que determina quais ocorrências de dados implicam na existência de outras em uma mesma transação (consulta SQL) [CABENA, 1998]. Por exemplo, um grupo de clientes moradores de uma região socialmente desfavorecida, ligados a um alimentador que possuem o hábito de praticar desvio.

Padrões Seqüenciais, como o nome já diz, são responsáveis por detectar padrões de seqüência entre transações; como um conjunto de itens encontra-se freqüentemente acompanhado durante um período de tempo por outro conjunto de itens [CABENA, 1998]. Por exemplo, determinar ao longo do tempo a sazonalidade e uma determinada atividade econômica, tento em vista a análise do comportamento de consumo.

2.7.3 Passos Fundamentais da Mineração

Na Figura 7 são apresentados os passos fundamentais para que uma Mineração seja bem sucedida [NAVEGA, 2007].

A Mineração de Dados parte geralmente de diversas fontes de dados, os quais são provenientes desses bancos, passam por uma limpeza e logo após se constrói um Banco de Dados com uma estrutura mais adequada para uma boa mineração de dados. Depois da construção do DB e da inclusão dos dados no mesmo, alguns dados relevantes são selecionados em função do objetivo. Posteriormente, os dados então passam pela MD, onde é utilizado algum *software* de Mineração e seus algoritmos de classificação, predição, regressão, segmentação, associação e padrões seqüenciais. A seguir se tem a avaliação, visualização dos resultados e a extração de conhecimento, para que possam ser utilizados nas tomadas de decisões. Vale destacar que o processo de Mineração tem início desde a primeira etapa do processo.

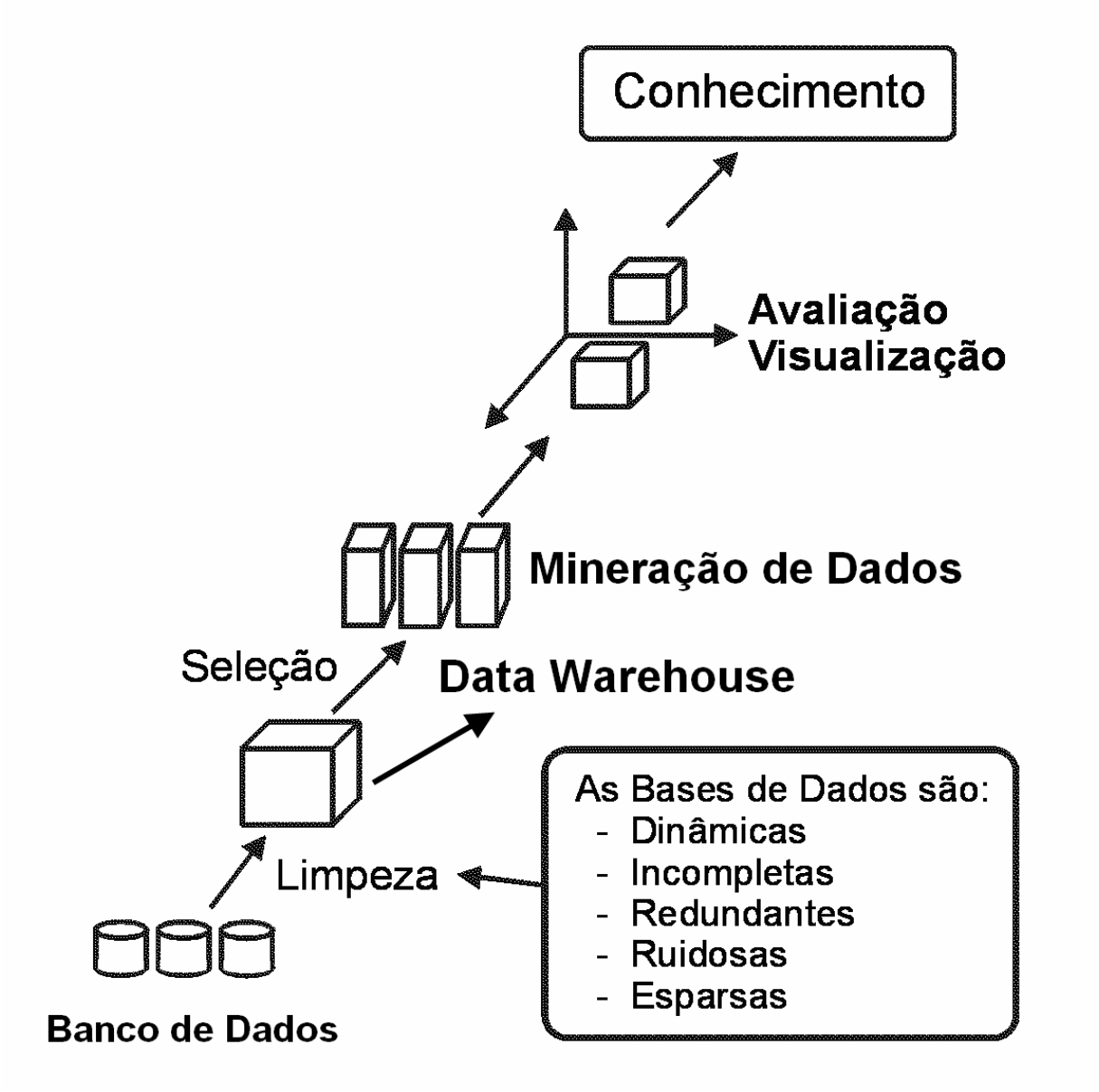


Figura 7 - Processo de Descoberta de Conhecimento em Base de Dados

Origem da figura: site <http://www.intelliwise.com/reports/i2002.htm>

2.8 Pacotes Computacionais para Mineração de Dados

A seguir, para conhecimento, são apresentados alguns softwares de Mineração de Dados mais usados no mercado:

Clementine

É o programa de mineração de dados do SPSS. Esta solução utiliza arquitetura aberta que interage com uma ampla variedade de fontes de dados e sistemas de informação. Isto significa que você pode acessar qualquer tipo de dados e fornecer previsões e recomendações. O site pode ser acessado no endereço: <http://www.spss.com/spssbi/clemetine/>.

DBMiner

Desenvolvido pela companhia com mesmo nome. Conhecido como visual *Data Mining* por utilizar recursos de computação gráfica (CG) para evidenciar padrões em bases de dados.

Disponível no site: <http://www.dbminer.com/>.

DB2 Intelligent Miner fro Data

Desenvolvido pela IBM, é uma ferramenta de mineração interligada diretamente com o banco de dados DB2 da IBM.

Disponível no site: <http://www-3.ibm.com/software/data/imine/fordata/>.

Enterprise Miner

Os usuários do SAS têm no *Enterprise Miner* sua opção para Mineração de Dados tradicionalmente utilizado na área de negócios, marketing e inteligência competitiva.

Disponível no site <http://www.sas.com/products/miner/index.html>.

Microsoft SQL Server 2000 Analysis Service

A Microsoft, dentre várias soluções para inteligência empresarial, oferece uma em Mineração de Dados: <http://www.microsoft.com/office/business/intelligence/default.asp>.

Oracle Data Miner

Desenvolvido pela Oracle, permite interligação direta com o banco de dados Oracle Enterprise 9i e 10i.

Encontra-se no endereço:

<http://www.oracle.com/solutions/business_intelligence/data-mining.html>.

***Statistica Data Miner*:**

Acrescenta as facilidades de MD ao tradicional pacote utilizado em aplicações de estatística. A solução da STARTSOFT inclui as funções típicas para a mineração de dados, a qual pode ser encontrada em: <http://www.startsoft.com>.

Weka (Waikato Environment for Knowledge Analysis ou Ambiente Waikato para a Análise do Conhecimento)

Software de domínio público desenvolvido em Java, pela Universidade de Waikato, Nova Zelândia, contém uma série de algoritmos de MD, por exemplo, *Fuzzy C-means, K-means,* entre outros. Escolhido dentre os tantos pacotes oferecido por ser um software gratuito, a equipe de desenvolvimento tem lançado periodicamente correções e atualizações do software, além de manter a lista de discussões acerca da ferramenta. Grande parte de seus componentes de software são resultantes de teses de dissertações de grupos de pesquisa desta universidade. Inicialmente, o desenvolvimento de software visava à investigação de técnicas e aprendizado de máquina, enquanto sua aplicação inicial foi direcionada para agricultura, área chave na economia da Nova Zelândia.

Possui uma interface gráfica amigável, dispõe de algoritmos que foram selecionados para a execução do método desenvolvido. Estes algoritmos fornecem relatórios com dados analíticos, estatísticos, minerados e também interage bem com o banco utilizado.

Também possui disponível uma abrangente documentação on-line do código fonte. Por ter o código em Java, este pode ser rodado em várias plataformas. Uma limitação da ferramenta é a fixação do volume de dados a ser manipulado.

O pacote encontra-se gratuito no endereço: <http://www.cs.waikato.ac.nz/ml/weka/>.

2.9 Escolha de algoritmos

A escolha é realizada conforme os padrões e resultados a serem encontrados, ou seja, conforme o foco ou objetivo da MD. Podem ser utilizados: Árvores de Decisão, Algoritmos

Estatísticos, Algoritmos Genéticos, Regras de Decisão, Redes Neurais Artificiais, Redes Bayesianas e Lógica Fuzzy [REZENDE et tal, 2003].

Alguns pesquisadores como CARDOSO (2003) em seus experimentos mostrou que não se utiliza um único bom algoritmo para todas as tarefas de Mineração de Dados. É aconselhada a utilização de mais de um algoritmo para a realização de uma tarefa a ser executada, até porque muitos exigirão um processamento enorme, o que poderá inviabilizar a mineração. Com a utilização de mais de um algoritmo se obtêm diversos modelos, com a possibilidade de fazer uma comparação entre eles e ver qual teve um melhor desempenho.

Nesta dissertação, o mais recomendado para a detecção de desvio de comportamento, segundo algumas pesquisas feitas em trabalhos com propósitos similares, seriam as Redes Neurais Artificiais como a MLP (MultiLayer Perceptron). Testes feitos com a rede MLP constataram que a mesma leva mais tempo de resposta, pois o custo computacional é muito elevado. Outras técnicas recomendadas são algoritmos de associação como o Apriori, algoritmos de classificação como Árvores de Decisão e também alguns algoritmos de agrupamento como Classificadores Bayesianos. Todos estes serão utilizados e posteriormente comparados, levando em consideração resultados, eficácia e tempo.

Os algoritmos em questão são encontrados na maioria dos softwares de Mineração de dados; neste caso, o Weka é usado para a obtenção dos resultados. O SGBD (Sistema de Gerenciamento de Banco de Dados) usado para a criação do Banco de Dados foi o mesmo utilizado pela concessionária.

Este capítulo foi elaborado para apresentar e descrever os conceitos que foram utilizados na dissertação, tendo em vista que para um bom trabalho de Mineração de Dados é imprescindível o uso das técnicas e ferramentas referidas.

O capítulo enfatizou a descoberta de conhecimento em base de dados, com o cruzamento de informações dos bancos de Faturamento e Consumo, desde a criação de uma DER, o entendimento de banco de dados, para a criação de um novo banco, usando métodos de geração de um *Data Warehouse*, passando por um conceito que engloba várias metodologias, entre elas MD, *DW*, BD e *BI* (Inteligência de Negócios).

Mostraram-se as etapas do *KDD*, como Seleção, Limpeza, Transformação e Análise de Dados da qual a Mineração de Dados faz parte, sendo ela a principal etapa do *KDD*. Foi apresentado, ainda, um histórico e as principais técnicas para correção, classificação, segmentação predição, associação, avaliação, preparação dos dados, que serão utilizadas posteriormente para Mineração de Dados.

Neste capítulo também foram apresentados os princípios fundamentais, tanto da Mineração como da Estatística, da Inteligência Artificial e do Aprendizado de Máquina, até as mais modernas ferramentas de Mineração e a escolha dos algoritmos existentes nas mesmas.

Capítulo 3 - Principais Técnicas de descoberta de Padrões

3 Aplicação de Técnicas e Descobertas de Padrões para Detecção e Desvio de Comportamento

Os Bancos de Dados são altamente suscetíveis a dados que apresentam erros, dados incompletos ou valores ausentes e dados inconsistentes. Isto decorre dos grandes volumes de dados, que muitas vezes são inseridos manualmente, acarretando em erros de digitação, ou ainda pelos erros nas medições, estas automatizadas. Com vistas a essas dificuldades, são aplicadas técnicas de pré-processamento e transformação de dados, para aumentar então a qualidade desses dados a serem minerados. Em algumas bibliografias relata-se que a fase de pré-processamento e de transformação de dados tende a consumir aproximadamente 70% do processo de *KDD*. Já a etapa Mineração é responsável pela seleção de métodos que serão utilizados para localizar padrões. Neste caso, será selecionado os métodos para a descoberta de padrão de consumo de energia, além do ajuste de parâmetros dos algoritmos na ferramenta de mineração. Pressupõe então que nesta etapa os dados tenham sido limpos, normalizados, etc. [SILVA, 2005].

A seguir abordam-se as técnicas para descoberta de padrões em grandes quantidades de dados.

3.1 Problema

Na década de 90 o *KDD* foi criado para designar conjunto de processos, técnicas e abordagens que proporcionam o contexto no qual a Mineração entrará. É a aplicação de métodos científicos modernos aos problemas do mundo dos negócios. Portanto, é preciso estar ciente de que o processo de descoberta de padrões não se faz através de hipóteses, mas sim de evidências e explicações sobre ela, podendo eventualmente levar a construção de um modelo [BRAGA, 2005].

A partir dessas evidências pode-se então construir um modelo para aplicação das técnicas e etapas pertinentes do *KDD* para então assim solucionarmos o problema, que neste caso é o desvio de comportamento de uso de energia.

3.2 Definição da População

Levando em consideração o conteúdo dos dados encontrados no banco, os mesmos podem ser classificados em três categorias distintas: demográfico, comportamental e psicológico (valores) [BRAGA, 2005]. Por exemplo, um cliente masculino de 35 anos, solteiro, consome em média 75 kW.h mensais de energia elétrica e acha que kW.h faturado em sua residência não condiz com o consumido. A Tabela 3 mostra essa classificação.

Tabela 3 – Tipos de Dados

Tipos de Dados		
Demográficos	**Comportamental**	**Psicológico (Perceptivo)**
Homem, 35 anos, solteiro.	**Consome mensalmente 75kW.h de energia elétrica**	**Acha que kW.h faturado não condiz com o consumido**

Nas empresas de distribuição de energia elétrica são muitas as fontes de dados que podem ser utilizadas em um projeto de Mineração de Dados: Banco de dados de clientes, por exemplo:

Tabela 4 – Exemplo da tabela de clientes

Tabela Clientes				
ID_CLIENTE	**LOGRADOURO**	**NUMERO**	**CEP**	**COD_MUNIC**

Banco de faturamento exemplificado na Tabela 5:

Tabela 5 – Exemplo da tabela de faturamento

Tabela de Faturamento				
NUC	**DT_FAT**	**TARIFA_CONT**	**CLASSE**	**TENSAO**

Conforme o objetivo da modelagem, ou finalidade da Mineração, o conjunto de dados que será utilizado deve mudar, sendo possível destacar a finalidade de combate a desvio.

Assim, por meio dos dados selecionados, define-se a população apropriada a ser inserida no *Data Warehouse* [BRAGA, 2005].

3.3 Amostragem

A quantidade de registros contidos em um Banco de Dados geralmente são infindáveis, não sendo preciso a utilização na sua totalidade para a criação de um modelo. Desse modo, é selecionada uma quantidade de dados que seja necessária para a sua amostra, utilizando-se para isto de técnicas específicas de amostragem. Nesse contexto, duas perguntas devem ser feitas: Qual o tamanho da amostra? Como selecioná-la? O tamanho depende da finalidade do modelo, número de parâmetros e poder de predição. Para estimação de proporções tem-se um exemplo na Tabela 6, explicitando que quanto maior a quantidade de amostras, menor a margem de erro [BRAGA, 2005].

O tamanho da amostra não só diminui a margem de erro, como também é importante na Mineração ter um número significativo de amostras, para que seja feita a seleção de atributos pelo *software* de mineração utilizado, assim como o processo de treinamento dos algoritmos, já que com uma amostra pequena fica muito difícil de obter resultados consistentes de uma Mineração.

* Nível de confiança.

Tabela 6 – Tabela de Amostragens

Tamanho da amostra (Universo)	100	200	400	750	1000	1500	3000	5000
Margem de Erro (λ*=99%)	13	9	7	5	4	3	2	2
Margem de Erro (λ*=95%)	10	7	5	4	3	3	2	1
Margem de Erro (λ*=90%)	8	6	4	3	3	2	2	1

Fator que deve ser levado em conta é a escolha dos elementos da amostra. Segundo o autor, são conhecidos cinco tipos de amostragem: aleatória simples, aleatória estratificada,

sistemática, por múltiplos estágios e por cotas. Caso a população seja bem homogênea, o mais adequado é amostragem aleatória simples. No entanto, se a população for segmentada, é recomendada à utilização das opções de amostra aleatória estratificada e amostra de cotas. [BRAGA, 2005].

Os tipos de amostragem, segundo BRAGA (2005) são cinco, especificadas logo a seguir:

Aleatória simples – é o tipo de amostragem que é seleciona por sorteio, de tal forma que cada unidade da população tenha igual chance de ser sorteada.

Aleatória estratificada – é a selecionada através de um sorteio de um subconjunto ou de um estrato da população.

Sistemática – é baseada na aleatória simples: são embaralhados os elementos da população e passa-se a selecioná-los a cada n/N elementos, onde temos n: tamanho da população e N: tamanho da amostra.

Múltiplos estágios – a amostra cuja população é representada por hierarquia campo de uma tabela, por exemplo, do mais amplo para o mais específico, desta forma: MUNICIPIO, BAIRRO, LOGRADOURO, NUMERO, NOME.

Cotas – a população é dividida em subgrupos e a seleção é feita arbitrariamente dentro de cada subgrupo e tantas vezes segundo sua população.

Neste estudo de caso as amostras utilizadas foram amostras de múltiplos estágios e cotas, pois o Banco de Dados utilizados e criado possui uma estrutura hierárquica e dividida em subgupos.

Por exemplo, a amostragem de clientes com determinada tensão, fornecimento consumo e tarifa:

- Tensão Fornecimento:
    - A1 TENSÃO DE FORNECIMENTO IGUAL OU SUPERIOR A 230 kV

- A2 TENSÃO DE FORNECIMENTO DE 88 KV A 138 kV
- A3 TENSÃO DE FORNECIMENTO DE 69 KV
- A3A TENSÃO DE FORNECIMENTO DE 30 KV A 44 kV
- A4 TENSÃO DE FORNECIMENTO DE 2,3 KV A 25 kV

- Classe Consumo:
  - TU011 Industrial
  - TU012 Comércio e Serviços Comercial
  - TU013 Fundações (ligadas ao estado)
  - TU014 Rural
  - TU015 Rural Irrigantes
  - TU016 Poder Público Federal
  - TU017 Poder Público Estadual

- Tarifa:
  - 2 Horo-Sazonal Verde
  - 3 Convencional
  - 4 Irrigante
  - 5 Amarelo

3.3.1 Triagem dos Dados

Após ter sido selecionada uma amostra, três tarefas devem ser seguidas. Estas tarefas fazem parte do processo de triagem de dados: tratamento de erros, valores aberrantes (*outliers*) e valores faltantes (*missing values*). Antes se deve lembrar que existem dois tipos de classificação de dados: os qualitativos e os quantitativos. [BRAGA, 2005].

Os qualitativos, por exemplo: Clientes separados por classe de consumo.

Os quantitativos, por exemplo: Quantidade de clientes horo-sazonal no cadastro da concessionária.

A detecção de um erro ou *outlier* nos dados qualitativos e quantitativos é fácil, pois basta verificar se os valores da amostra correspondem aos valores reais. Se um erro for detectado para algum elemento, este deverá ser descartado ou substituído pela 'moda' (que mais se repete). [BRAGA, 2005].

Um problema de Mineração de Dados reside no fato de ter como meta encontrar a exceção e não a regra ou, dito de outra forma, o objetivo é a detecção de desvio. No caso dessa dissertação, deve incluir testes que detectem desvios significativos do padrão usual do consumo dos clientes de uma determinada classe.

Também em dados quantitativos dados errados podem ser substituídos, neste caso pela média ou pela mediana [BRAGA, 2005].

Veja o exemplo a seguir:

Tabela 7 – Exemplo de substituição de outliers

Renda (sm)	**5**	**4,5**	**6**	**4,8**	**5,1**	**8**	**9**	**10**	**11**
Consumo (kWh) ao mês	**80**	**70**	**120**	**400**	**100**	**200**	**350**	**400**	**140**

Se tivéssemos dúvidas sobre o erro do consumo de energia de 400 kWh, ele poderia ser substituído pela média local dos pares próximos: (80+70+120+100)/4 = 92,5.

3.4 Transformação dos Dados

Muitas vezes as variáveis se encontram de uma forma pouco conveniente, então são apresentadas técnicas que podem ser úteis. Estas são chamadas de transformação de dados do projeto de Mineração de Dados. [BRAGA, 2005].

A Transformação de dados está dividida em sete tipos de aplicações, são elas: sumarização, razões, codificação, codificação simbólica, redução de variáveis, parametrização e transformações matemáticas.

Nesta dissertação, as aplicações usadas foram:

1. Sumarização - substituição de uma média diária de Potência Reativa por uma média mensal;

2. Redução de variáveis - a eliminação de dados redundantes e irrelevantes.

3. Transformações matemáticas - como exemplo de transformação pode-se citar a operação de soma dos campos (R.DM_LIDA +R.DM_LIDA_PTA + R.DM_LIDA_FPTA) da tabela SGC_RESUMO e apresentá-los em um único campo chamado DEMANDA.

Neste capítulo, foram abordadas algumas técnicas e descobertas de padrões em banco de dados, técnicas estas que são imprescindíveis para se ter um bom resultado em uma Mineração de Dados, uma vez que foi apresentado um problema que constitui desvio. Foi feita uma definição da população conforme os dados disponibilizados para essa dissertação. Estes dados passaram por uma triagem, sendo descartados valores aberrantes (valores que não condizem com a realidade) e valores faltantes (campos em branco). Logo após foi feita a transformação desses dados, como sumarização, redução de variáveis consideradas irrelevantes para o processo.

Capítulo 4 – Classificação e Predição para Detecção de Desvio de Comportamento de uso

4 Classificação e Predição para Detecção de Desvio

Num projeto de Mineração de Dados existem seis etapas. Dentre estas, a duas deve-se dar maior importância: a Classificação e a Predição. A utilização de diferentes algoritmos é de extremamente relevante para a avaliação de resultados, pois no processo de Mineração de Dados são ponderados não apenas a precisão de um modelo, mas também a velocidade, a robustez e a capacidade de interpretação dos dados [PETERMANN, 2006].

Dentre os vários algoritmos existentes para os mais diversos tipos de tarefas, foram selecionados os mais recomendados (o de Classificação e o de Predição) para a tarefa de detecção de desvio segundo a bibliografia e trabalhos pesquisados. Estes algoritmos selecionados encontram-se à disposição na ferramenta Weka. São eles:

Apriori – o algoritmo recomendado para associação e criação de padrões dos consumidores com desvio de comportamento;

Árvores de Decisão – os algoritmos selecionados foram o ID3 e o J48.

Classificadores Bayesianos – o classificador selecionado foi o Bayes Net.

Os fatores que foram levados em consideração para a escolha desses algoritmos foram: a disponibilidade, o desempenho e o custo computacional como ferramentas de classificação a serem utilizadas na detecção de desvio [PETERMANN, 2006].

O desempenho de cada um dos algoritmos foi analisado, testado previamente e apresentado neste capítulo. Apriori, Árvore de decisão e os Classificadores Bayesianos foram os que obtiveram um melhor desempenho junto aos dados selecionados como amostra.

O custo computacional foi avaliado, tendo em vista os recursos necessários para a execução dos algoritmos em função do tempo. Em testes preliminares foram utilizadas as Redes Neurais *MLP´s* (*MultiLayer Perceptron*), tendo sido descartadas em função do tempo

de processamento. Estes algoritmos acabam exigindo muito do *Hardware,* tornando inviável o uso do mesmo.

Nas seções a seguir serão detalhados os algoritmos de classificação e predição. A aplicação destes e os resultados obtidos com os mesmos serão observados no Capítulo de estudo de caso.

4.1 Apriori

A palavra apriori significa etapa para se chegar ao conhecimento através do conhecimento dedutivo; já o algoritmo tem por característica sua simplicidade e versatilidade aplicada a grandes volumes de dados, levando em consideração o fator de confiança, pois quanto menor o fator confiança melhor. Outra característica é sua ágil criação de regras.

Segundo CARVALHO et. tal o algoritmo Apriori realiza a mineração de dados em dois passos: geração e poda. No primeiro passo, o algoritmo executa uma varredura no arquivo, com o propósito de criar os conjuntos de combinações entre valores dos atributos existentes e editados no arquivo. No segundo passo, o algoritmo leva em consideração os conjuntos de dados existentes no arquivo com uma não menor freqüência pré-fixada no valor mínimo, conhecido como grandes conjuntos. Neste caso este valor é pré-fixado na ferramenta de Weka.

> Este algoritmo utiliza o princípio de que cada subconjunto de um conjunto de itens freqüentes também deve ser freqüente. Esta regra é utilizada para reduzir o número de causas a serem comparadas com cada transação do banco de dados. Todas as causas geradas que contenham algum subconjunto que não seja freqüente são eliminadas [Lange, 2007].

O trabalho tem por objetivo demonstrar a viabilidade do emprego do algoritmo Apriori ao sistema para detectar distúrbios ou anomalias na medição de demanda de energia elétrica. Este estudo pode ser importante não somente para mostrar estas anomalias, mas também para auxiliar na tomada de decisão com a exploração das regras de associações.

Para a análise do banco de dados em que está armazenado uma série de informações (demanda, tipo de consumidor, tipo de contribuinte, local da instalação...), é possível fazermos uma série de associações importantes entre estas informações, de tal forma que um item implique atuação do outro, ou seja, através do aumento da demanda de um determinado consumidor de uma região em tempo chuvoso constata-se uma anomalia que deve ser

investigada. A aplicação do algoritmo tem por objetivo encontrar estas regras de associação que são relevantes entre estes itens. O tratamento destas questões segue um modelo matemático, em que as regras de associação devem atender a um suporte e confiança mínima, especificado por quem está tomando a decisão [Han et. Tal, 2001].

4.2 Classificadores Bayesianos

Segundo HRUSCHKA (1997) *apud* PETERMANN, em 1973 foi publicada a Regra de Bayes, produzida por Thomas Bayes. Esta regra possibilita que a probabilidade de algum evento poderia ser dada através do conhecimento humano. Eventos onde a freqüência não pode ser medida, a probabilidade se dá através do conhecimento de um especialista.

Na década de 70 a estatística bayesiana passou a ter uma utilizaç4ão junto a sistemas de IA, nesta época ainda não bem definida [MELLO, 2002].

A Rede Bayesiana é um gráfico acíclico (*DAG – Directed Acyclic Graph*), diferente de uma árvore de decisão, pois cada nó possui uma distribuição de probabilidade condicionada. As variáveis de domínio são representadas por nó x e os arcos representam a dependência condicional entre cada nó [CHENG et al.,1999]. A Figura 8 representa essa rede.

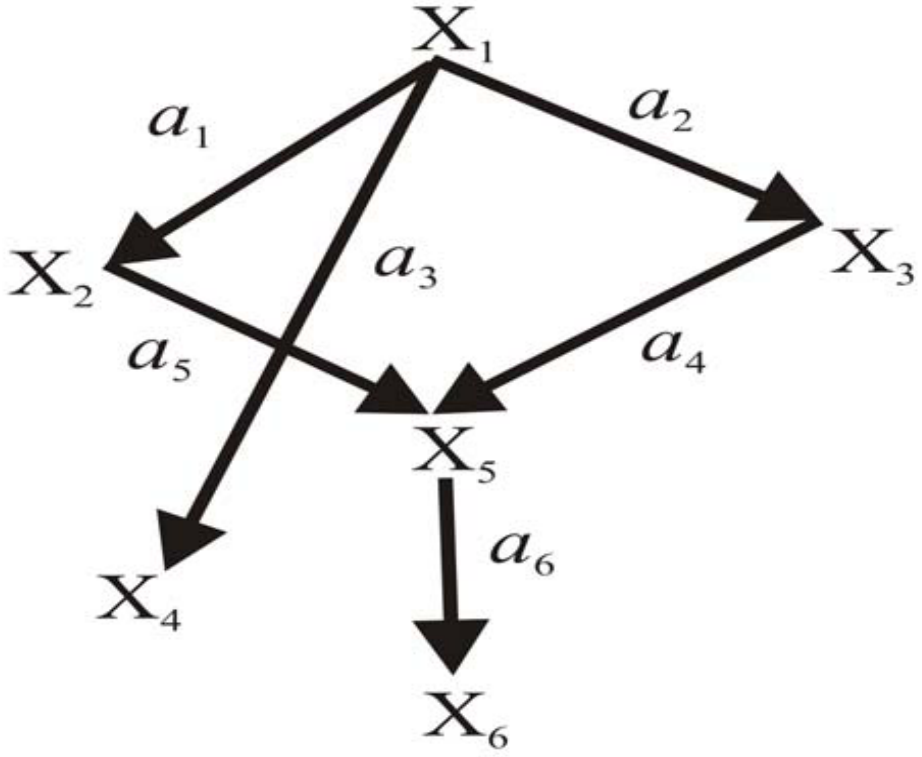


Figura 8 – Rede Bayesiana para a distribuição P(X1,X2,X3,X4,X5,X6)

Fonte: MELLO, 2000

Segundo PEARL (1988) *apud* PETTERMAN ao visualizar um Gráfico *DAG* e verificar se ele é uma rede Bayesiana, encontra-se a interdependência condicional de todos os nós de cada variável x e seus descendentes, exceto seus nós pais. Esta condição faz com que se reduza significativamente o custo computacional, uma vez que há uma distribuição conjunta das probabilidades.

Entre as Redes Bayesianas uma das que se mais destaca é o Classificador BayesNet, amplamente utilizado nos trabalhos publicados, tendo sido a mesma selecionada.

As redes Bayesianas têm como característica representar uma classe ou atributo no nó pai e não permitem que os nós filhos possuam arco ou ligação entre si. Seu processo de classificação é eficiente, pois geralmente seus atributos ou classes são independentes entre si. Esta rede vem sendo utilizada normalmente em processos de classificação e de predição [MELLO, 2002].

As redes Bayesianas estão disponível na maioria dos softwares de mineração. Possuem uma facilidade de implementação de seus algoritmos e são muito eficientes como classificadoras pela boa previsão e precisão [CHENG et al.,1999].

As redes Bayesianas estão disponível na maioria dos softwares de mineração, possuem uma facilidade de implementação de suas algoritmos uma fácil implementação, são muito eficientes como classificador pela boa previsão e precisão [CHENG et al.,1999].

4.3 Árvore de Decisão (*Decision Trees*)

Ross Quinlan, da Universidade de Sidney, desenvolveu um algoritmo chamado ID3, em 1983. O ID3 e suas evoluções (ID4, ID6, C4.5, See 5) são bem adaptadas para o uso em conjunto com as árvores de decisão. Estas são responsáveis por gerar as regras, as quais são ordenadas pelo grau de importância, produzindo assim através dos fatos um modelo de árvore de decisão que afeta diretamente os itens de saída. [COUTINHO, 2003].

As Árvores de Decisão são procedimentos hierárquicos cujo método possui a função de definir e predizer as classes através de variáveis preditoras. [BRAGA, 2005].

Dependendo da complexidade de um problema, o mesmo é dividido em sub-problemas mais simples. Esta técnica é aplicada a cada outro sub-problema, por isso a Árvore de Decisão utiliza uma técnica ou estratégia chamada dividir-para-conquistar. [GAMA, 2000].

Esta capacidade de subdividir os problemas vem das características de divisão do espaço, definidas pelos atributos em sub-espaços, onde cada subespaço é associado a uma classe.

Segundo GARCIA (2000), as Árvores de Decisão consistem de nós que representam os atributos; ramos, estes provenientes dos nós e que recebem os valores possíveis para estes atributos (o ramo descende ou corresponde a um valor deste atributo) e de nodos folha (folha da árvore), que representam as diferentes classes de um conjunto de treinamento. Deste modo, cada folha está associada a uma classe. Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.

Um exemplo de árvore de decisão para uma aplicação de correlação que indicaria fraude ou desvio de comportamento de consumidores de energia elétrica, teria uma forma

mais ou menos como a apresentada na Figura 9. As regras estão exemplificadas na Tabela 8, mostrados a seguir.

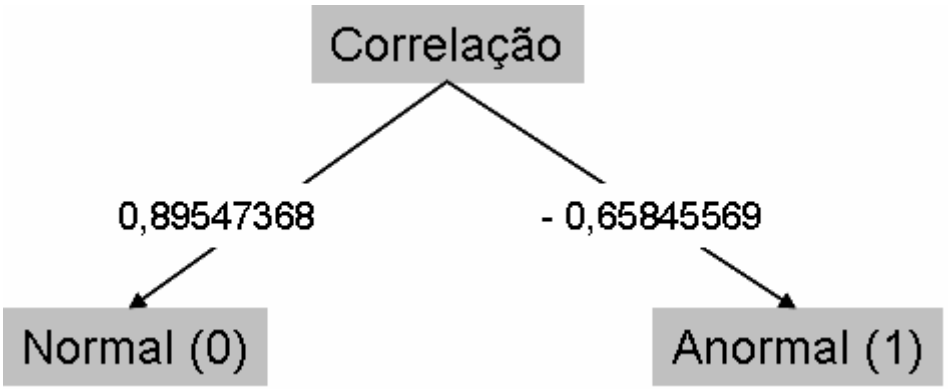


Figura 9 – Árvore de decisão de correlação

Tabela 8 - Tabela Exemplo de condições da Árvore de Decisão

SE correlação > 0,1
ENTÃO cliente = Normal.
SE correlação < = 0,1
ENTÃO cliente = Anormal.

Para o tema em questão – detecção de desvio de comportamento – foi analisado o comportamento do algoritmo J48 perante esta tarefa, algoritmo este que é um dos principais algoritmos de árvore de decisão, também disponibilizado no software Weka, este algoritmo foi testado previamente para esta situação comentada anteriormente e se obteve bons resultados.

Outro algoritmo seleciona para a utilização no estudo de caso foi algoritmo ID3 (*Inductive Decision Tree*, um dos primeiros algoritmos de árvore de decisão), que teve uma redução nos erros de poda, nas regras de verificação pós-poda, trabalhando com atributos

contínuos, com valores ausentes e com o aumento do desempenho computacional [PETERMANN, 2006].

Pode-se observar que todos os algoritmos selecionados nessa dissertação têm a mesma função: classificação e predição; conseqüentemente, necessitam de um treinamento supervisionado. Na Sessão 6 será traçado um comparativo entre as três redes selecionadas.

Foi apresentado neste capítulo um estudo sobre os tipos de métodos e ferramentas de classificação e predição e, dentre os métodos apresentados, como Apriori, Árvore de Decisão e Classificadores Bayesianos, foram escolhidos algoritmos em meio aos diversos existentes em cada método, quais sejam: o algoritmo Apriori, algoritmo J48 e ID3 e o algoritmo BayesNet, que serão utilizados no processo para detecção de desvios. Estes três tipos de algoritmos foram selecionados através de testes. Foi feita uma análise levando em consideração sua disponibilidade (pois todos são encontrados no pacote Java da ferramenta de Mineração de Dados Weka, a qual foi escolhida e utilizada neste estudo de caso), desempenho (foram avaliados os algoritmos com melhores desempenhos e maior velocidade de resposta) e custo computacional (devido ao tamanho do banco de dados utilizado e desenvolvido no estudo de caso).

A escolha de três métodos para classificação e predição e escolha de três algoritmos distintos tem como objetivo fazer um comparativo entre os mesmos, para assim fazer um confronto dos resultados e analisar qual irá apresenta melhor desempenho e resultado para o estudo de caso em questão.

Capitulo 5 - Metodologias para Detecção de Desvio de Comportamento

5 Metodologia

5.1 Introdução

Este capítulo tem como objetivo apresentar um método para detecção de desvio em diversas tipologias de clientes. Os dados disponíveis para o estudo de caso são constituídos pelos Clientes Horo-sazonais Irrigantes, que fazem parte, na sua totalidade, da região da Fronteira Oeste do Estado do Rio Grande do Sul, mais especificamente plantadores de arroz. Estes clientes foram selecionados por serem os clientes que representam um maior impacto à concessionária, do ponto de vista de consumo de energia. São também clientes que possuem curvas típicas, fato este importante para o uso deste método de análise proposto.

As distribuidoras de energia elétrica realizam as medições dos seus clientes de maior impacto, clientes esses classificados como consumidores do Grupo "A", assinalados como clientes horo-sazonais.

Essas informações são coletadas por medidores de energia instalados no consumidor e configurados para receberem as informações de consumo e demanda em um intervalo de tempo selecionado e configurado pela concessionária. Estas informações são transferidas para um computador, no qual constituem um Banco de Dados.

Para este tipo de clientes foi desenvolvido um método que define perfis de comportamento nos períodos de safra, sendo possível classificá-los em clientes com comportamentos normais (clientes sem suspeita de desvio) ou clientes com comportamentos anormais, classificados como problema (clientes com suspeita de desvio).

Em 2003 iniciou-se o desenvolvimento de um projeto de P&D em conjunto com uma concessionária de energia, focando o impacto da demanda dos clientes rurais horo-sazonais da região da Fronteira Oeste do Estado do Rio Grande do Sul. Como foi referido, tais clientes em sua maioria são plantadores de arroz que usam a energia para irrigação. As leituras de memória de massa destes clientes foram disponibilizadas tendo sido utilizados dados do mesmo projeto para a realização da Mineração de Dados nesta dissertação.

A concessionária em questão utilizou-se de informações em intervalos de 5 minutos, totalizando 9000 registros mensais por cliente. Os registros são armazenados em arquivos chamados "memória de massa" nos medidores.

Os arquivos públicos[5] gerados pelo medidor dos clientes da região da fronteira foram disponibilizados para o Grupo de Pesquisa de Gestão de Energia da PUCRS, em arquivos no formato binário. Os arquivos são lidos através de um software e inseridos no banco de dados desenvolvido para essa dissertação, banco esse chamado de PROPUSDM (PROPUS nosso servidor de dados e *DM* de *Data Mining*).

Outra fonte de dados fornecida pela concessionária contempla as informações do banco de dados de faturamento da mesma; estas foram obtidas através do banco de dados, um formato chamado *Dump*, que é gerado através de uma exportação feita neste banco de dados alocado num arquivo cuja extensão é DMP. O responsável pela administração do banco de dados da concessionária executa uma instrução *SQL* com os parâmetros necessários para a extração somente dos dados relevantes para o projeto.

Vale destacar que o banco de dados desenvolvido para a mineração possui um número menor de tabelas e a quantidade de dados que compõem as mesmas também são menores, mas não menos importante que o banco de dados na sua totalidade, uma vez que esta redução da quantidade de tabelas e de dados se deve à normalização e limpeza, para que este banco ficasse mais adequado para uma boa mineração.

A Figura 10 mostra o *software* desenvolvido, no módulo curvas típicas, onde se faz a importação dos arquivos públicos, após concluir uma inserção de um arquivo no banco.

---

[5] Arquivos Públicos é um arquivo texto padronizado pela NBR 14522 de 2000. Esta norma define o padrão de intercâmbio de informações no sistema de medição de energia elétrica brasileiro, de forma a se alcançar a compatibilidade entre os sistemas e equipamentos de medição de energia elétrica de diferentes procedências e fabricantes [SIADAGE, 2007].

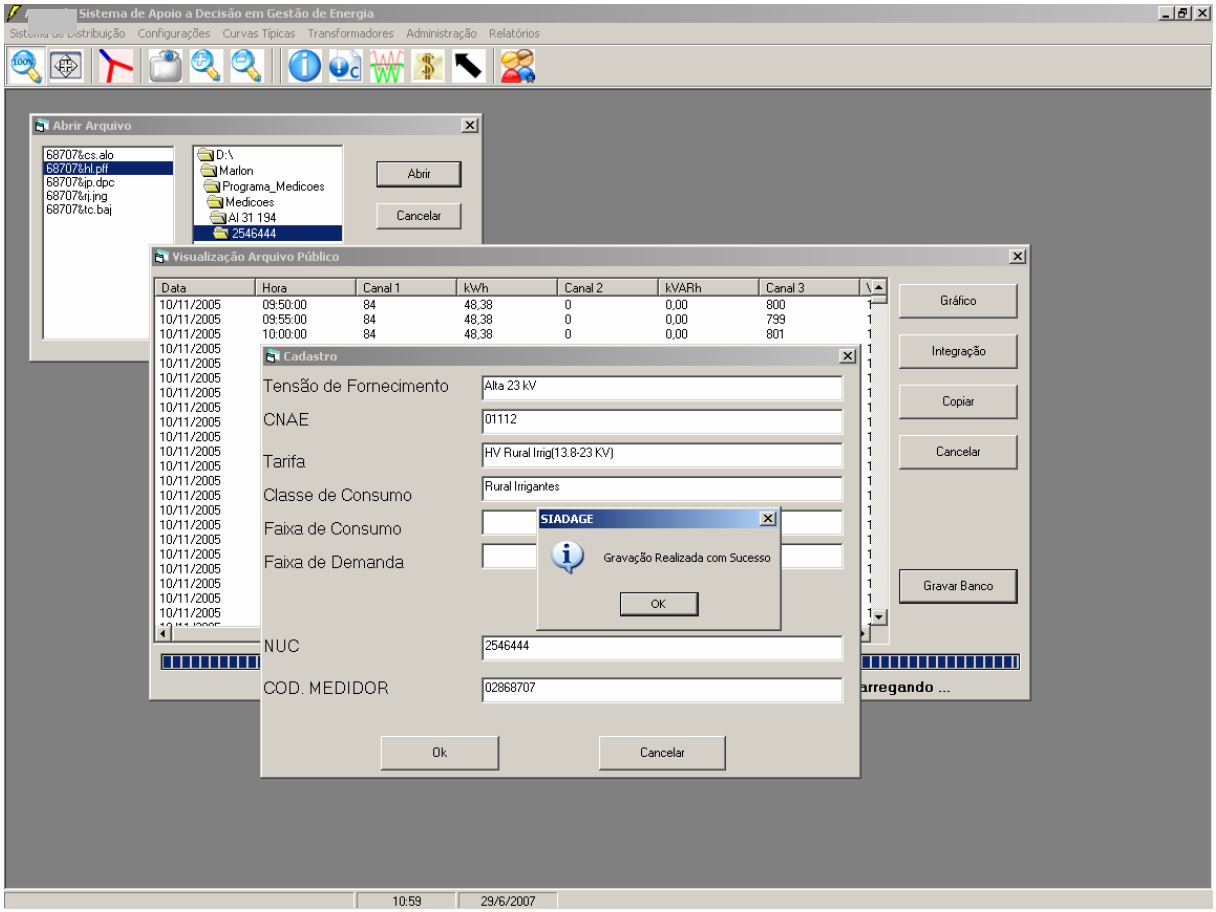


Figura 10 - Inserção dos Arquivos Públicos

Após terem sido processados pelo *software* que realiza a interpretação e inserção dos arquivos públicos no banco de dados, os mesmos resultam no conteúdo dos atributos abaixo listados e inseridos em duas tabelas: uma com os dados referentes à medição e outra referente aos valores das medições respectivamente. Estas tabelas são chamadas de MEDICOES e MEDICOES_VAL com os seguintes campos:

- ID_MEDIDOR – número do medidor;
- DATA_INICIO – data do início da medição;
- DATA_FIM – data do fim da medição;
- INTERVALO – intervalo no qual foi feito a medição, neste caso de 5 em 5 minutos;
- CANAL 1 – valores brutos do canal 1;

- CANAL 2 – valores brutos do canal 2;
- CANAL 3 – valores brutos do canal 3;
- POTENCIA_ATIVA – valor calculado do canal 1;
- POTENCIA_REATIVA – valor calculado do canal 2;
- TENSAOHORA_MEDICAO - – valor calculado do canal 3;
- HORA_MED – data e hora da medição;
- FATOR_POTENCIA – razão entre a energia elétrica ativa e a raiz quadrada da soma dos quadrados das energias elétricas ativas e reativas, consumidas num mesmo período especificado.
- POT_APARENTE – é a raiz da soma dos quadrados da potência ativa e reativa.

Estes dados são de muita importância para o método, já que a partir destas informações serão estratificadas as curvas típicas e os dias típicos, para então determinar um perfil de consumo padrão e assim detectar a possível anormalidade, que poderá ou não se caracterizar como desvio.

5.2 Método Utilizado na Detecção de Desvio de Comportamento de Uso de Energia Elétrica

O método para identificação de desvio de comportamento de energia elétrica tem como objetivo detectar padrões de comportamento para cada consumidor analisado. Este método tem início na aquisição dos dados, tanto dos arquivos de memória de massa, quanto dos arquivos DMP (*Dump* gerados pelo banco da concessionária). Após essa etapa é feita a obtenção e a avaliação dos dados, onde os mesmos passam por uma análise em que selecionam-se os atributos relevantes. Logo em seguida é feita uma subdivisão e uma limpeza nos dados, preparando-os para que posteriormente sejam inseridos no banco de dados criado para recebê-los.

Após as etapas citadas no parágrafo anterior, os dados são selecionados e pré-processados. Em seguida é efetuada uma transformação e a normalização dos mesmos. Posteriormente é realizada a Mineração de Dados utilizando o software Weka, através dos algoritmos adequados, que foram selecionados e apresentados no Capítulo anterior. Então os resultados serão interpretados e avaliados, e após estas etapas vem a extração de conhecimento, procurando identificar uma possível anormalidade, que indicará o desvio de comportamento.

A metodologia empregada para detectar desvio de comportamento no sistema de fornecimento de energia elétrica da concessionária foi elaborado através dos seguintes parâmetros:

- Leitura da potência

  Normal, baixa, muito baixa e alta

- Região (norte, sul, leste e oeste)

- Horário de ponta considerado das 17h às 20horas

- Dias da semana: úteis e feriados

- Potência instalada:

  Até 100 kW de demanda

  Maior que 100 KW e menor que 200 KW de demanda

  Maior que 200 KW

- Com transformador instalado ou sem transformador

- Histórico do consumo de energia elétrica.

A seguir os modelos das tabelas utilizadas:

Tabela 9 – Tabela de Regiões

Região			
Norte	Sul	Leste	Oeste
n1	s1	l1	o1
n2	s2	l2	o2
n3	s3	l3	o3

Na Tabela 9, temos alguns exemplos de localizações de clientes e conforme a região onde ele sua propriedade esta localizada.

Tabela 10 – Tabela de Potências

Potência Instalada em kw		
<=100	>100<=200	>200

Já na Tabela 10 mostramos a classificação do cliente quanto à potência instalada, como menor que 100kW (mei100), entre 100kW e 200kW (ma100) e maior que 200kW (ma200).

Tabela 11 – Tabela Variada

	Horário Ponta	Dias úteis	Transforma-dor	Histórico
Sim				
Não				

A Tabela 11 aponta se o cliente possui a medição no horário de ponta, se a medição é de um dia útil, se ele possui transformador na propriedade e se tem um histórico de consumo bom (normal) ou ruim (problema).

Tabela 12 – Classe de Consumo

Classe de consumo			
ad	bd	cd	id

Na Tabela 12 apresentamos a classe de consumo destes clientes, onde: Monofásica “ad”, Bifásica “bd”. Trifásica “cd” e ainda Trifásica com medição indireta representado por “id”.

Através destas tabelas poderemos fazer uma composição do tipo de consumidor que iremos encontrar ao fazermos a mineração de dados, ou seja, o consumidor que está instalado em uma determinada região será avaliado com perfil de demanda, isto é:

Consumidor da região norte, sul, leste ou oeste

Será avaliado pela potência <= 100, .100<=200 ou >200

Trabalhando no horário de ponta ou fora de ponta

Em dia útil ou feriado

Com transformador ou não

Pertencente a classe de :

Monofásica “ad”

Bifásica “bd”

Trifásica “cd”

Ou trifásica com medição indireta

Com histórico bom ou problemático

Com estes dados poderemos completar ainda o estudo através da potência demanda, isto é, se a potência demandada lida está:

Normal, Alta, Baixa ou muito baixa.

Com estas informações poderemos traçar um perfil destes consumidores no banco de dados fornecido pela concessionária.

Na aplicação da ferramenta Weka visualizamos os 12 atributos na Figura 11.

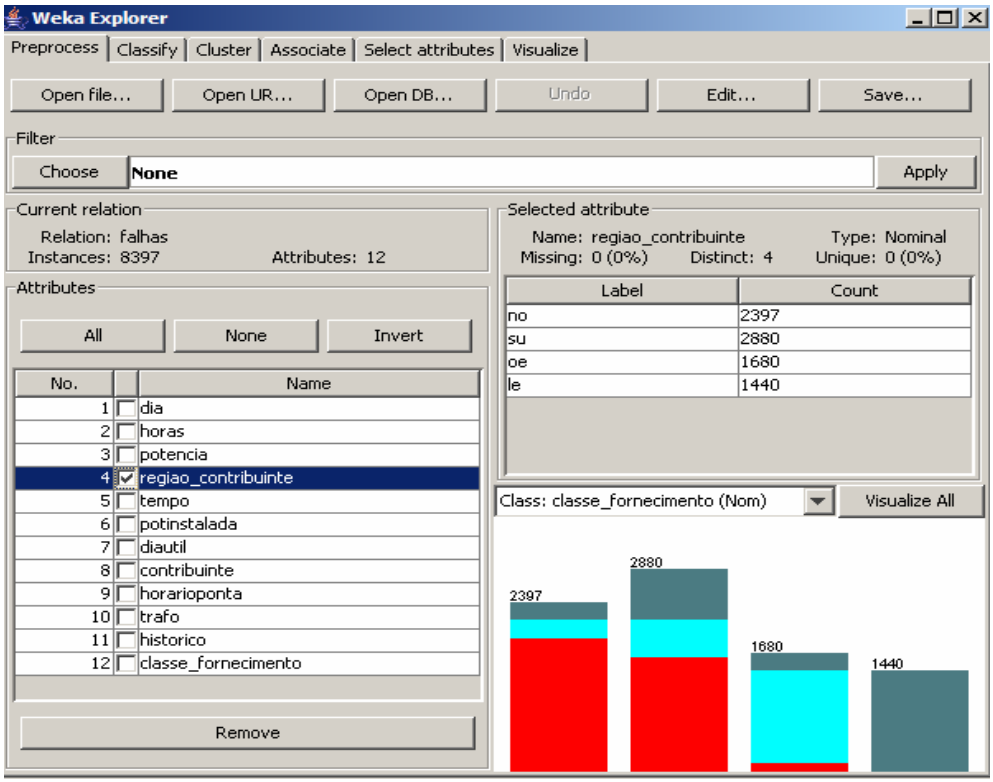


Figura 11 – Atributos utilizados na MD

Nesta Figura 11 podemos visualizar os atributos considerados significativos, bem como a região em que poderemos encontrar estes consumidores. Em nosso estudo identificaremos os consumidores através da localização geográfica: norte, sul leste e oeste, como descrita na tabela acima.

O período utilizado para a análise e mineração foi o mês de outubro, pois nesta época a demanda de energia elétrica é considerada muito significativa, pois é a época em que o plantio de arroz tem início.

Através da Figura 12 poderemos visualizar as horas que foram computadas para a verificação de possíveis distúrbios de comportamento de demanda energética.

Em nossa base de dados estes tempos são fornecidos de 5 em 5 minutos. Porém, com o atributo histórico de mês anterior e variação da demanda, ou seja, desvio de demanda contratada com relação à demanda consumida, estes tempos podem ser reduzidos. Desta forma, tomamos tempos em horários estratégicos com o objetivo de detectar anormalidades

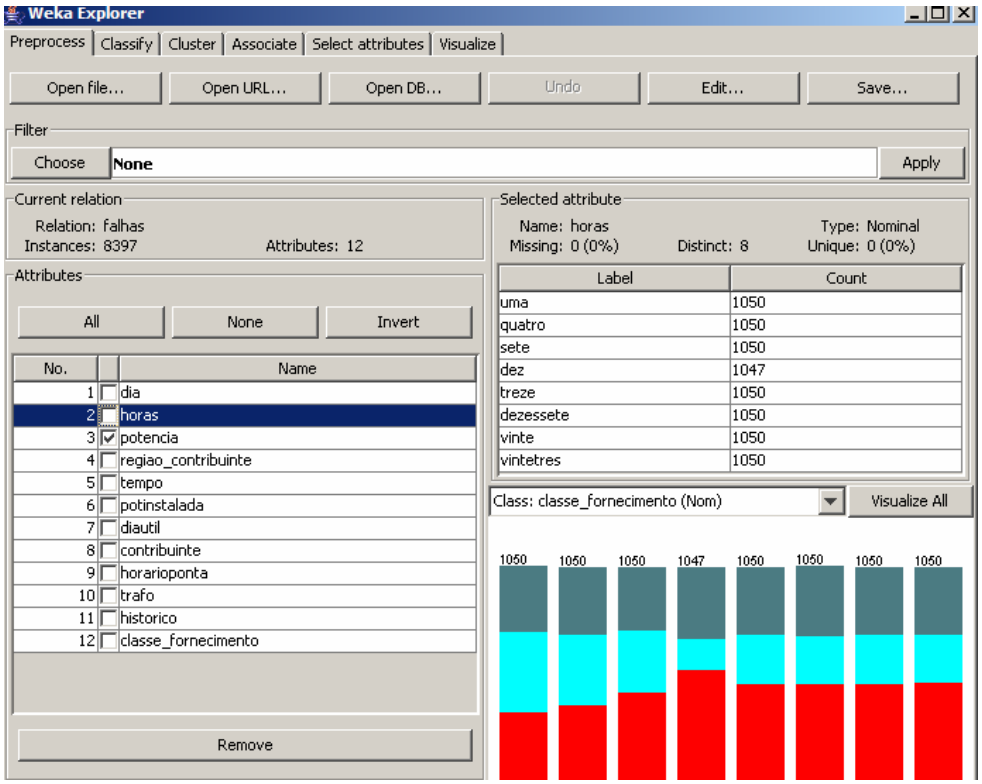


Figura 12 – Tempos computados na MD

No Figura 13 podemos visualizar os atributos em que identificamos a potência muito baixa (mb), potência normal, potência alta e potência baixa.

Este atributo nos informa as características e desvios possíveis de determinados consumidores; porém, com vistas a um estudo mais aprofundado teremos que executar algoritmos para identificar estes desvios de demanda de energia que possam informar se consistem em casos normais ou realmente desvios de padrões de consumo de energia.

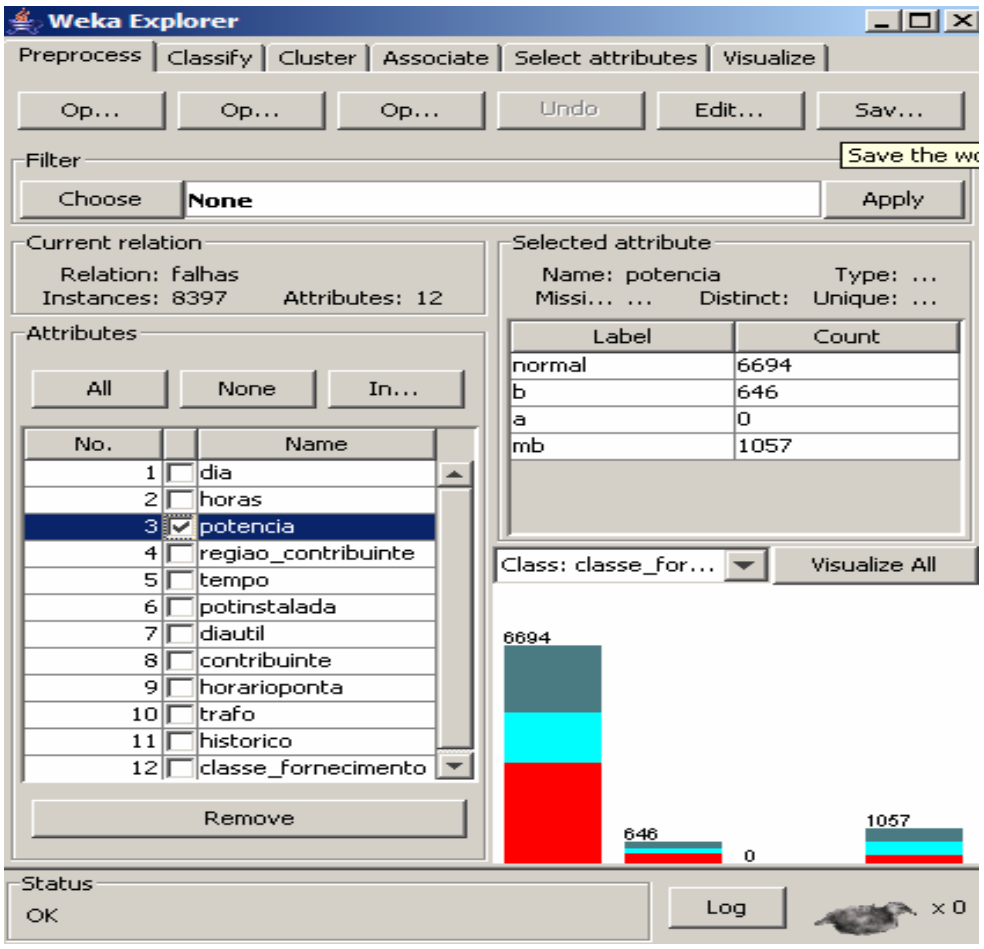


Figura 13 - Identificar potência demandada

5.3 Metodologia para a Mineração Dados

5.3.1 Regra de associação do algoritmo Apriori

O trabalho tem por objetivo demonstrar a viabilidade do emprego do algoritmo Apriori ao sistema de desvios de padrões na demanda de energia elétrica de consumidores arrozeiros. Este estudo pode ser importante não somente na detecção de desvios de padrões de consumidores de energia elétrica, mas auxiliar, ainda, na tomada de decisão com a exploração das regras de associações.

Para a análise do banco de dados em que estão armazenadas as informações (região, horário de ponta, histórico, contribuinte, etc), é possível fazermos uma série de associações importantes entre estas informações, de tal forma que um item implique atuação do outro, ou seja, o consumidor de uma determinada região em dia chuvoso poderá ter uma demanda

normal. A aplicação do algoritmo tem por objetivo encontrar estas regras de associação que são relevantes entre estes itens. O tratamento destas questões segue modelo matemático, em que as regras de associação devem atender a um suporte e à confiança mínima, especificados por quem está tomando a decisão. O suporte corresponde à freqüência com que ocorrem os padrões em toda a base de dados, como por exemplo o histórico do contribuinte; sendo problemático gera a possibilidade de que tenha desvio de padrão no seu consumo de energia.

Aplicação do algoritmo Apriori para a descoberta das principais regras.

```
Best rules found:

 1. potencia=nor 7875 ==> tempo=dry historico=nm 7875    conf:(1)
 2. tempo=dry 7875 ==> potencia=nor historico=nm 7875    conf:(1)
 3. potencia=nor tempo=dry 7875 ==> historico=nm 7875    conf:(1)
 4. potencia=nor historico=nm 7875 ==> tempo=dry 7875    conf:(1)
 5. tempo=dry historico=nm 7875 ==> potencia=nor 7875    conf:(1)
 6. tempo=dry 7875 ==> historico=nm 7875    conf:(1)
 7. potencia=nor 7875 ==> historico=nm 7875    conf:(1)
 8. potencia=nor 7875 ==> tempo=dry 7875    conf:(1)
 9. tempo=dry 7875 ==> potencia=nor 7875    conf:(1)
10. historico=nm 8397 ==> potencia=nor tempo=dry 7875    conf:(0.94)
```

A regra 1 informa que a potência do consumidor é normal em tempo seco e o histórico de consumo de energia elétrica é normal.

A regra 4 verifica que o consumidor tem uma potência normal e se o seu histórico de demanda também é normal, isto é, corresponde com a demanda contratada.

A regra 10 informa que o histórico do consumidor é normal, que ele possui uma potência normal em tempo bom.

5.3.2 Algoritmo árvore de decisão

O software Weka disponibiliza para uso dos conceitos de árvore de decisão as ferramentas ID3 e J48 e outros algoritmos, a partir da elaboração da tabela de atributos com os dados e a verificação da possibilidade de existência de regras sobre desvios de padrões de consumidores no consumo de energia elétrica nos consumidores desta região.

Iniciaremos agora a aplicação de uma classificação utilizando o algoritmo de árvore de decisão (J48) para identificar qual classe de demanda e podemos visualizar através da matriz denominada de confusão a característica encontrada no banco de dados.

```
=== Confusion Matrix ===

    a    b    c   <-- classified as
 3840    0    0 |    a = mei100
    0 1920    3 |    b = mal00
    0    0 2634 |    c = ma200
```

Este estudo mostrou a característica de demanda destes consumidores, isto é, foram classificados de acordo com a sua potência de demanda.

Aplicando o algoritmo ID3 poderemos visualizar e fazermos uma comparação entre matrizes.

```
=== Confusion Matrix ===

    a    b    c   <-- classified as
 3840    0    0 |    a = mei100
    0 1922    1 |    b = mal00
    0    0 2631 |    c = ma200
```

O número de incidência entre as duas matrizes não foram relevantes nesta aplicação.

O algoritmo ID3 é um dos algoritmos que implementa árvore de decisão, considerado um algoritmo recursivo, ao passo que busca, sobre um conjunto de atributos, aqueles que melhor se encaixam nas raízes das sub-árvores a serem construídas. Inicialmente, todos os atributos, menos o classificatório, são reunidos em um conjunto. Em seguida, o melhor atributo é escolhido e passa a ser a raiz da sub-árvore em construção. Para cada possível valoração deste atributo, é criada uma aresta até as futuras sub-árvores obtidas com a recursividade deste algoritmo. Os dois únicos critérios de parada são quando não há mais instâncias ou atributos a serem analisados.

O algoritmo J48 é uma melhora do Id3, ou seja, além de possuir as mesmas características, ele possui a vantagem de poder lidar com a poda (prunning) da árvore para evitar o sobre-ajustamento, com a ausência de valores, com a valoração numérica de atributos e com a presença de ruídos nos dados [QUILAN, 1986].

Capítulo 6 – Estudo de Caso

6 Resultados Obtidos

Este capítulo tem como objetivo apresentar os resultados obtidos com método desenvolvido no capítulo anterior, com a utilização da ferramenta Weka, agindo sobre uma Base de Dados real fundamentada no banco de dados da concessionária. Esta foi consolidada em um *Data Warehouse* criado para suprir as necessidades deste trabalho, tendo sido ainda desenvolvido um *Data Mart* com as tabelas referentes à tipificação das curvas de clientes, que são as mesmas tabelas nas quais foi realizada a Mineração.

Serão apresentadas também algumas das etapas que foram descritas anteriormente, para detecção de desvio de comportamento, as aplicações de técnicas de descobertas de padrões, bem como a classificação e a predição. Técnicas estas que, em conjunto, serão utilizadas para identificar o comportamento de clientes consumidores de energia elétrica horo-sazonais e serão empregadas também para verificar a existência de alguma anormalidade, podendo ou não ser indício de fraude.

Após a utilização da metodologia desenvolvida, com aplicação do método estatístico de correlação e a utilização do *software* de mineração, a concessionária poderá ser sinalizada de que há um indício de um possível desvio. A mesma poderá se certificar disto através do envio de especialistas para a localidade onde o cliente indicado pela mineração se encontra, para assim confirmar ou não a ocorrência de ilegalidade.

6.1 Aplicação de testes para análise

Para determinação dos resultados corretos para a metodologia de detecção de desvios de padrões de demanda de energia elétrica, serão avaliados os dados referentes ao período do mês de Outubro, como mencionado anteriormente.

6.1.1 Geração de regras de associação a partir do algoritmo Apriori

```
Scheme:       weka.associations.Apriori -N 400 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation:     falhas
Instances:    8397
Attributes:   12
              dia
              horas
              potencia
              regiao_contribuinte
              tempo
              potinstalada
              diautil
              contribuinte
              horarioponta
              trafo
              historico
              classe_fornecimento
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.35 (2939 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 24

Size of set of large itemsets L(3): 36

Size of set of large itemsets L(4): 27

Size of set of large itemsets L(5): 9

Size of set of large itemsets L(6): 1
```

1. A maioria dos contribuintes tem uma demanda normal de potência

2. A maioria dos contribuintes tem um bom histórico em sua contas de energia

3. Os contribuintes com demanda menor igual a 100 kw com potência normal

4. Os contribuintes com demanda menor igual a 2 00 kw com potência normal

5. Os contribuintes com demanda maior que 200 kw com potência normal

6.1.2 Geração de árvores de decisão com base nas regras obtidas no período de 2005

Neste mesmo estudo agora vamos aplicar um algoritmo de classificação que utilizará apenas os parâmetros relevantes e necessários para a obtenção das estratégias de ação com respeito ao desvio de padrões.

Nesse sentido e com base nas regras de associação obtidas no item 6.1.1, observa-se que o perfil dos contribuintes é, em sua maioria, ter uma demanda normal de potência.

Contribuintes com demanda menor e igual a 200 KW possuem uma potência normal.

A maioria dos contribuintes tem um bom histórico em sua conta de demanda de energia.

Geração de árvore de decisão com base nas regras obtidas neste período.

Parâmetros relevantes que na associação do algoritmo apriori não foram manifestadas

Deste modo não serão considerados para o algoritmo de classificação os parâmetros:

- **Classe de fornecimento**
- **Transformador**
- **Potência instalada**

Além dos parâmetros acima, foi observado nas regras obtidas que o dia útil e horário de ponta não geraram qualquer regra de associação. Assim, serão também excluídos para o algoritmo de classificação os parâmetros:

- **dia**
- **dia útil**
- **horário de ponta**

Ainda segundo as regras de associação, a maioria das solicitações teve como motivação desvios de padrões. Por outro lado, existem no BD parâmetros obtidos pelos registros no banco de dados de contribuintes com histórico problemático em suas contas de consumo de energia elétrica.

Geração de árvore de decisão com base nas regras obtidas neste período com os parâmetros relevantes, que na associação do algoritmo Apriori não foram manifestados.

```
=== Run information ===
Scheme: weka.classifiers.trees.Id3
Relation:falhas-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R6-
weka.filters.unsupervised.attribute.Remove-R10-weka.filters.unsupervised.attribute.Remove-R8-
weka.filters.unsupervised.attribute.Remove-R7
Instances:    8397
Attributes:   7
              horas
              potencia
              regiao_contribuinte
              tempo
              potinstalada
              contribuinte
              historico
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
              Id3


potencia = normal
|  contribuinte = n1
|  |  horas = uma: bom
|  |  horas = quatro: bom
|  |  horas = sete: bom
|  |  horas = dez: bom
|  |  horas = treze: bom
|  |  horas = dezessete: bom
|  |  horas = vinte: bom
|  |  horas = vintetres: bom
|  contribuinte = s1
|  |  horas = uma: bom
|  |  horas = quatro: bom
|  |  horas = sete: bom
|  |  horas = dez: bom
|  |  horas = treze: bom
|  |  horas = dezessete: bom
|  |  horas = vinte: bom
|  |  horas = vintetres: bom
|  contribuinte = o1
|  |  horas = uma: bom
|  |  horas = quatro: bom
|  |  horas = sete: bom
|  |  horas = dez: bom
|  |  horas = treze: bom
|  |  horas = dezessete: bom
|  |  horas = vinte: bom
|  |  horas = vintetres: bom
|  contribuinte = le1: problema
|  contribuinte = n2
|  |  horas = uma: bom
|  |  horas = quatro: bom
```

| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = s2
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = o2
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = le2
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = n3
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = s3
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = o3
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = le3
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom

| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = n4
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = s4
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = o4
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = le4
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = n5
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = s5
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = o5
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom

| | **horas = vinte: bom**
| | **horas = vintetres: bom**
| **contribuinte = le5**
| | **horas = uma: bom**
| | **horas = quatro: bom**
| | **horas = sete: bom**
| | **horas = dez: bom**
| | **horas = treze: bom**
| | **horas = dezessete: bom**
| | **horas = vinte: bom**
| | **horas = vintetres: bom**
| **contribuinte = n6**
| | **horas = uma: bom**
| | **horas = quatro: bom**
| | **horas = sete: bom**
| | **horas = dez: bom**
| | **horas = treze: bom**
| | **horas = dezessete: bom**
| | **horas = vinte: bom**
| | **horas = vintetres: bom**
| **contribuinte = s6**
| | **horas = uma: bom**
| | **horas = quatro: bom**
| | **horas = sete: bom**
| | **horas = dez: bom**
| | **horas = treze: bom**
| | **horas = dezessete: bom**
| | **horas = vinte: bom**
| | **horas = vintetres: bom**
| **contribuinte = o6**
| | **horas = uma: bom**
| | **horas = quatro: bom**
| | **horas = sete: bom**
| | **horas = dez: bom**
| | **horas = treze: bom**
| | **horas = dezessete: bom**
| | **horas = vinte: bom**
| | **horas = vintetres: bom**
| **contribuinte = le6**
| | **horas = uma: bom**
| | **horas = quatro: bom**
| | **horas = sete: bom**
| | **horas = dez: bom**
| | **horas = treze: bom**
| | **horas = dezessete: bom**
| | **horas = vinte: bom**
| | **horas = vintetres: bom**
| **contribuinte = n7**
| | **horas = uma: bom**
| | **horas = quatro: bom**
| | **horas = sete: bom**
| | **horas = dez: bom**
| | **horas = treze: bom**
| | **horas = dezessete: bom**
| | **horas = vinte: bom**
| | **horas = vintetres: bom**
| **contribuinte = n8**
| | **horas = uma: bom**
| | **horas = quatro: bom**
| | **horas = sete: bom**
| | **horas = dez: bom**
| | **horas = treze: bom**
| | **horas = dezessete: bom**
| | **horas = vinte: bom**
| | **horas = vintetres: bom**

| contribuinte = n9
| | horas = uma: problema
| | horas = quatro: problema
| | horas = sete: problema
| | horas = dez: problema
| | horas = treze: problema
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: problema
| contribuinte = n10
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = s7
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = s8
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = s9
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = s10
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = s11
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = s12
| | horas = uma: bom

| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = o7
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
potencia = b
| contribuinte = n1: bom
| contribuinte = s1: bom
| contribuinte = o1: bom
| contribuinte = le1: problema
| contribuinte = n2: bom
| contribuinte = s2: bom
| contribuinte = o2: bom
| contribuinte = le2: bom
| contribuinte = n3: bom
| contribuinte = s3: bom
| contribuinte = o3: bom
| contribuinte = le3: bom
| contribuinte = n4: bom
| contribuinte = s4: bom
| contribuinte = o4: bom
| contribuinte = le4: bom
| contribuinte = n5: bom
| contribuinte = s5: bom
| contribuinte = o5
| | horas = uma: bom
| | horas = quatro: bom
| | horas = sete: bom
| | horas = dez: bom
| | horas = treze: bom
| | horas = dezessete: bom
| | horas = vinte: bom
| | horas = vintetres: bom
| contribuinte = le5: bom
| contribuinte = n6: bom
| contribuinte = s6: bom
| contribuinte = o6: bom
| contribuinte = le6: bom
| contribuinte = n7: bom
| contribuinte = n8: bom
| contribuinte = n9: problema
| contribuinte = n10: bom
| contribuinte = s7: bom
| contribuinte = s8: bom
| contribuinte = s9: bom
| contribuinte = s10: bom
| contribuinte = s11: bom
| contribuinte = s12: bom
| contribuinte = o7: bom
potencia = a: null
potencia = mb: problema
Time taken to build model: 0.27 seconds

=== Stratified cross-validation ===

```
=== Summary ===

Correctly Classified Instances        8110               96.5821 %
Incorrectly Classified Instances      281                3.3464 %
Kappa statistic                       0.8397
Mean absolute error                   0.0645
Root mean squared error               0.1854
Relative absolute error               27.6755 %
Root relative squared error           54.3436 %
UnClassified Instances                6                  0.0715 %
Total Number of Instances             8397

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
 1          0.247     0.963      1        0.981      bom
 0.753      0         0.998      0.753    0.858      problema

=== Confusion Matrix ===

   a    b   <-- classified as
 7259    2 |    a = bom
  279  851 |    b = problema
```

No relatório mostrado pelo algoritmo Id3 temos:

```
potencia =
|  contribuinte = le1: problema
contribuinte = n9
|  |  horas = uma: problema
|  |  horas = quatro: problema
|  |  horas = sete: problema
|  |  horas = dez: problema
|  |  horas = treze: problema
|  |  horas = dezessete: bom
|  |  horas = vinte: bom
|  |  horas = vintetres: problema
potencia = b
|  contribuinte = le1: problema
|  contribuinte = n9: problema
```

Com relação à potência, a árvore mostra que o contribuinte le1, ou seja, localizado na região leste do município, apresentou em sua leitura problema com relação à demanda de potência.

Para o contribuinte n9, localizado na região norte do município, a leitura mostra problemas, isto é, em oito leituras realizadas seis apresentaram problemas no histórico destes consumidores.

Com relação à potência, a árvore revelou que a potência destes contribuintes estava baixa.

Aplicando o algoritmo J48, poderemos visualizar que esta rotina faz uma poda demasiada, não informando a formação da árvore em detalhes como vimos no algoritmo ID3.

Na figura abaixo podemos visualizar a formação da árvore de decisão montada em que a potência é descrita como normal, ou seja, a potência demandada está de acordo com a potência contratada. O ramo denominado de “b” é considerado a potência baixa e pode ser em dias chuvosos. O ramo designado por “a” é a ligação do contribuinte em monofásica e que não teve nenhum evento. Já o ramo designado com o nome de “mb” (muito baixa), são contribuintes que apresentam em seu histórico com problemas em registros anteriores.

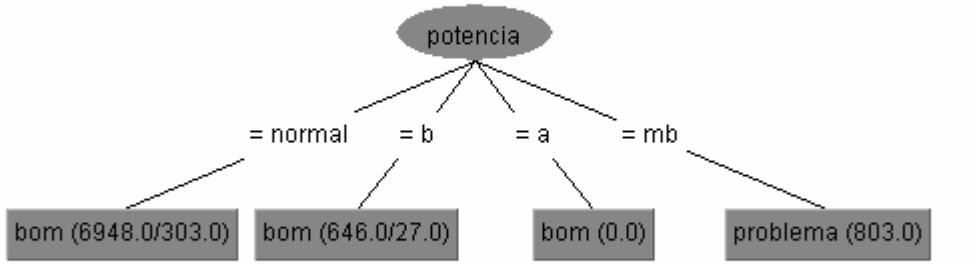


Figura 14 – Árvore de decisão J48

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: falhas-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R6-weka.filters.unsupervised.attribute.Remove-R10-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.Remove-R7
Instances: 8397
Attributes: 7
horas
potencia
regiao_contribuinte
tempo
potinstalada
contribuinte
historico
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
----

potencia = normal: bom (6948.0/303.0)
potencia = b: bom (646.0/27.0)
potencia = a: bom (0.0)
potencia = mb: problema (803.0)

Number of Leaves : 4

Size of the tree : 5

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	**8067**	**96.07 %**
Incorrectly Classified Instances	**330**	**3.93 %**
Kappa statistic	**0.8081**	
Mean absolute error	**0.0752**	
Root mean squared error	**0.1939**	
Relative absolute error	**32.2061 %**	
Root relative squared error	**56.7674 %**	
Total Number of Instances	**8397**	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	0.291	0.957	1	0.978	bom
0.709	0	1	0.709	0.83	problema

=== Confusion Matrix ===

a	b	<-- classified as
7264	0	a = bom
330	803	b = problema

6.1.3 Aplicando o algoritmo Bayes Net

No algoritmo Net bayes poderemos visualizar o Gráfico montado abaixo representado pela Figura 15.

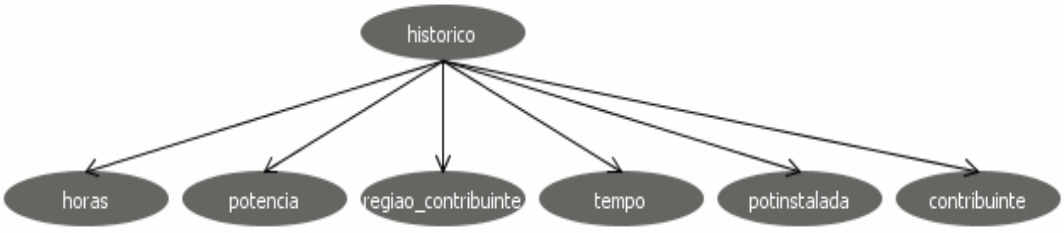


Figura 15 – Representação do Algoritmo Bayes Net

=== Run information ===

Scheme: weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
Relation: falhas-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R6-weka.filters.unsupervised.attribute.Remove-R10-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.Remove-R7
Instances: 8397
Attributes: 7
horas
potencia
regiao_contribuinte
tempo
potinstalada
contribuinte
historico
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Bayes Network Classifier
not using ADTree
#attributes=7 #classindex=6
Network structure (nodes followed by parents)
horas(8): historico
potencia(4): historico
regiao_contribuinte(4): historico
tempo(2): historico
potinstalada(3): historico
contribuinte(35): historico
historico(2):
LogScore Bayes: -74925.74056425689
LogScore BDeu: -74725.22983497397
LogScore MDL: -75258.33209647048
LogScore ENTROPY: -74802.03279267221
LogScore AIC: -74903.03279267221

Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	**8118**	**96.6774 %**
Incorrectly Classified Instances	**279**	**3.3226 %**
Kappa statistic	**0.8417**	
Mean absolute error	**0.0552**	
Root mean squared error	**0.1814**	
Relative absolute error	**23.6446 %**	
Root relative squared error	**53.0878 %**	
Total Number of Instances	**8397**	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.999	0.243	0.964	0.999	0.981	bom
0.757	0.001	0.995	0.757	0.86	problema

=== Confusion Matrix ===

a	b	<-- classified as
7260	4	a = bom
275	858	b = problema

Considerando o atributo hora, agora para determinar os contribuintes que aparecem com potência muito baixa e em que horário este evento ocorre, poderemos visualizar se estão com potência muito baixa em horário de ponto ou fora de ponta.

6.1.4 Aplicando o algoritmo ID3 para visualiza a árvore.

```
=== Run information ===

Scheme:       weka.classifiers.trees.Id3
Relation:     falhas-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R6-
weka.filters.unsupervised.attribute.Remove-R10-weka.filters.unsupervised.attribute.Remove-R8-
weka.filters.unsupervised.attribute.Remove-R7
Instances:    8397
Attributes:   7
              horas
              potencia
              regiao_contribuinte
              tempo
              potinstalada
              contribuinte
              historico
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Id3

tempo = dry
| contribuinte = n1
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vintetres
| contribuinte = s1
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: quatro
| contribuinte = o1
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vintetres
| contribuinte = le1
| | potencia = normal: quatro
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: dezessete
| contribuinte = n2
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
```

| | potencia = a: null
| | potencia = mb: vintetres
| contribuinte = s2
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = o2
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = le2
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: quatro
| contribuinte = n3
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = s3
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = o3
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = le3
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: vinte
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: uma
| contribuinte = n4
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = s4
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte

```
| contribuinte = o4
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = le4
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: vinte
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: uma
| contribuinte = n5
| | potencia = normal
| | | historico = bom: treze
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: uma
| contribuinte = s5
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = o5
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: dezessete
| contribuinte = le5
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: dezessete
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: uma
| contribuinte = n6
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: quatro
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: uma
| contribuinte = s6
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = o6
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = le6
| | potencia = normal
```

| | | historico = bom: dezessete
| | | historico = problema: vintetres
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = n7
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = n8
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = n9
| | historico = bom: vinte
| | historico = problema
| | | potencia = normal: uma
| | | potencia = b: null
| | | potencia = a: null
| | | potencia = mb: vintetres
| contribuinte = n10
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: vinte
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: uma
| contribuinte = s7
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: treze
| contribuinte = s8
| | potencia = normal
| | | historico = bom: vintetres
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = s9
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = s10
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: vinte
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: dez
| contribuinte = s11
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: vinte

```
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: dez
| contribuinte = s12
| | potencia = normal
| | | historico = bom: vintetres
| | | historico = problema: vinte
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: uma
| contribuinte = o7
| | historico = bom: dezessete
| | historico = problema: uma
tempo = chuvoso
| potencia = normal: null
| potencia = b
| | contribuinte = n1: uma
| | contribuinte = s1: uma
| | contribuinte = o1: uma
| | contribuinte = le1
| | | potinstalada = mei100: null
| | | potinstalada = ma100: uma
| | | potinstalada = ma200: dez
| | contribuinte = n2: uma
| | contribuinte = s2: uma
| | contribuinte = o2: uma
| | contribuinte = le2: uma
| | contribuinte = n3: uma
| | contribuinte = s3: uma
| | contribuinte = o3: uma
| | contribuinte = le3: uma
| | contribuinte = n4: uma
| | contribuinte = s4: uma
| | contribuinte = o4: uma
| | contribuinte = le4: uma
| | contribuinte = n5: uma
| | contribuinte = s5: uma
| | contribuinte = o5
| | | historico = bom: uma
| | | historico = problema: uma
| | contribuinte = le5: uma
| | contribuinte = n6: uma
| | contribuinte = s6: uma
| | contribuinte = o6: uma
| | contribuinte = le6: uma
| | contribuinte = n7: uma
| | contribuinte = n8: uma
| | contribuinte = n9: uma
| | contribuinte = n10: uma
| | contribuinte = s7: uma
| | contribuinte = s8: uma
| | contribuinte = s9: uma
| | contribuinte = s10: uma
| | contribuinte = s11: uma
| | contribuinte = s12: uma
| | contribuinte = o7: uma
| potencia = a: null
| potencia = mb
| | potinstalada = mei100: null
| | potinstalada = ma100
| | | regiao_contribuinte = no: uma
| | | regiao_contribuinte = su: null
| | | regiao_contribuinte = oeste: null
| | | regiao_contribuinte = le: uma
| | potinstalada = ma200: quatro
```

Time taken to build model: 0.2 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	**508**	**6.0498 %**
Incorrectly Classified Instances	**7889**	**93.9502 %**
Kappa statistic	**-0.0738**	
Mean absolute error	**0.2214**	
Root mean squared error	**0.3363**	
Relative absolute error	**101.2168 %**	
Root relative squared error	**101.698 %**	
Total Number of Instances	**8397**	

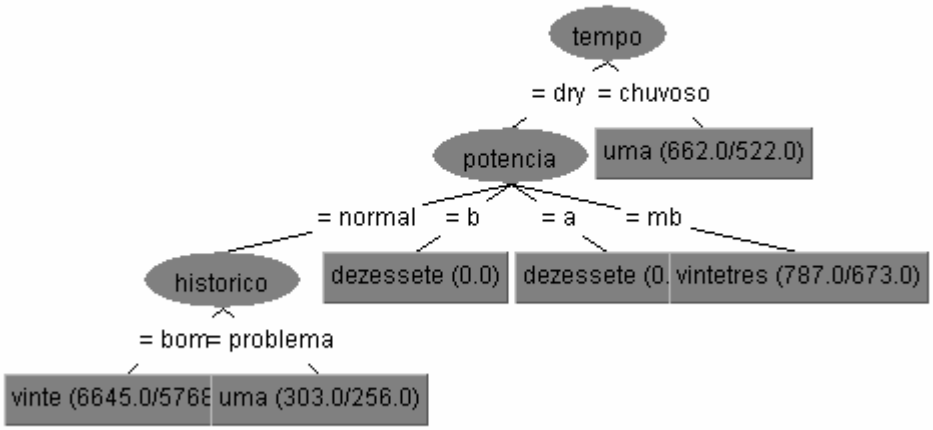
=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.11	0.116	0.12	0.11	0.115	uma
0.007	0.055	0.017	0.007	0.01	quatro
0.004	0.039	0.014	0.004	0.006	sete
0.007	0.053	0.018	0.007	0.01	dez
0.03	0.07	0.057	0.03	0.039	treze
0.135	0.283	0.064	0.135	0.087	dezessete
0.09	0.232	0.053	0.09	0.067	vinte
0.101	0.224	0.06	0.101	0.076	vintetres

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	<-- classified as
116	102	41	46	72	267	207	199	a = uma
153	7	47	58	68	252	240	225	b = quatro
146	48	4	55	69	290	214	224	c = sete
148	36	35	7	73	291	226	231	d = dez
149	57	28	49	31	293	223	220	e = treze
95	52	45	66	74	142	316	260	f = dezessete
82	56	52	57	75	345	95	288	g = vinte
81	56	42	60	82	344	279	106	h = vintetres

Agora iremos aplicar o algoritmo ID3 para visualizar a árvore:



Aplicando o algoritmo Bayes Net com os mesmos parâmetros utilizados para gerar a árvore de decisão, poderemos visualizar o seguinte gráfico mostrado na Figura 16:

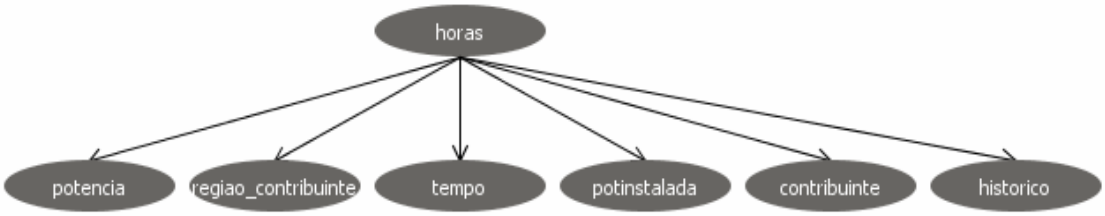


Figura 16 - Representação do Algoritmo Bayes Net

6.2 Análise referente a aplicação do algoritmo Apriori

Observa-se com base no histórico de demanda de energia elétrica dos contribuintes no período do mês de Outubro e a partir das regras de associação obtidas pelo algoritmo Apriori, que o perfil do contribuinte se apresenta com potência normal, não sendo significativo para responder a desvio de comportamento de carga.

6.3 Análise referente à aplicação do algoritmo ID3

1-Neste algoritmo foi detectado que o contribuinte n1 (região norte) apresentou no seu histórico uma anomalia, já que a potência solicitada se manteve muito baixa em relação à potência demandada em um horário de trabalho considerado normal. Desta forma, evidencia-se um motivo para se manter em alerta ou acompanhar os próximos históricos deste consumidor.

```
tempo = dry
| contribuinte = n1
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = mb: vintetres
```

2- Neste algoritmo foi detectado que o contribuinte s1 (região sul) apresentou no seu histórico uma anomalia, uma vez que a potência solicitada se manteve muito baixa em relação à potência demandada em um horário de trabalho continuo. Com base neste informação, tem-se um motivo para se manter em alerta, ou acompanhar os próximos históricos deste consumidor.

```
| contribuinte = s1
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: quatro
```

3- Neste algoritmo foi detectado que o contribuinte n2 (região norte) apresentou no seu histórico uma anomalia, visto que a potência solicitada se manteve muito baixa em relação à potência demandada em um horário de trabalho contínuo. Desta forma, verificamos um motivo para se manter em alerta, ou acompanhar os próximos históricos deste consumidor.

```
| contribuinte = n2
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vintetres
```

4- Neste algoritmo foi detectado que o contribuinte s2 (região sul) apresentou no seu histórico uma anomalia, levando em conta que a potência solicitada se manteve muito baixa em relação à potência demandada em um horário de trabalho contínuo. Assim, constatamos um motivo para se manter em alerta, ou acompanhar os próximos históricos deste consumidor.

| contribuinte = s2
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte

5- Neste algoritmo foi detectado que o contribuinte 03(região oeste) e ele3 (região leste) apresentou no seu histórico uma anomalia, pois a potência solicitada se manteve muito baixa em relação à potência demandada em um horário de trabalho contínuo. Desta forma temos um motivo para se manter em alerta, ou acompanhar os próximos históricos deste consumidor.

| contribuinte = o3
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: uma
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: vinte
| contribuinte = le3
| | potencia = normal
| | | historico = bom: dezessete
| | | historico = problema: vinte
| | potencia = b: null
| | potencia = a: null
| | potencia = mb: uma

Com base no que foi exposto, podemos concluir que estes consumidores apresentaram em seu histórico de demanda problemas e que, ao fazermos a mineração de dados, estes problemas foram identificados pelo algoritmo, facilitando a identificação rápida e segura.

A utilização destes algoritmos vem comprovar a sua eficiência na detecção de desvios de comportamento no tocante à demanda de energia elétrica.

Capitulo 7 – Conclusões

7 Conclusões Finais

O objetivo desta dissertação foi apresentar um método para detecção de desvio de comportamento de consumidores de energia elétrica, através de algoritmos de Classificação como o algoritmo Baysiano (BayesNet), algoritmos de árvore de Decisão (ID3 e J48) e algoritmos de Associação como Apriori, a partir de testes e da coleta dos resultados, traçando um comparativo entre os mesmos.

O método desenvolvido para detecção de desvio de comportamento foi aplicado a clientes horosazonais agrícolas da região da fronteira do estado do Rio Grande do Sul, onde encontram-se os maiores arrozeiros do país, sendo estes os clientes que representam um maior impacto à concessionária de energia elétrica.

Os dados referentes a estes clientes foram adquiridos de duas maneiras: por arquivos *DUMP* fornecidos pela concessionária e por arquivos de memória de massa adquiridos através de medidores instalados diretamente na propriedade destes clientes, também fornecidos pela concessionária, só que de forma bruta. Os dados posteriormente foram armazenados em um *Data Warehouse,* montado para o recebimento dos mesmos. Este *DW* foi bem estruturado e considerado adequado para o processo, assim como o *Data Mart* de treinamento, por ser baseado em uma base real de uma concessionária, com algumas alterações para agilizar mineração, o que temos como um dos fatores que contribuíram para o resultado satisfatório deste trabalho.

Com relação às ferramentas utilizadas, podemos salientar que o SGBD é um dos mais utilizados no mercado atual; a ferramenta *Case,* utilizada para a criação dos diagramas, também é uma das melhores e o software de Mineração Weka é o mais utilizado no meio acadêmico e um dos mais citados pelos especialistas na área, por ser uma ferramenta de distribuição livre com um ótimo poder de Mineração.

Os algoritmos selecionados para este método aliados à boa estrutura de banco de dados e às demais ferramentas utilizadas foram de fundamental importância para os resultados gerados.

Os algoritmos de árvore de decisão utilizados foram comparados, sendo que o algoritmo ID3 teve um resultado melhor como relação ao algoritmo J48, pois a poda utilizada neste algoritmo é maior e deixa de lado algumas evidências de desvio e comportamento de consumo de energia.

Os algoritmos foram aplicados e foi traçado um comparativo dos mesmos, sendo que todos resultaram em boas respostas de mineração e poucas diferenças entre os resultados.

As dificuldades encontradas referem-se à inexistência de dados relacionados a outras tipologias de clientes para que fosse feita uma estimativa de que classe de consumo e em que classe social é mais provável a sinalização de desvio de comportamento.

A elaboração deste trabalho teve como principais metas ressaltar a aplicação da informática e a sua importância, a utilização da Tecnologia da Informação no desenvolvimento de um banco e a aplicação da mineração de dados utilizando a ferramenta Weka para identificar determinadas anomalias na demanda de consumidores de energia elétrica de um sistema de dados de uma concessionária. A partir disso, buscamos identificar padrões ou tendências de cada região no tocante a anomalias de demanda de energia elétrica e sua influência na concessionária do sistema elétrico. Ademais, procuramos destacar de que maneira a obtenção destes padrões pode favorecer a identificação de pontos críticos, com a integração do banco de dados da concessionária do sistema elétrico e a Mineração de Dados na descoberta destes desvios de comportamentos na demanda informada pelo cliente e a demanda registrada no banco de dados da concessionária.

Neste estudo foi feita uma explanação da importância de tal controle, utilizando como referência as normas recomendadas pelo programa Sistema de energia elétrica da concessionária e as informações armazenadas no banco de dados, o que permitiu fazermos a aplicação de algoritmos de Mineração de dados para a descoberta de comportamentos ou tendências de uma determinada região ou consumidor de energia elétrica, quanto à demanda presente, bem como o seu histórico de demanda energética, analisando a apresentação de diferenças consideráveis na sua demanda.

Recomendações:

- Inserir dados de clientes com outras tipologias como Clientes residenciais, pois o *Data WareHouse* desenvolvido para este estudo de caso suporta estas tipologias de cliente, basta a construção de um *Data Marts* com as tabelas envolvidas neste processo de Mineração, já que não seriam usadas as mesmas tabelas partindo do principio que estes clientes não possuem Medidor.

- Através da inserção de outras tipologias, pode-se fazer um comparativo entre as várias tipologias de clientes e, após a mineração, utilizando métodos estatísticos para saber em que região, em que época do ano ocorrem mais desvios de comportamento de consumo de energia, bem como em qual classe acontecem com mais freqüência estes desvios e o impacto que os mesmos causam à concessionária.

- Tendo em vista os clientes industriais, por exemplo, basta a inserção dos dados no *Data WareHouse* para que os mesmos sejam Minerados.

- A utilização de outros algoritmos de classificação e agrupamento, para se fazer um maior comparativo de resultados.

- Utilização de uma interface gráfica web para melhor visualizar os resultados com a utilização e gráficos on-line.

Bibliografia

AUSTERN, M. H. **"Generic Programming and STL. Using and Extending the C++ Standart Template Library**". Canada 1999.

BOSE, I. e MAHAPATRA, R., **"Business Data Mining – A Machine Learning perspective"**: Information e Management 39, pp. 211-225, 2001.

BRADZIL, P. B., **"Construção de Modelos de Decisão a partir de dados",** 1999. Disponível em: < http://www.liacc.up.pt/~pbrazdil/Ensino/ML/DecTrees.html >. Acesso em 02/05/2007.

BRAGA, Antônio P., CARVALHO, André P. L. F., LUDERMIR, Teresa B., **"Redes Neurais Artificiais: Teoria e Aplicações".**Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora S.A., 2000.

BRAGA, Luis Paulo V., **"Introdução à Mineração de Dados",** 2ª Edição revisada e ampliada. Rio de Janeiro: E - Papers Serviços Editoriais, 2005.

BRAGA, Rosena T., **"Engenharia Reversa ou Reengenharia".** Disponível em

< http://www.inf.ufpr.br/silvia/ES/reengenharia/reengenharia.pdf > . Acesso em 27/09/2007.

BUENO. André D., "**Introdução ao Processamento Paralelo e ao Uso de Clusters de orkstacions em Sistemas GNU/LINUX, Parte I: Filosofia",** 2002. Disponível em: <http://www.rau-tu.unicamp.br/nou-rau/softwarelivre/document/?view=83>. Acesso em 10/10/2007.

CABENA, P. et al., **"Discovering Data Mining : From concept to implementation"**. Upper Saddle River – NJ: Prentice Hall, 1998.

CABRAL, J. E.; PINTO, J. O. P.; GONTIJO, E. M.; REIS, J.,**''Rough Sets Based Fraud Detection in Electrical Energy Consumers"**. WSEAS International Conference on Mathematics and Computers in Physics, Cancun, México, Abr. 2004.

CARDOSO, Cristina C., **"Modelo de Previsão Baseado em Agrupamento e Base de Regras Nebulosas"**. Campinas.

CARVALHO, Juliano V., SAMPAIO, Marcus C., MONGOVI, Giuseppe, "Utilização de Técnicas de "Data Mining" para o Reconhecimento de Caracteres Manuscritos". Universidade

Federal da Paraíba. Disponível em <http://www.inf.ufsc.br/sbbd99/anais/SBBD-Completo/20.pdf>. Acesso em 03/03/2008.

CARVALHO, L. A.V., **"Data Mining: A mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração"**. São Paulo: Editora Érica, 2001.

CASSOL, Ivo, SCHNEIDER, Junior Flávio, BRONDANI, Gilberto, MADRUGA, Sérgio Rossi, ZANELA, José, **"Diagnóstico Sócio-Econômico do Entorno de Santa Maria-RS",** 2005. Disponível em <www.planalto.gov.br/sri/CooperacaoInternacional/ Docs_CoopItaliana/DiagnosticoSantaMaria.doc>. Acesso em 30/11/2007.

CHENG, Jie; GREINER, Russell. "**Learning Bayesian Belief Network Classifiers: Algorithms and System. Edmonton**", Canadá: Department of Computing Science, University Alberta, 1999. Disponível em <www.cs.ualberta.ca/~jcheng/Doc/cscsi.pdf>. Acesso em 15/10/2007.

COREY, M., et al.(Eds), **"Oracle Data Warehouse 8i"**, Oracle Press, 2001. COUTINHO, Fernando V., **"Data Mining".** Disponível em <http://www.dwbrasil.com.br/html/dmining .html>. Brasília, 2003. Acesso em 30/04/2007.

DIN - Departamento de Informática - UEM - Universidade Estadual de Maringá. GSI - Grupo de Sistemas Inteligentes - Mineração de Dados, 1998. Disponível em: <http://www.din.uem.br/ia/mineracao/tecnologia/ferramentas.html>. Acesso em 02/05/2007.

FANDERUFF, D., **"Dominando o Oracle 9i",** São Paulo: Person Education do Brasil, 2003.

FAYYAD, U. M.; PIATESKY-SHAPIRO, G.; SMYTH, P. "From Data Mining to Knowledge Discovery: An Overview". In Advances in Knowledges Discovery and Data Mining, AAAI Press, 1996.

FERNANDES, Marcelo A. C.; NETO, Adrião D. D.; BEZERRA, João B. **"Aplicação das Redes RBF na Detecção Inteligente de Sinais Digitais".** IV Congresso Brasileiro de Redes Neurais, São José dos Campos, 1999.

GAMA, J., **"Árvores de Decisão"**, 2000. Disponível em: <http://www.liacc.up.pt /~jgama/Mestrado/ECD1/Arvores.html> Acesso em: 03/05/2007.

GARCIA, S. C., **"O uso de Árvores de Decisão na descoberta de conhecimento na Área da Saúde"**. SEMANA ACADÊMICA, 2000.Universidade Federal do Rio Grande do Sul.

GARDNER, S. R., **"Building the Data Warehouse"**, Communications of the ACM,1998.

GOEBEL, M.; GRUENWALD, L., **"A Survey of Data Mining and Knowledge Discovery Software Tools"**. New Zealand: SIGKDD Explorations, 1999.

GOLDSHMIDT, Ronaldo; PASSOS, Emmanuel, **"Data Mining: Um Guia Prático".** Rio de Janeiro: Elsevier, 2005.

GPGE Grupo de Pesquisa de Gestão de Energia, **"Relatório Final Gestão de Energia em Programas Anuais de Eficiência Energética e Promoção do Uso Racional de Energia".** Porto Alegre, 2007.

GPPD Grupo de Processamento Paraleilo e Distribuído. 1998. Disponível < http://www.inf.ufrgs.br/gpesquisa/procpar/intro.html >. Acesso em: 10/10/2007.

HAN, Jiawei; KAMBER, Micheline, **"Data Mining Concepts and Techniques"**, San Francisco, EUA: Morgan Kaufmann, 2001.

HARRISON, T. H., **"Intranet Data Warehouse"**. [s.l.]: Ed. Berkeley, 1998.

HRUSCHKA, Estevam R.; TEIXEIRA, W., **"Propagação de Crença em Redes Bayesianas",** Brasília: UnB, 1997 (Relatório de Pesquisa CIC/UNB – 02/97).

INGARGIOLA, G. , "**Building Classification Models: ID3 and C4.5.",** 1996. Disponível em: http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html> Acesso em 03/05/07.

INMON, Willian H., **"Como Construir o Data Warehouse"**, Rio de Janeiro: Editora Campus, 1997.

JANNUZZI, Gilberto M.; SWISHER, Joel N. P., **"Planejamento Integrado de Recursos Energéticos"**, Campinas: Editora Autores Associados, 1997.

KLÖSGEN, W.; ZYTKOW, Jan M., **"Handbook of Data Mining and Knowledge Discovery"**, New York: Oxford University Press, 2002.

KRÓSE, B. J. A.; VAN DER SMAGT, P. P., **"An Introduction to Neural Networks"**. Amsterdam, University of Amsterdam, 1993.

LANGE, Luis Carlos; "Mineração de Dados em Sistema Eficiente de Iluminação Pública incluindo Parâmetros Sócio-Comportamentais". Porto Alegre, Agosto de 2007.

LEMOS, Elaine P., **"Análise de Crédito Bancário com o uso de *Data Mining*: Redes Neurais e Árvore de Decisão"**. Dissertação (Mestrado em Ciências) – Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Área de concentração em Programação Matemática, Universidade Federal do Paraná, 2003.

LOPES, Paulo A., **"Probabilidades e Estatística"**. Rio de Janeiro: Reichmann & Affonso Editores, 2001.

LUGER, George F., **"Inteligência Artificial – Estruturas e Estratégias para a resolução de problemas complexos",** Porto Alegre: Bookmann, 2004.

MELLO, Carlos Eduardo R., SILVA, Geraldo Zimbrão, SOUZA, Jano M., **"Extensão do WEKA para Métodos de Agrupamento com Restrição de Contigüidade",** em **"IX Brazilian Symposium on GeoInformatics",** Campos do Jordão, Brasil**,** INPE, p. 277-282. Novembro 2007.

MELLO, Luis Cesar. "**Um Assistente de Feedback para o Serviço de Filtragem do Software Direto"**. Dissertação (Mestrado em Ciências da Computação) – Programa de Pós-Graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

NAVEGA, Sergio, **"Princípios Essenciais do Data Mining"**, Publicado nos Anais do Infoimagem 2002. Disponível em <http://www.intelliwise.com/reports/i2002.htm> Acesso em: 08/03/2007.

PATRICIO, Cristin Mara M. M., **"Detecção de Fraude ou Erro de Medição em Grandes Consumidores de Energia Elétrica Utilizando Rough Sets Baseado em Dados Históricos e em Dados em Tempo Real"**, Campo Grande, Julho de 2005.

PEARL, Judea, **"Probabilistic Reasoning in Intelligent Systems".** San Mateo, EUA: Morgan Kaufman, 1988.

PETERMANN, Rafael J., **"Modelo de Mineração de Dados para Classificação de Clientes e, Telecomunicações"**, Porto Alegre Outubro de 2006.

POZO, Aurora Trinidad Ramirez, Universidade Federal do Paraná, Departamento de Informática. Disponível em < http://www.inf.ufpr.br/aurora/> Acesso em: 20/12/2007.

QUILAN, J. R. **"Induction of Decision Tree : Machine learning"**, 1986.

REIS, J.; CONTIJO, E. M.; MANIZA, E.; CABRAL, J. E.; PINTO, J. O. P., **"Fraud Indentification in Eletricity Company Customers using Decision Tree"**, in 2004 IEEE International Conference on Systems, Man and Cybernetics, 2004.

REZENDE, S. O. ; PUGLIESI, J. B. ; MELANDA, E. A. ; PAULA, M. F. , **"Mineração de Dados. In: Solange Oliveira Rezende. (Org.). Sistemas Inteligentes - Fundamentos e Aplicações**". Barueri, SP: Editora Manole Ltda, 2003.

SAUER, Ildo, **"Metodologia de Análise de Cadastro de Concessionárias"**, São Paulo, 2000.

SILBERSCHATZ, Abraham; KORTH, Henry F., SUDARSHAN S., **"Sistemas de Banco de Dados"**, São Paulo: Makron Books, 1999.

SILVA, Marcelino Pereira dos Santos, **"Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka"**. Mossoró, RN, 2005. Universidade do Estado do Rio Grande do Norte (UERN).

SIMEÃO, J. de M., **"Bancos de Dados Geográficos e Redes Neurais Artificiais: Tecnologias de Apoio à Gestão de Território"**. São Paulo, 1999. Universidade de São Paulo.

SOARES, Silviane L., **"Aplicação de Técnicas de Mineração de Dados em Gestão de Sistemas de Energia Elétrica"**, Porto Alegre, Março de 2005.

STEINER, M. T. A., **"Redes Neurais"**. Universidade Federal do Paraná. Métodos Numéricos em Engenharia - Pesquisa Operacional, 1999.

TAFNER, M. A., **"Redes Neurais Artificiais: Aprendizado e Plasticidade"**. Revista Cérebro & Mente, 2(5), março/maio. 1998.

VINHAS, Lúbia; QUEIROZ, Gilberto R.; FERREIRA, Karine R.; CÂMARA, Gilberto; PAIVA, João Argemiro C., **"Programação Genérica Aplicada a Algoritmos Geográficos"**. São José dos Campos, SP, 2000. INPE – Instituto Nacional de Pesquisas Espaciais.

WILEY, John; SONS**, "Data Mining Methods and Models"**, New Jersey, 2006.

WITTEN, Ian H.; FRANK, Eib**."Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations".** San Francisco, EUA: Morgan Kaufmann Publishers Inc., 2000.

ZANUSSO, Maria Bernadete, Departamento de Computação e Estatística - DCT Universidade Federal de Mato Grosso do Sul. Disponível em < http://www.dct.ufms.br/~mzanusso/ensino/ensino.htm >Acesso em: 09/01/2008.

Glossário

Índices Pluviométricos no Estado

Em BARATTO (1992) *apud* CASSOL et tal é dito que anualmente no estado do Rio Grande do Sul a média de precipitação é de 1600mm e uniformemente distribuídas pelas estações dos ano, sendo que a média de chuva no verão é em média 400mm, no outono a média oscila entre 400mm e 450mm, o inverno a média fica em torno de 400mm e já a primavera é responsável por 450mm, segundo o diagnóstico os meses mais chuvosos ou seja com maior índice pluviométrico são os meses de abril, maio, setembro e outubro.

As anormalidades climáticas que vem ocorrendo nos últimos anos devem-se ao fenômeno do El Niño. Fenômeno que causa elevação das temperaturas em determinados períodos na superfície líquida do oceano Pacífico, provocando assim um aumento da temperatura nas massas de ar responsáveis pela dinâmica do sistema atmosférico geral do globo terrestre, e isso fazendo com que as condições de tempo na maior parte do planeta tenham alterações no seu clima e conseqüente mente nas chuvas. No estado do Rio Grande do Sul, este fato acarreta no estacionamento de frentes, que geralmente entram pela Argentina, isso faz com que aumente os índices pluviométricos [CASSOL 2005].

Estes fatos devem ser levados em consideração pela concessionária, pois se o Cliente tem uma redução no seu consumo de energia, isso tem impacto na estratificação de sua curva típica e também faz com que sua correlação diminua. Isto pode fazer com que a concessionária sinalize que este cliente possa ser um possível fraudador. Então antes da concessionária mandar um especialista para fiscalizar este clientes esse fato deve ser levado em consideração.

Banco de Dados

Um sistema gerenciador de banco de dados (SGBD) é um conjunto de dados inter-relacionados e uma coleção de programas para acessar estes dados. O princípio básico de um SGBD é proporcionar um ambiente conveniente e eficaz para recuperação e armazenamento das informações.

Como as informações vindas da concessionária estão organizadas em bancos de dados nada mais conveniente do que estudá-los, para um melhor entendimento de como funcionam esses bancos e como se relacionam suas tabelas.

Um sistema de banco de dados é projetado para armazenar grandes quantidades de informação. O gerenciamento de informações implica a definição das estruturas e armazenamento destas informações e o fornecimento de mecanismos para sua manipulação. Além disso, o sistema de banco de dados precisa proporcionar segurança ao armazenamento de informações, diante de falhas do sistema ou acesso não autorizado. [SILBERSCHATZ et al, 1999].

O sistema de banco de dados tem como objetivo principal proporcionar aos usuários uma visão abstrata dos dados (visando a facilitar a interação dos usuários com o sistema) isto é, o sistema esconde determinados detalhes de como os dados são mantidos e como estes estão armazenados. Isto é feito por três níveis de abstração:

Nível Físico: é o mais baixo dos níveis de abstração que descreve como estes dados estão de fato armazenados.

Nível Lógico: este é o nível intermediário de abstração que descrevem quais dados estão armazenados o banco e quais os inter-relacionamentos entre eles.

Nível de Visão: o mais alto nível de abstração descreve apenas parte do banco de dados. Cada usuário normalmente utiliza parte do banco, por isso o nível de visão é restrito, pois cada usuário o vê de um ponto de vista diferente ou acessa o banco de uma maneira.

Anexo 1

Os Anexos 2 e 3 forma retirados do Relatório Final do *software* SIADAGE (GPGE, 2007).

Aquisição dos Dados

> Para a aquisição dos dados contidos nos arquivos públicos[6] foi desenvolvida uma ferramenta pelo Grupo de Pesquisa de Gestão de Energia (GPGE) da Pontifícia Universidade Católica do Rio Grande do Sul. A ferramenta segue as condições da NBR 14522 para leitura dos arquivos. Trata-se de um processo manual, isto é, os arquivos têm que ser selecionados para que a ferramenta proceda a importação.
>
> É importante ressaltar que a ferramenta de importação dos arquivos públicos verifica cada um deles levando em consideração as seguintes condições:
>
> se o nome do arquivo tem a quantidade de caracteres correta;
>
> se o número do medidor do cliente consta na tabela SGC_ME_FORNECIMENTOS (disponibilizada pela concessionária por dump).
>
> Caso alguma dessas duas situações não seja validada, o arquivo não é importado, portanto as medições não são inseridas no banco de dados de medições. O Manual do Sistema apresenta as orientações para utilização da ferramenta de importação dos arquivos públicos.
>
> Cada arquivo público possui o número do medidor que o produziu, e através deste registro é possível vincular as informações nele contidas ao cliente. Este vínculo é possível, pois consta na tabela SGC_ME_FORNECIMENTOS o número do medidor.

A Figura 17 ilustra o processo de aquisição, geração e posterior consulta das curvas de carga típicas executado pelo SIADAGE:

[6] É um arquivo texto padronizado pela NBR 14522 de 2000. Esta norma define o padrão de intercâmbio de informações no sistema de medição de energia elétrica brasileiro, de forma a se alcançar a compatibilidade entre os sistemas e equipamentos de medição de energia elétrica de diferentes procedências e fabricantes.

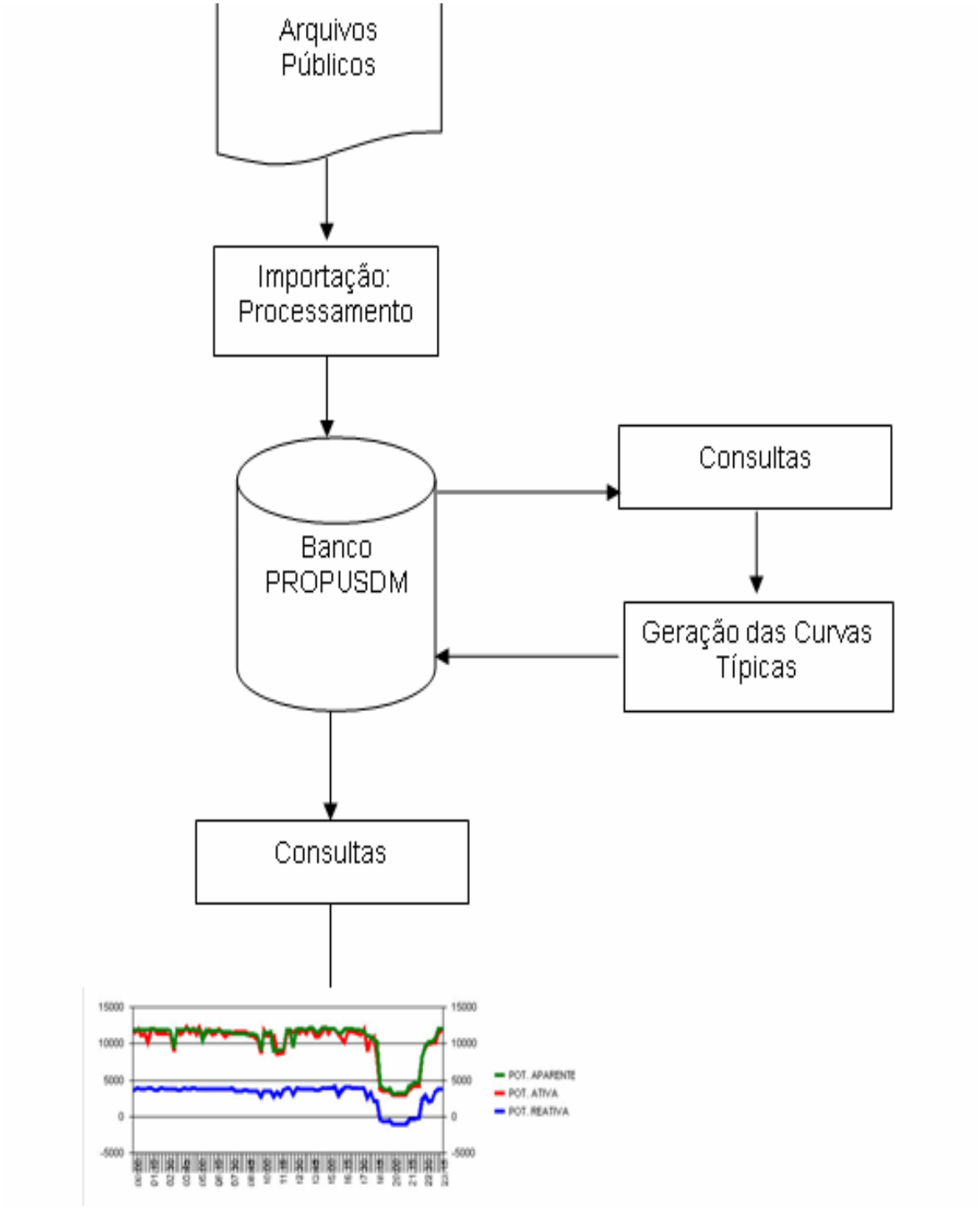


Figura 17 - Processo de Aquisição, Criação e Consulta das Curvas de Carga.

O tratamento da curva de carga dos clientes é realizado pelo SIDAGE[7], que transforma os dados de medição armazenados na memória de massa dos registradores em unidades de potência ativa (kW) ou reativa (kVar). É importante ressaltar novamente que os arquivos que contém os dados de medição devem estar no formato público proposto pela ABNT. O período de integração utilizado é de 15 minutos, assim, cada dia será representado por 96 pontos de carga. Contudo, o SIADAGE está preparado para utilizar qualquer período de integração definido pelo usuário. Constam do Manual do Sistema as informações completas de utilização do módulo Curvas Típicas de Carga do SIADAGE.

Importar Arquivo Público de Clientes

Para importar os arquivos públicos dos clientes horo-sazonais proceda da seguinte forma:

no menu principal clique em Curvas Típicas;

selecione Abre Arquivo Público;

na caixa de dialogo Abrir Arquivo Figura 18 selecione na janela da direita a pasta onde se encontra o arquivo público desejado, selecione o arquivo público desejado;

em seguida clique no botão Abrir.

[7] O Sistema de Apoio a Decisão em Gestão de Energia (SIADAGE) é um sistema geo-referenciado que fornece aos coordenadores, gerentes de planejamento e eficiência energética uma ferramenta de decisão para estratégias de gestão de energia dentro da área de concessão da concessionária de energia.

A tomada de decisão acontece através de simulações ambientadas no mercado consumidor e na rede distribuição da concessionária, de forma integrada. Este processo permite simular ações de gestão de energia, com base em experiências de campo armazenadas no seu banco de dados resultando informações, técnicas e econômicas, das ações de eficiência energética.

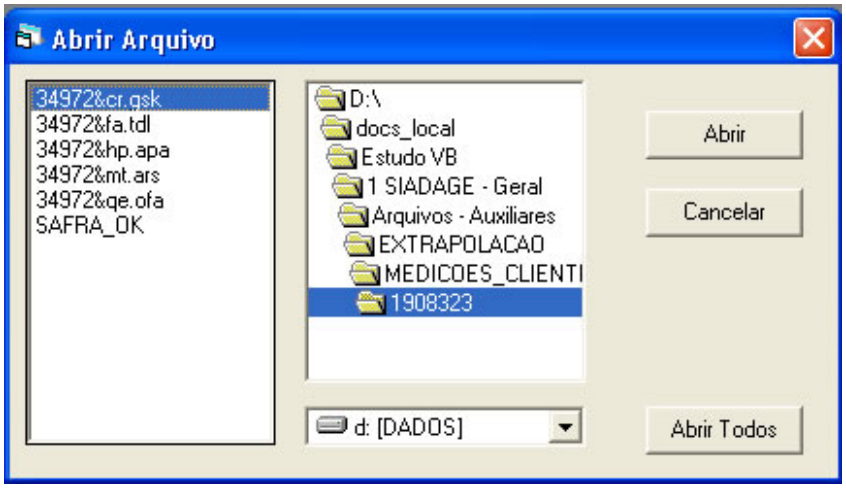


Figura 18 – Seleção de Arquivos Públicos de Clientes.

Pode-se visualizar o arquivo público – depois de importado – com diferentes intervalos de integração, bastando para isso clicar no botão Integração e na caixa de diálogo digite o intervalo de integração desejado Figura 19 e Figura 20.

Para gravar as informações contidas no arquivo público importado clique no botão Gravar Banco, terminado o processo de gravação uma mensagem de aviso é exibida.

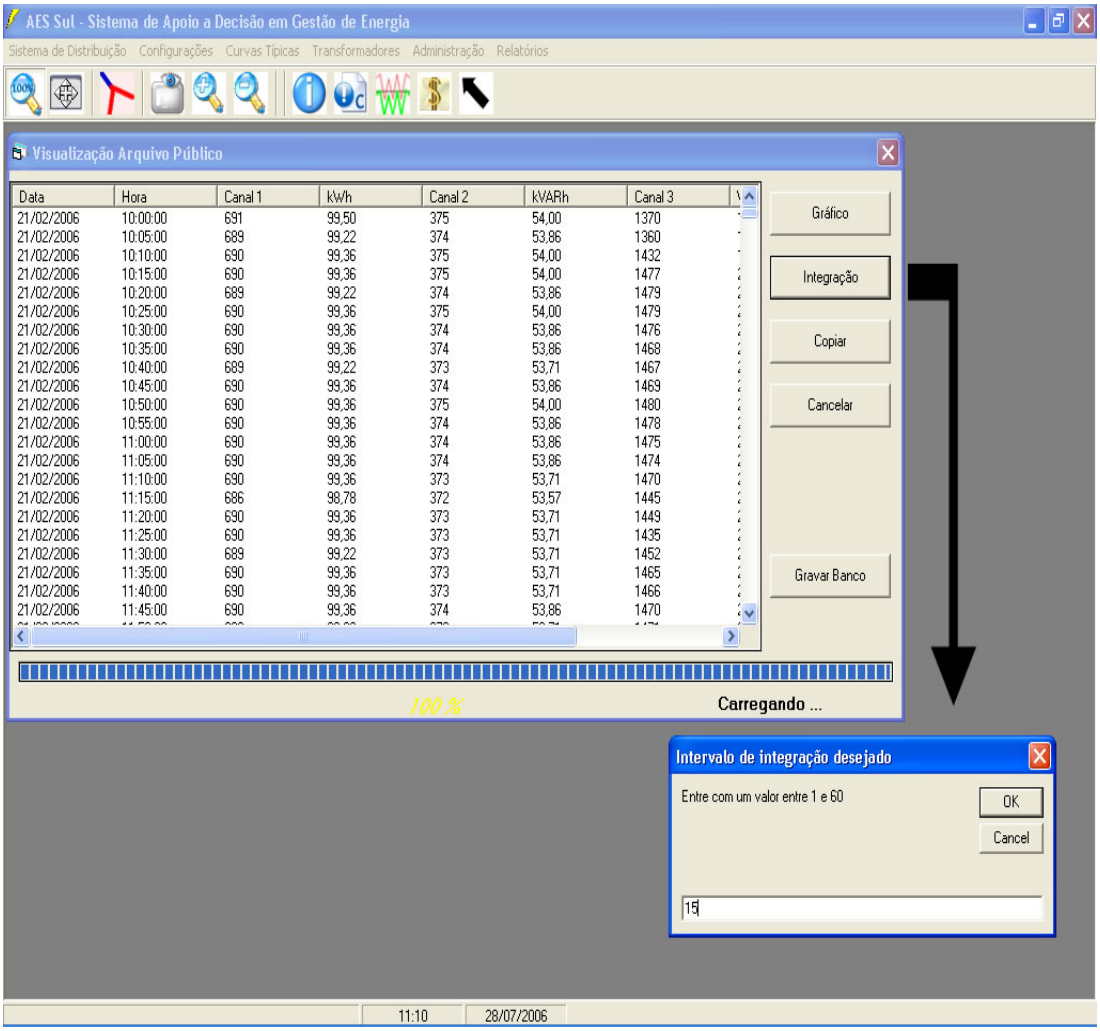


Figura 19 – Visualização do Arquivo Público do Cliente

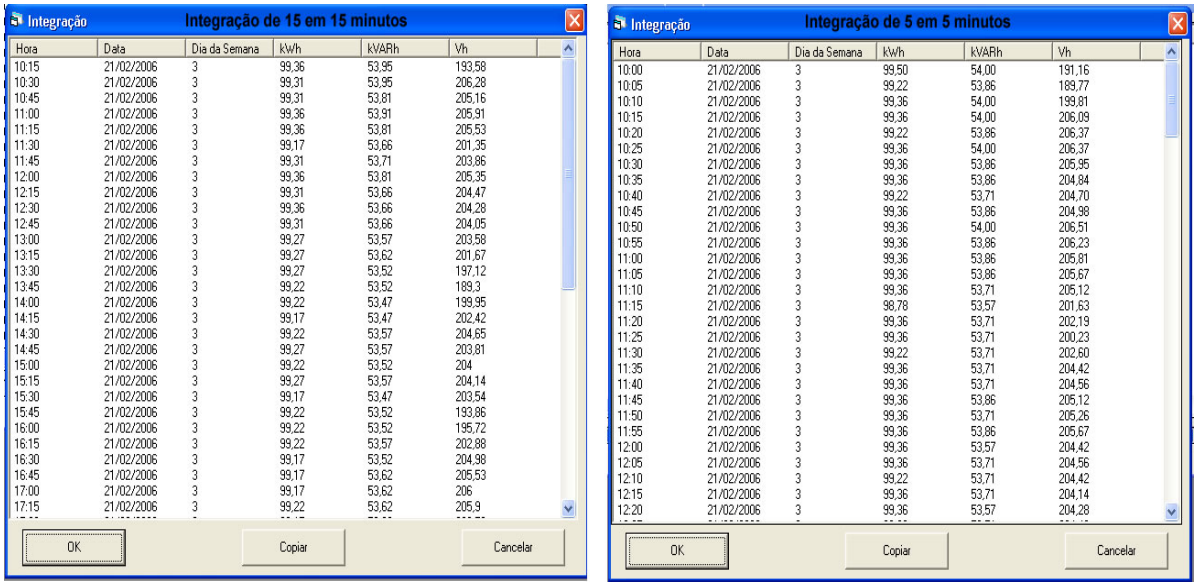


Figura 20 – Integrações de 5 em 5 e de 15 em 15 minutos.

É importante ressaltar que o SIADAGE verifica se o arquivo tem o comprimento de caracteres correto e se o número do medidor do cliente está cadastrado no banco de dados da concessionária. Caso alguma dessas duas situações não seja validada, o arquivo não é inserido no banco de dados de medições e uma mensagem de advertência será exibida.

O software foi desenvolvido com o objetivo de fazer a leitura de arquivos no formato público e posterior alimentação de um banco de dados racle com as informações lidas.

Para o desenvolvimento do software utilizou-se o Visual Basic 6.0 conectado ao banco de dados instalado no servidor PROPUS.

Tabelas do Banco de Medições

A seguir, segue uma breve descrição das tabelas utilizadas nos processos de carregamento das leituras, consulta e geração das curvas típicas. Para maiores detalhes dessas tabelas, deve-se consultar o tópico Estrutura do banco de dados PROPUSDM.

O módulo de carga, geração e consulta das curvas típicas de carga utiliza as seguintes tabelas do banco PROPUSDM:

Tabela 13 – Tabelas do Banco PROPUSDM

Nome da Tabela	Descrição
SGC_RESUMO	A aplicação utiliza esta tabela para relacionar o arquivo público de medição com os dados cadastrais do cliente
CLI_CADASTRO	Nesta tabela a aplicação salva todos os clientes que possuem medição cadastrada. É uma tabela resumo onde sempre que o arquivo público de um cliente é carregado, as informações do NUC e classificação são gravadas. Contém NUCs repetidos, pois o identificador único é o ID_CLI_CADASTRO.
TARIFAS	Armazena as informações das tarifas disponíveis para cadastro.
TENSAO_FORNEC	Armazena as informações das tensões de fornecimento disponíveis para cadastro.
CNAE	Armazena as informações das atividades econômicas disponíveis

Nome da Tabela	Descrição
	para cadastro.
APMEDIDA_AP	Faz o vínculo entre a tabela de clientes e a tabela de medições. Armazena o número de série do medidor do cliente.
MEDICOES	Armazena as informações de medição do cliente.
MEDICOES_VAL	Armazena as informações da memória de massa dos arquivos públicos de cada cliente.
ARQ_MEDICOES	Armazena o arquivo público de medição propriamente dito em formato binário e compactado (ZIP). Além disso, guarda o nome do arquivo de leitura de cada cliente já cadastrado para evitar redundância de dados.
ESTAÇÕES	Armazena as informações das estações do ano disponíveis para cadastro.
FERIADOS_MUNICIPAIS	Armazena as informações dos feriados municipais de cada cidade da área de concessão da concessionária
CIDADES	Armazena as informações de todas as cidades presentes na área de concessão.
FERIADOS_NACIONAIS	Armazena as informações dos feriados nacionais.
MEDICOES_ALIMENTADOR	Armazena todas as medições de potência dos alimentadores concessionária.
MEDICOES_TRAFO	Armazena todas as medições de potência dos transformadores de potência da concessionária.

Padronização – medição de energia elétrica

Este padrão define os seguintes itens:

1. Comunicação convencional medidor/leitor

2. Comunicação direcional medidor/leitor

3. Comunicação remota com medidores (síncrona)

4. Arquivo de dados lidos (formato público)

5. Arquivo de parâmetros para medidor

6. Arquivo de programa para medidor

7. Saída do consumidor

8. Comunicação leitora direcional/computador

9. Comunicação leitora/computador

10. Comunicação remota com medidores (assíncrona)

11. Códidos e grandezas do mostrador

12. Códigos de comandos e operação da leitora

Formato Público

A padronização para o nome do arquivo no formato público é a seguinte:

NNNNN&XX.XXX

ONDE:

NNNNN são os 5 dígitos menos significativos do número de série do medidor XX.XXX é o resultado do cálculo SS+MMx60+HHx3600+(DD)x24x3600+(MM)x31x24x3600* transformado para a base 20, onde A=0 até T=19.

*Segundo, minuto, horas, dia e mês do arquivo gerado são relativos à hora e data da leitura.

O Arquivo Público é um arquivo texto que pode ser dividido basicamente em dois blocos principais:

Bloco Inicial (CABEÇALHO) – Valores Absolutos com parâmetros de configuração

Bloco de Acumuladores (MEMÓRIA DE MASSA) com os registros de medição

Bloco Inicial - Cabeçalho

4159982416100115454424050004154500240500154500240500160000130400000000000000000000000000000000000000001800180018002100210021000000000000000000031968000000001500000000000000000000000000000000000000000000000000000000000000000000000000000000000000000004000001000004000001000004000001000001 (300 Caracteres por bloco) – Fragmento inicial

Posição	Tamanho	Descrição
0001	08	Número de série do registrador
0009	06	Número de série do leitor
0015	02	Hora atual
0017	02	Minuto atual
0019	02	Segundo atual
....	....	....
2378	23	Não utilizados

Características Principais (Resumo):

Os valores numéricos são apresentados no formato MSB anterior ao LSB;

O Bloco inicial apresenta até 8 blocos de 300 caracteres cada sem quebra de linha (CF ou LF);

No algoritmo implementado o cabeçalho foi lido de forma integral, pois o número de caracteres é fixo e demarcado. Item por Item.

Bloco de Acumuladores

CONT001

000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000 (288 CARACTERES POR BLOCO)

24 medições de canal por bloco.

(Independente do intervalo de Integração)

Canal 1 – 24 registros

Canal 2 – 24 registros

Canal 3 – 24 registros

Posição	Tamanho	Descrição
0001	04	'CONT' se desde a última reposição de demanda
		'SALV' se relativos ao à última reposição de demanda
0005	02	Reservados
0007	02	Número do bloco

Formatação dos Blocos de Memória de Massa:

contador real = contador arquivo   Se contador arquivo < 2048

contador real = 2048 - contador arquivo   Se contador arquivo > 2048

OBS: Para uma medição de 30 dias e intervalo de integração de 5 minutos o contador apresenta numeração máxima de 440 blocos.

Se Número do Bloco for '01', '04', '07' etc.,

Posição   Tamanho   Descrição

0009	04	contador do canal 1 enésimo intervalo
0013	04	contador do canal 2 enésimo intervalo
0017	04	contador do canal 3 enésimo intervalo
0021	04	contador do canal 1 enésimo + 1 intervalo
0025	04	contador do canal 2 enésimo + 1 intervalo
0029	04	contador do canal 3 enésimo + 1 intervalo

OBS: Cada bloco contador possui um ordenamento dos canais.

Se Número do Bloco for '02', '05', '08' etc.,

0009	04	contador do canal 2 enésimo intervalo
0013	04	contador do canal 3 enésimo intervalo
0017	04	contador do canal 1 enésimo + 1 intervalo
0021	04	contador do canal 2 enésimo + 1 intervalo
0025	04	contador do canal 3 enésimo + 1 intervalo

Se Número do Bloco for '03', '06', '09' etc.,

0009	04	contador do canal 3 enésimo intervalo
0013	04	contador do canal 1 enésimo + 1 intervalo
0017	04	contador do canal 2 enésimo + 1 intervalo
0021	04	contador do canal 3 enésimo + 1 intervalo
0029	04	contador do canal 1 enésimo + 2 intervalo

001: "CONT" se for leitura do período atual

"SALV" se for leitura do período anterior

005: "001" - número do bloco de dados da memória de massa

008: - 5 caracteres reservados para uso futuro

013: - 4 caracteres que indicam o valor do canal 1 em seu 1° período.

017: - 4 caracteres que indicam o valor do canal 2 em seu 1° período.

021: - 4 caracteres que indicam o valor do canal 3 em seu 1° período.

024: - 4 caracteres que indicam o valor do canal 1 em seu 2° período.

029: - 4 caracteres que indicam o valor do canal 2 em seu 2° período.

033: - 4 caracteres que indicam o valor do canal 3 em seu 2° período.

...

289: - 4 caracteres que indicam o valor do canal 1 em seu 24° período.

293: - 4 caracteres que indicam o valor do canal 2 em seu 24° período.

297: - 4 caracteres que indicam o valor do canal 3 em seu 24° período.

Constantes de medição

Os valores que aparecem nos blocos são multiplicados por uma constante do medidor para resultar no valor absoluto (real) de leitura. Dados do cabeçalho.

Posição	Tamanho	Descrição
0261	06	Numerador da constante de multiplicação do canal 1
0267	06	Denominador da constante de multiplicação do canal 1
0273	06	Numerador da constante de multiplicação do canal 2
0279	06	Denominador da constante de multiplicação do canal 2
0285	06	Numerador da constante de multiplicação do canal 3

0291 06 Denominador da constante de multiplicação do canal 3

Cálculo das constantes multiplicativas:

K1 = (NUM1 / DEN1) * (PULSOS)

K2 = (NUM2 / DEN2) * (PULSOS)

K3 = (NUM3 / DEN3) * (PULSOS)

PULSOS: Número de pulsos que preenchem o intervalo de 1 hora segundo o intervalo de integração da memória de massa.

Exemplo:

Intervalo de Integração = 5 minutos

Valor do canal 1 = 0007 NUM1 = 400 DEN1 = 100

K1 = (400 / 100) * (60 / 5) = 48

Valor absoluto = 7 * 48 = 336 (real)

Valor do canal 2 = 0031 NUM2 = 400 DEN2 = 100

K2 = (400 / 100) * (60 / 5) = 48

Valor absoluto = 31 * 48 = 1488 (real)

Valor do canal 3 = 0000 NUM3 = 400 DEN3 = 100

K3 = (400 / 100) * (60 / 5) = 48

Valor absoluto = 0 * 48 = 0 (real)

Características Principais (Resumo):

Os valores numéricos são apresentados no formato MSB anterior ao LSB;

O Bloco de acumuladores apresenta um número indefinido de sub-blocos “CONT” ou “SALV”, dependendo do período de medição e tempo de integração dos registros;

No algoritmo implementado a memória de massa foi lida de uma em uma linha. (Número variável de caracteres) – 288 CARACTERES POR LINHA.

Anexo 2

Geração da Curvas Típicas de Carga dos Clientes

As curvas típicas de carga são geradas para os clientes mais característicos ou para grupos de clientes similares. A representação das curvas é feita de forma normalizada (p.u.) em função dos diferentes valores de potência entre os clientes. Concluí-se então que curvas típicas de carga refletem um processo de variação de consumo de energia elétrica em um determinado período de tempo (usualmente diário), para consumidores individualizados ou grupos de consumidores semelhantes (empresas de ramos indústrias específicos, etc.), sem a estimação do valor real da carga.

Outras características são levadas em consideração pelo SIADAGE para a construção das curvas típicas, são elas:

- os diferentes períodos do ano;
- a variação de consumo para diferentes dias da semana e dias atípicos (dias úteis, sábados, domingos e feriados);
- as condições climáticas e as condições sócio-econômicas das regiões em análise.

O SIADAGE para a identificação de curvas de carga para cada cliente ou um grupo de clientes, executa o seguinte procedimento:

1. seleciona os três dias mais representativos, ou seja, um dia útil, um sábado e um domingo;
2. identifica as curvas de carga para os dias selecionados, representando o comportamento típico;
3. associa as curvas ao cliente ou grupo de cliente.

O procedimento descrito acima é ilustrado na Figura 21:

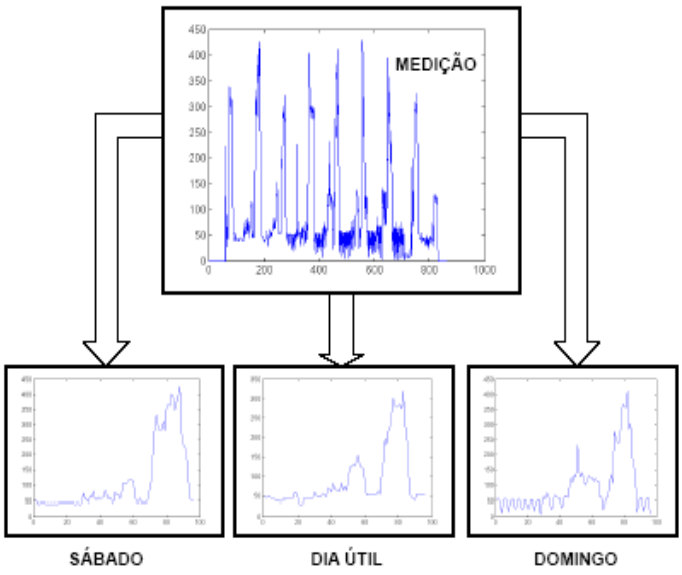


Figura 21 - Processo de separação das curvas típicas

O SIADAGE disponibiliza outros tipos de caracterização para a geração das curvas de carga típicas, caracterizações importantes para o conhecimento do usuário no que se refere ao conhecimento do consumo do cliente ou grupos de clientes sempre objetivando a melhor ação de GLD a ser adotada, são elas:

por Atividade Econômica (levando em consideração o CNEA)

por tarifa;

por classe de consumo;

por classe de demanda;

por épocas especificas ou períodos do ano (por exemplo, período seco ou úmido, primavera ou verão, etc.).

A Figura 22 ilustra a hierarquia das caracterizações das curvas de carga típicas:

- Classificação / Filtro

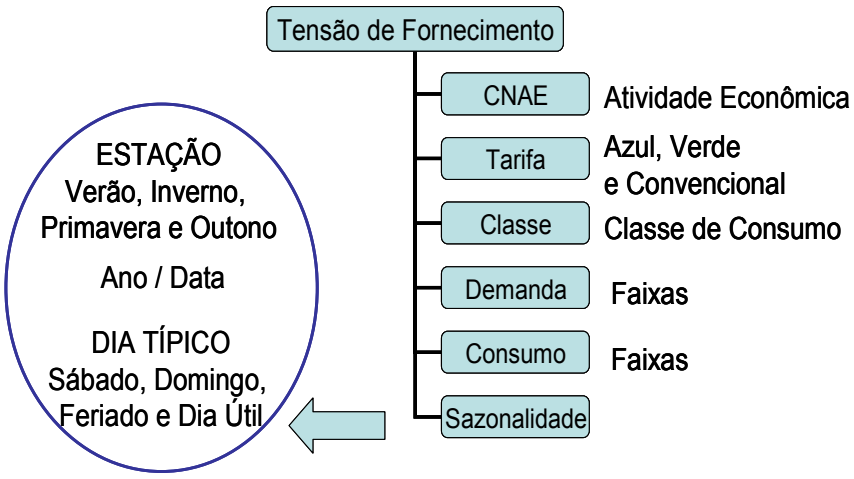


Figura 22 - Agregação e Síntese das Informações

Detalhamento do Processo de Formação das Curvas Típicas

O SIDAGE determina para cada ordenada "It":

$$I_t = \frac{1}{5} \cdot \left[2 \cdot I_{ts} + 2 \cdot Me\{I_{ts}\} + Mo\ \{I_{ts}\}\right]$$

Equação 1 - Estimação para cada ordenada "It".

Onde:

It é o valor médio estimado de cada ponto da amostra.

Its é o valor instantâneo de cada ponto da amostra.

Me é a mediana de cada ponto da amostra.

Mo é a moda de cada ponto da amostra.

Sendo os valores da amostra normalizados.

A utilização da Equação, na estimação da ordenada da curva típica de carga, torna possível obter resultados melhores fundamentados, em comparação com o uso apenas da média dos valores iniciais. Isto está associado ao fato de que no primeiro caso os valores aleatórios excessivos ou erros de medidas podem distorcer os resultados, levando em consideração o pequeno volume de uma amostra. Por isso, com esta abordagem pode-se eliminar ou reduzir a influência de medidas aleatórias ou incertas.

Os passos envolvidos no processo de geração das curvas típicas de carga são resumidos a seguir:

definição do intervalo de integração ou período de análise diário;

filtrar curvas conforme dia útil, sábados, domingos e feriados;

calcular a média estimada e o desvio padrão estimado para cada curva típica de carga;

obter uma curva típica de carga da média estimada e outra do desvio padrão para cada cliente, grupo de clientes ou elemento do sistema elétrico;

normalização das curvas obtidas pela demanda máxima para todos os consumidores ou elementos do sistema.

As curvas típicas geradas em kVA são compostas pelo valor individual de cada cliente monitorado, e não a partir da composição dos resultados finais obtidos em kW e kVAr

Anexo 4

Publicações Relacionadas ao Desenvolvimento da Dissertação

- Resumos expandidos publicados em anais de congressos

MINUSSI, M. M. ; SANTOS, F. L. S. ; IBIAS, M. V. G. ; KAEHLER, J. W. M. . Mineração de Dados para detecção de fraude nas empresas de distribuição de energia elétrica. In: Congrega - 4ª Jornada de Pós-Graduação, Pesquisa e Extensão, 2006, São Gabriel. Mineração de Dados para detecção de fraude nas empresas de distribuição de energia elétrica. São Gabriel : Fundação Attila Taborda - FAT, 2006. p. 170-171.

- Resumos publicados em anais de congressos

MINUSSI, M. M. ; SANTOS, F. L. S. ; IBIAS, M. V. G. ; KAEHLER, J. W. M. . Mineração de Dados para detecção de fraude nas empresas de distribuição de energia elétrica. In: Mostra de Pesquisas da Pós-Graduação da PUCRS 2006, 2006, Porto Alegre. Mineração de Dados para detecção de fraude nas empresas de distribuição de energia elétrica. Porto Alegre : EDIPUC, 2006.

- Apresentação de Trabalho/Congresso

MINUSSI, M. M. ; ETCHICHURY, F. C. ; IBIAS, M. V. G. ; KAEHLER, J. W. M. . Aplicação de Data Mining em Banco de Dados Histórico de Medições de Cargas. Congresso Internacional de Distribuição Elétrica.Aplicação de Data Mining em Banco de Dados Históricos de Medições de Carga. 2006.

Este capítulo apresentou um breve resumo sobre cenário atual no setor de distribuição de energia elétrica, no Brasil, salientando o aspecto que os dados dessas companhias não tem um aproveitamento adequado para uma detecção de fraude. Também foi apresentado o problema foco dessa dissertação que é o roubo de energia elétrica sendo que, em média 15% da energia gerada no país é furtada. Com base nestes fatos, verificou-se a necessidade da criação de uma metodologia com base nos dados da concessionária AES - Sul, aplicando

métodos para fazer uma Mineração de Dados e objetivando a detecção de fraudes e aplicar estas técnicas no estudo de caso e apresentando os resultados na conclusão desta dissertação.