UNIVERSIDADE FEDERAL DE UBERLÂNDIA

FACULDADE DE ENGENHARIA ELÉTRICA

PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

SISTEMA INTELIGENTE BASEADO EM ÁRVORE DE DECISÃO, PARA APOIO AO COMBATE ÀS PERDAS COMERCIAIS NA DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

Dissertação apresentada à Universidade Federal de Uberlândia por José Reis Filho para a obtenção do título de Mestre em Engenharia Elétrica.

Professor Antônio Carlos Delaiba, Dr. (orientador)
Professor Keiji Yamanaka, Ph.D.
Professor Kleiber David Rodrigues, Dr.
Professor João Onofre Pereira Pinto, Ph.D.

Uberlândia
2006

SISTEMA INTELIGENTE BASEADO EM ÁRVORE DE DECISÃO, PARA APOIO AO COMBATE ÀS PERDAS COMERCIAIS NA DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

JOSÉ REIS FILHO

Dissertação apresentada por José Reis Filho à Universidade Federal de Uberlândia como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica.

Professor Antônio Carlos Delaiba, Dr. Orientador	Professor Darizon Alves de Andrade, Ph.D. Coordenador do Curso de Pós-Graduação

Para José Reis e Custódia, meus pais, pelos exemplos de honestidade e trabalho.

Para Cláudia, minha esposa, pelo o apoio e dedicação em todos os dias.

Para minhas filhas Fernanda e Juliana, pelo amor e carinho nos momentos difíceis.

AGRADECIMENTOS

A Deus nosso pai e criador;

A Concessionária de energia pela oportunidade de crescimento pessoal, intelectual e profissional;

Ao meu orientador Prof. Delaiba, pela confiança na realização do trabalho;

Ao Prof. João Onofre coordenador do projeto, pelos resultados alcançados.

Este trabalho não seria possível sem a colaboração de alguns colegas, os quais não poderia deixar de mencionar e agradecer:

Alexandra Maria Almeida Carvalho Pinto
Prof. Dr. Evandro Mazina Martins
Nery de Oliveira Lima Neto
José Edson Cabral Júnior
Edgar Marques Gontijo

“Não ande somente pelos caminhos já trilhados,
eles só o levarão onde alguém já esteve”.

Graham Bell

RESUMO

O aumento das perdas comerciais nas concessionárias distribuidoras de energia tem sido motivo de grande preocupação das empresas. Os principais motivos desse aumento, são ocasionados por dois grandes problemas enfrentados pelas empresas de distribuição de energia, que são as fraudes efetuadas pelos consumidores e também os problemas decorrentes em medidores de energia.

Atualmente para identificar essas situações são realizadas inspeções nas unidades consumidoras. Devido ao elevado número de unidades, tais inspeções são efetuadas sem uma pré-análise eficiente de comportamento dos clientes, acarretando baixas taxas de acertos.

Por outro lado as concessionárias de distribuição possuem armazenadas em seus bancos de dados uma grande quantidade de informações de seus clientes. Essas informações podem ser utilizadas na identificação de perfis de comportamento das unidades consumidoras. Porém devido a grande quantidade de informações torna-se necessário um processo automatizado para identificação dos perfis.

O objetivo deste trabalho é desenvolver um sistema de suporte ao combate às perdas comerciais para apoio ao setor de distribuição de energia elétrica. Tal sistema será baseado em Descobrimento de Conhecimento em Banco de Dados (DCBD), que trata da descoberta de informações em banco de dados aumentando as possibilidades de inspeções bem sucedidas em campo.

Será utilizada a técnica de Árvore de Decisão como ferramenta de mineração de dados. Trata-se de uma técnica que se baseia em inteligência artificial que busca implementar em máquinas, habilidades humanas realizando o processo de aprendizagem, utilizando métodos de classificação.

ABSTRACT

The increase in commercial losses in electric utility companies has been a reason of great concern for these companies. The main motives of the increase in these losses are two: fraud practiced by the consumers; and problems in the energy meters.

Nowadays, to identify one of the two problems mentioned above, *in-site* inspections are required. However, due to the high number of consumer unities, such inspections are done without any previous analysis of the consumer behavior, which results in a low rate of problem identification.

On the other hand, electric utility companies have a database with much information about their consumers. So, this information can be used to identify the behavior profile of those consumers that are likely to be frauding or having problems with their energy meters. However, due to high quantity of data, it is demanding the use of an automatic process for identification of such behavior profiles.

The goal of this work is to develop a decision support system to combat commercial losses in distribution power systems. Such system is based on Knowledge Discovery in Database – KDD, which refers to discovering of knowledge in database, which may increase the rate of successful *in-site* inspections.

The tool used to do the data mining stage of the KDD is Decision Tree. This is an artificial intelligence technique that tries to emulate human abilities in a computer system, and it learns from data and it is used for classification type of problems.

SISTEMA INTELIGENTE, BASEADO EM ÁRVORE DE DECISÃO, PARA APOIO AO COMBATE ÀS PERDAS COMERCIAIS NA DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

Sumário

CAPÍTULO I 12

INTRODUÇÃO 12

1.1 CONTEXTUALIZAÇÃO 12
1.2 REVISÃO BIBLIOGRÁFICA 17
1.3 PROBLEMAS ASSOCIADOS ÀS PERDAS COMERCIAIS 23
1.4 DISPONIBILIDADE DE DADOS 24
1.5 OBJETIVO DA DISSERTAÇÃO 26
1.6 ORGANIZAÇÃO DO TRABALHO 27

CAPÍTULO II 28

PERDAS NO SETOR ELÉTRICO 28

2.1 INTRODUÇÃO 28
2.2 HISTÓRICO DAS PERDAS 31
2.3 PERDAS TÉCNICAS NA DISTRIBUIÇÃO 36
2.3.1 Condutores de rede primária de distribuição 40
2.3.2 Transformadores 40
2.3.3 Condutores de rede secundária 41
2.3.4 Ramais de ligação ou ramais de serviço 42
2.3.5 Medidores 43
2.3.6 Conectores 44
2.3.7 Equipamentos 45
2.3.8 Perdas diversas 46
2.4 PERDAS COMERCIAIS NA DISTRIBUIÇÃO 47
2.4.1 Ligações clandestinas 48
2.4.2 Intervenções indevidas no padrão e na medição 50
2.4.2.1 Irregularidade no ramal de ligação 50
2.4.2.2 Irregularidade no ramal de entrada 51
2.4.2.3 Irregularidade no disjuntor 51
2.4.2.4 Irregularidade no medidor 52
2.4.2.5 Religação à revelia 53
2.4.3 Medidores 53
2.4.4 Medições indiretas 56
2.4.5 Perdas comerciais de origem administrativas 56
2.4.6 Falta de medição 58
2.4.7 Cargas especiais sem medição 58
2.4.8 Perdas na transformação 59
2.4.9 Perdas em iluminação pública 60
2.4.10 As perdas no ponto de vista jurídico 62
2.5 COMBATE ÀS IRREGULARIDADES 65
2.5.1 Inspeções de varredura 66
2.5.2 Inspeções de consumo zero 67
2.5.3 Inspeções de unidades consumidoras inativas 67

2.5.4 Inspeções a partir de denúncias *68*
2.6 PROCEDIMENTOS DE INSPEÇÃO 68
2.7 COMENTÁRIOS FINAIS 71

CAPÍTULO III 73

PROCESSO DE DCBD (DESCOBRIMENTO DE CONHECIMENTO EM BANCO DE DADOS) E MINERAÇÃO DE DADOS 73

3.1 INTRODUÇÃO 73
3.2 DESCOBRIMENTO DE CONHECIMENTO EM BANCO DE DADOS 75
3.2.1 Seleção dos dados *75*
3.2.2 Pré-processamento de dados *76*
3.2.3 Transformação dos dados *77*
3.2.4 Mineração de dados *85*
3.2.5 Interpretação do conhecimento descoberto *90*
3.2.6 Consolidação do conhecimento descoberto *91*
3.3 ÁRVORE DE DECISÃO 91
3.4 COMENTÁRIOS FINAIS 97

CAPÍTULO IV 99

DESENVOLVIMENTO DO SISTEMA DE IDENTIFICAÇÃO DE FRAUDES E ERROS DE MEDIÇÃO 99

4.1 INTRODUÇÃO 99
4.2 PROCESSO DE SELEÇÃO DE DADOS 100
4.3 BANCO DE DADOS 103
4.4 DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS 105
4.4.1 Montagem do banco de dados - Seleção e coleta de dados *105*
4.4.1.1.Preparação dos dados: pré-processamento e consolidação dos dados 106
4.4.1.2 Pré-processamento dos dados 108
4.4.1.3 Consolidação dos dados 109
4.4.1.3.1 Relacionamento de consumo com inspeção 109
4.4.1.3.2 Relacionamento de consumo e inspeção com trafos 111
4.4.1.3.3 Concentração de registros 112
4.4.1.3.4 Seleção de clientes normais e fraudadores 114
4.4.2 Transformação dos dados *115*
4.4.3 Mineração de dados de dados utilizando Árvore de Decisão *116*
4.4.4 Estudo de casos *121*
4.4.5 Análise final dos casos simulados *140*
4.4.6 Padrões e Modelos – Avaliação *140*
4.5 COMENTÁRIOS FINAIS 141

CAPÍTULO V 142

CONCLUSÕES E PROPOSTAS DE TRABALHOS FUTUROS 142

5.1 CONSIDERAÇÕES FINAIS 142
5.2 TRABALHOS FUTUROS 145
5.3 ARTIGOS PUBLICADOS 146

LISTA DE FIGURAS

Figura 2.1 Perdas anuais globais..... 33
Figura 2.2 Perdas globais 2004..... 33
Figura 2.3 Perdas técnicas e comerciais – global 2004..... 34
Figura 2.4 Perdas na distribuição 2004..... 35
Figura 2.5 Perdas técnicas e comerciais na distribuição..... 35
Figura 2.6 Diagrama unifilar de um sistema de distribuição..... 36
Figura 2.7 Ilustrações de Ligações Clandestinas..... 49
Figura 2.8 Ilustrações de irregularidade no ramal de ligação..... 50
Figura 2.9 Ilustrações de irregularidade no ramal de entrada..... 51
Figura 2.10 Ilustrações de irregularidade no disjuntor..... 52
Figura 2.11 Ilustrações de irregularidade no medidor..... 53
Figura 3.1 Diagrama de blocos do processo de DCBD..... 74
Figura 3.2 Agrupamento para identificação de *outliers*..... 80
Figura 3.3 Regressão linear para identificação de *outliers*..... 80
Figura 3.4 Percentual de variação para identificação de *outliers*..... 81
Figura 3.5 Redução de dados por amostragem estratificada..... 84
Figura 3.6 Modelo de árvore de decisão..... 92
Figura 3.7 Redução de dados com uso de Árvore de Decisão..... 93
Figura 4.1 Árvore de Decisão com 5 atributos..... 119
Figura 4.2 Parte da Árvore de Decisão com 5 atributos..... 120
Figura 4.3 Matriz de eficiência do sistema..... 124
Figura 4.4 Matriz de eficiência do sistema – caso 01..... 126
Figura 4.5 Matriz de eficiência do sistema – caso 02..... 127
Figura 4.6 Matriz de eficiência do sistema – caso 03..... 129
Figura 4.7 Matriz de eficiência do sistema – caso 04..... 130
Figura 4.8 Matriz de eficiência do sistema – caso 05..... 131
Figura 4.9 Matriz de eficiência do sistema – caso 06..... 133
Figura 4.10 Matriz de eficiência do sistema – caso 07..... 134
Figura 4.11 Matriz de eficiência do sistema – caso 08..... 135
Figura 4.12 Resposta do sistema com variação dos critérios..... 137

LISTA DE TABELAS

Tabela 2.1 Perdas nos componentes do sistema de distribuição..... 39
Tabela 2.2 Irregularidades com perda em medidores..... 54
Tabela 2.3 Perdas estimadas por fases..... 55
Tabela 2.4 Perdas estimadas por origem..... 55
Tabela 2.5 Perdas estimadas por classe..... 55
Tabela 3.1 Discretização de consumo de energia elétrica..... 83
Tabela 3.2 Componentes da Árvore de Decisão..... 92
Tabela 4.1 Lista de atributos..... 101
Tabela 4.2 Informações quantitativas do banco de dados..... 109
Tabela 4.3 Registros de uma unidade consumidora anônima da tabela *CI*..... 111
Tabela 4.4 Unidades consumidoras da tabela *CIT* agrupadas pelo número de registros..... 112
Tabela 4.5 Unidades consumidoras da tabela *CIT* agrupadas pelo número de registros..... 113
Tabela 4.6 Unidades consumidoras da tabela *CIT* agrupadas pelos resultados de inspeção. 115
Tabela 4.7 Conjunto de atributos disponíveis para o processo de mineração..... 115
Tabela 4.8 Análise quantitativa das fraudes – caso 01..... 126
Tabela 4.9 Análise quantitativa das fraudes – caso 02..... 127
Tabela 4.10 Análise quantitativa das fraudes – caso 03..... 128
Tabela 4.11 Análise quantitativa das fraudes – caso 04..... 130
Tabela 4.12 Análise quantitativa das fraudes – caso 05..... 131
Tabela 4.13 Análise quantitativa das fraudes – caso 06..... 132
Tabela 4.14 Análise quantitativa das fraudes – caso 07..... 134
Tabela 4.15 Análise quantitativa das fraudes – caso 08..... 135
Tabela 4.16 Análise com critério 10 a 100 – caso 09..... 137
Tabela 4.17 Relação Normal/Fraudador – NF 1/1, 1/2, 1/3, 1/4, 1/5..... 139
Tabela 4.18 Relação Normal/Fraudador – NF 2/1, 2/2, 2/3, 2/4, 2/5..... 139
Tabela 4.19 Relação Normal/Fraudador – NF 3/1, 3/2, 3/3, 3/4, 3/5..... 139
Tabela 4.20 Relação Normal/Fraudador – NF 4/1, 4/2, 4/3, 4/4, 4/5..... 139

SISTEMA INTELIGENTE, BASEADO EM ÁRVORE DE DECISÃO, PARA APOIO AO COMBATE ÀS PERDAS COMERCIAIS NA DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

CAPÍTULO I

INTRODUÇÃO

1.1 Contextualização

As crescentes mudanças no cenário energético nacional têm exigido maior habilidade nas tomadas de decisões, seja para definir os investimentos futuros ou até mesmo os de curto prazo. Desta forma, a correta aplicação dos recursos técnicos e financeiros nas empresas tornou-se imprescindível e condição vital para a sobrevivência das concessionárias de distribuição de energia elétrica.

Dentro deste contexto, as distribuidoras intensificaram a preocupação com o aumento das perdas na sua área de atuação. Principalmente com aquelas perdas denominadas não técnicas também conhecidas como perdas comerciais.

Essas perdas na maioria das situações podem ser provocadas intencionalmente por consumidores, através de procedimentos irregulares ou ainda por falhas e defeitos nos medidores de energia.

Estima-se que o montante dessas perdas chegue a 6% do faturamento das concessionárias distribuidoras de energia.

A dificuldade em concretizar ações eficazes para a solução das perdas comerciais tem feito as concessionárias buscarem diversas alternativas, porém muitas vezes com baixa eficiência no resultado.

Uma dessas alternativas poderia ser o aumento no número de inspeções em unidades consumidoras *in loco*, contudo, não seria possível inspecionar todos os consumidores atendidos pela empresa.

Aumentar o número de inspeções tornaria o processo inviável na relação custo/benefício, principalmente em função de majoração de recursos a serem destinados a visitas de técnicos em unidades consumidoras, não tende a promover melhorias significativas nos resultados.

A questão principal para a minoração das perdas está na seleção adequada dos consumidores a serem inspecionados.

Atualmente, o processo de inspeção na maioria das vezes é realizado através de inspeções em consumidores selecionados por técnicos especializados nesta tarefa.

Outra maneira também utilizada é a varredura, na qual uma área é escolhida e uma equipe de técnicos percorrem ponto por ponto tentando identificar possíveis irregularidades.

Tipicamente, a seleção dos consumidores a serem inspecionados é baseada nos dados cadastrais do consumidor e no seu perfil de consumo. Com base na sua experiência, o especialista realiza consultas à base de dados e posteriormente seleciona manualmente os resultados da consulta para identificar os consumidores que devem ser submetidos à inspeção.

Muitos clientes que são inspecionados sentem-se desconfortáveis com a visita dos fiscais, por entenderem que existe desconfiança por parte da concessionária. Isso gera um grau de insatisfação e conflito entre a sociedade e a empresa de distribuição de energia.

O presente contexto sinaliza para a necessidade de se investigar alternativas que permitam selecionar melhor o candidato a ser inspecionado.

O processo de identificação dos consumidores baseado no cadastro e nos dados de perfil de consumo, na prática, é artesanal. Desta forma acaba impedindo que os técnicos tenham como avaliar detalhadamente um grande número de consumidores candidatos. Como resultado, o índice de sucesso é baixo, ficando na faixa de 5% a 10% do total de serviços de inspeções realizadas.

Para se ter uma idéia, a empresa a ser avaliada possui cerca de 620 mil consumidores e suas equipes de inspeção são capazes de realizar na ordem de 120 mil operações de inspeção por ano. Isto sem levar em consideração que pode ser necessário realizar várias inspeções em um mesmo consumidor no período de um ano. O montante da perda comercial calculada em 2004 para a empresa analisada, chegou próximo ao montante de 300.000 MWh, representando um valor estimado em R$ 65 milhões de reais.

Outro motivo de preocupação para as concessionárias é com relação às reposições tarifárias para recompor as perdas, pois, atualmente a tarifa não mais remunera totalmente tais eventos. Assim os resultados apurados das perdas que estejam acima dos índices estabelecidos pelo órgão regulador do setor elétrico, a ANEEL (Agência Nacional de Energia Elétrica), não são mais remuneradas, desta forma torna-se essencial a tomada de ações para recuperação destas receitas.

Além dos aspectos financeiros para a concessionária, existe também a questão do impacto nos reajustes e revisões tarifárias. Estes por sua vez estão diretamente correlacionados com as referidas perdas. A situação ideal é a prática constante da modicidade tarifária para a população.

Apesar da grande necessidade na redução das perdas comerciais, o problema a cada dia torna-se mais grave. As atuações dos clientes têm evoluído constantemente, criando métodos de irregularidade de difícil percepção.

Outra questão que dificultou de forma significativa a identificação dessas situações, aconteceu no ano de 2001. Com início do racionamento as metodologias de análise para determinação de inspeções das unidades tornaram-se mais complexas.

Houve assim, o aumento do grau de dificuldade para localização das possíveis unidades com fraudes e/ou problemas nos medidores de energia, pois durante um período de 8 meses foram efetuadas ações diversificadas por parte dos consumidores para atingir as suas metas de redução do consumo de energia.

As metas foram elaboradas pelas concessionárias por determinação do governo federal. Tal redução de consumo foi determinada para contornar uma situação de crise muito delicada que atravessava o setor elétrico.

Em função de um planejamento inadequado por parte dos governantes da nação, veio à tona a informação que o sistema elétrico nacional não estava preparado para a demanda de energia requerida pelo país.

Essa situação que gerou estagnação no crescimento e no desenvolvimento do setor produtivo brasileiro impactando de forma negativa na economia.

Após o período de racionamento, uma diversidade de mudanças havia ocorrido no perfil de comportamento dos consumidores de energia elétrica.

A população percebeu a necessidade de economizar energia, e mais ainda percebeu que havia um desperdício grande do produto na sua rotina diária.

Essa conscientização provocou mudanças de hábitos no consumo, ocasionando uma redução nos seus custos com energia elétrica.

Por outro lado as distribuidoras tiveram que buscar meios de recompor a sua situação financeira que foi afetada pela crise. Desta forma, para diminuir as perdas, as empresas intensificaram os esforços para recuperações de receita ocasionadas pelas fraudes.

Geralmente as empresas possuem especialistas que indicam quais unidades devem ser alvo de inspeção. Esta decisão baseia-se em alguns fatores: região com alta incidência de fraudes, denúncias, média de consumo baixa, entre outros.

Em razão do grande número de unidades consumidoras é praticamente impossível a avaliação do comportamento de cada uma pelo especialista.

Encontrando um perfil que indique um comportamento suspeito, o especialista pode recomendar que este seja inspecionado. O ideal é que o processo de descoberta destes padrões de comportamento seja realizado de maneira automática, por alguma ferramenta computacional que analise os dados e extraia conhecimento.

Hoje já existem diversos segmentos na sociedade que utilizam as técnicas de mineração de dados na detecção de fraudes. Dentre as diversas áreas que tem buscado soluções para minimizar seus problemas de perdas podemos citar: empresas de cartões de crédito, água, telefonia, distribuição de energia, dentre outros.

Seria inviável para qualquer ramo de negócios investigar grandes volumes de informações utilizando pessoas, por maior que fosse a equipe disponível.

A ajuda da inteligência computacional veio a corroborar com a necessidade de averiguar e processar dados de maneira rápida e confiável, onde seu manuseio seria humanamente impraticável.

Neste contexto, as técnicas de mineração de dados têm um papel preponderante por estarem aptas a lidar com grandes quantidades de dados e serem aplicadas em trabalhos investigativos.

As técnicas de Inteligência Artificial (IA) buscam encontrar e interpretar padrões em dados incrementando habilidades do ser humano em sistemas computacionais.

A aplicação da mineração de dados neste trabalho será com o objetivo de alcançar melhorias nos índices das perdas comerciais.

Através da investigação das características das unidades consumidoras poderá obter padrões de comportamento que indiquem a possibilidade de fraude ou ainda problemas em medidores de energia.

Este estudo poderá contribuir com o aprimoramento das técnicas hoje usadas na detecção de fraudes em energia elétrica e a exploração científica do processo de mineração de dados como ferramenta para descoberta de conhecimento no domínio de distribuição de energia.

Árvore de Decisão é uma técnica de (IA) que realiza o processo de aprendizagem, utilizando métodos de classificação. Amplamente utilizada em algoritmos de classificação, Árvore de Decisão é uma representação simples do conhecimento. É um meio prático de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados.

Neste trabalho, Árvore de Decisão é utilizada em um banco de dados de uma empresa concessionária de energia visando a identificação de clientes potencialmente fraudadores e ou com problemas em medidores de energia.

São feitos então experimentos com a análise dos resultados de maneira cíclica e evolutiva para avaliação da técnica.

Deseja-se alcançar regras de classificação que sejam capazes de determinar tais perfis com uma taxa de acerto médio de 30%. Seqüencialmente é efetuada a construção de um sistema automático de classificação.

1.2 Revisão bibliográfica

De maneira geral fraudes são eventos decorrentes em quase todos os segmentos de negócios, entretanto alguns ramos de atividades são mais evidenciados: cartões de crédito,

telefonia, (fixa e móvel), consumos de água e energia, seguros (vida, imobiliários, automóveis etc), planos de saúde, bancos, imposto de renda, etc, são exemplos destes ramos de atividades.

Uma pesquisa, envolvendo aproximadamente 1.000 empresas brasileiras do ramo industrial, revelou que somente 50% das perdas por fraudes no ano de 2004 foram recuperadas (KPMG, 2004). Segundo a mesma pesquisa, 71% das empresas indicaram como a principal causa a precariedade do sistema de controle interno. Pode-se notar a partir dos índices apurados na pesquisa a gravidade do problema enfrentado e a enorme necessidade de ações ou mecanismos para detecção mais eficientes.

Existem disponíveis no mercado alguns programas comerciais para detecção de fraudes. O Clementine2, comercializado por SPSS Inc., disponibiliza ferramentas de classificação, agrupamento e predição, podendo ser utilizado na detecção de vários tipos de fraudes, porém por ser um software genérico para mineração de dados, sua performance para setores específicos não é satisfatória.

Já o programa Falcon Fraud Manager, comercializado por Fair Isaac3, é direcionado especificamente para detecção de fraudes em cartões crédito e utiliza modelos baseados na técnica de redes neurais artificiais. A adaptação deste software para o setor elétrico apresenta os mesmos problemas que o Clementine 2.

A área de cartões de crédito concentra a maioria dos trabalhos publicados sobre detecção de fraudes utilizando técnicas de inteligência artificial, em (Kou et al., 2004) encontra-se uma revisão dos principais métodos.

As fraudes em cartões de crédito e instituições financeiras não são divulgadas, pois tal fato poderia comprometer a credibilidade destas instituições. Por isto, investigações sobre estes tipos de fraudes são pouco conhecidas e não são publicadas com detalhamento conforme [Kou, 2004].

As técnicas e metodologias para detecção de fraudes em cartões de crédito são baseadas no histórico completo de transações dos portadores ou apenas nas informações recentes e inerentes a uma nova transação

Uma das análises é baseada nos aspectos relativos à informação geográfica na utilização de cartões, como aquisições de grande vulto solicitando envio para locais distantes, ou o uso imediato do cartão em dois locais distantes em pequeno intervalo de tempo.

Essas fraudes podem ser divididas em duas categorias: fraude offline e online.

A offline é executada através do roubo físico do cartão de crédito e sua posterior utilização diretamente na aquisição de bens. Em sua maioria, as instituições são capazes de bloquear o uso destes cartões antes mesmo de seu uso ilícito.

A online é executada via telefone, internet ou em compras sem a presença do dono do cartão, portanto sem a assinatura manual do comprador [Kou, 2004].

Uma outra forma de categorizar as fraudes em cartões de crédito é segundo [Bolton e Hand, 2001] através de fraude comportamental ou da fraude na aplicação.

A fraude na aplicação ocorre quando um indivíduo obtém um cartão através de dados falsos, e a fraude comportamental, mais freqüente e mais difícil de ser identificada ocorre quanto o fraudador obtém dados de um cartão válido e os utiliza, especialmente em aquisições sem a presença do proprietário.

Na pesquisa de [Bolton e Hand, 2001] as fraudes comportamentais são investigadas através de métodos não supervisionados, uma vez que métodos supervisionados têm dificuldade em detectar comportamentos fraudulentos não encontrados previamente nos históricos de transações. Pode-se compará-los com métodos probabilísticos que necessitam de dados confiáveis para treinamento dos classificadores.

O aprendizado não-supervisionado encontra-se o comportamento normal de cada portador de cartão de crédito a partir de seu histórico de transações. Variações na freqüência

ou no valor das transações, por exemplo, podem direcionar para desvios em comportamento, indicando possíveis fraudes. A principal adversidade deste método é controlar o número de falsos alarmes, ou seja, diferenciar transações legais incomuns (exceções) de transações fraudulentas.

O CARDWATCH é um programa de mineração de dados voltado para a detecção de fraudes em cartões de crédito, baseando-se em uma rede neural artificial de alimentação direta (feedforward) [Aleskerov et al., 1997]. A partir de dados gerados por simulações (dados artificiais), alcançou-se uma taxa de acerto de 85% na detecção de fraudes.

As empresas de telecomunicações é um dos segmentos empresarias que constantemente são vítimas de fraudes. Devido a grande expansão na última década de linhas telefônicas principalmente no que se refere a aparelhos móveis, a ampliação do número de fraudes cresceu proporcionalmente a esse mercado.

Estima-se que são perdidos de 4% a 6% dos lucros entre as empresas de telecomunicações. Essas fraudes além da questão financeira causam outros impactos para as operadoras de telecomunicações, entre os principais, está a perda da capacidade de transmissão de dados/voz.

Nas áreas dos serviços públicos de energia elétrica e abastecimento de água a natureza das fraudes é bastante similar.

Diferentemente dos seguimentos de cartões de crédito, ou chamadas telefônicas, neste tipo de serviço, existe uma medição instalada *in loco*, onde se verifica periodicamente o consumo do usuário.

O serviço é utilizado de forma continua e o seu registro é monitorado por equipamentos de medição instalados nas dependências do usuário. Para que as empresas prestadoras deste serviço efetuem a cobrança é necessária à obtenção da leitura nos equipamentos de medição

seja de água ou energia. Apura-se a diferença entre a leitura obtida no mês de referencia e a leitura do mês anterior.

Normalmente, as fraudes são efetuadas a partir de adulteração dos dispositivos de medição, ou da conexão direta na rede de distribuição por parte dos consumidores.

Na tentativa de reduzir perdas, empresas concessionárias têm investido em automação, na implantação de sistemas de informatização integrados, na instalação de equipamentos e no combate às fraudes.

Na Sanasa, empresa de distribuição de água de Campinas, as fraudes contribuíram em 5% dos 26,6% de perdas na distribuição, no ano de 2000 [Passini, 2002]. O uso de mineração de dados para identificar fraudes surgiu por existirem dados históricos armazenados há mais de dez anos, que poderiam ser investigados para descoberta de informações válidas e desconhecidas, contribuindo para identificação de perfis de comportamento que pudessem levá-los aos fraudadores.

No trabalho [Passini, 2002] foi utilizado o programa DB2 Intelligent Miner, comercializado pela IBM7A, baseado em árvore de decisão para a detecção de fraudes em ligações de água. Na fase de treinamento do sistema, foram selecionados 80% dos consumidores fraudadores já conhecidos. Uma grande quantidade de testes foi realizada alternando-se os parâmetros de entrada do modelo selecionado, verificando para cada teste o percentual de erros e acertos.

O projeto da Sanasa tinha como motivação para uso de mineração de dados o combate às perdas de água, com foco nas irregularidades nas ligações de água e tinha como objetivo uma redução de 51% para 41% das visitas improcedentes para detecção de fraudes.

Os resultados alcançados ficaram aquém dos esperados, no entanto, sabia-se que o modelo ainda precisava ser melhorado. A performance ruim se deve provavelmente ao fato de o IBM7A ser um software de mineração de dados genérico.

Em [Eller, 2003], a pesquisa na área de energia elétrica voltou-se para a construção de uma arquitetura de sistemas capaz de realizar o gerenciamento de perdas comerciais de energia. Esta arquitetura está baseada na utilização de Redes Neurais para a identificação de potenciais fraudadores através de classificação. Os resultados apresentados demonstraram uma melhora na identificação de fraudadores em relação aos seus processos artesanais anteriores de amostragem e visita a campo.

No trabalho [Reis et al., 2004] é apresentado um sistema de pré-seleção de consumidores de energia elétrica para inspeção, com o objetivo de detectar fraudes e erros de medição. A partir do banco de dados de uma empresa de distribuição de energia elétrica, foram selecionados cinco atributos (dentre os 52 disponíveis) e 40.000 registros (de um total de 600.000). O sistema é baseado em uma árvore de decisão CART [Breiman et al., 1993], a qual foi treinada com 20.000 registros selecionados aleatoriamente. Os testes do sistema com os 20.000 registros remanescentes resultou em uma taxa de acerto de 40% para fraudadores, 35% a mais que a taxa alcançada pela empresa em questão.

Em (Cabral et al., 2004) foram utilizados alguns conceitos de Rough Sets para a identificação de padrões de comportamento fraudulentos em dados históricos. Um conjunto de clientes e seus respectivos atributos foram organizados em um Sistema de Informação, onde foram aplicados os conceitos de aproximação inferior, reduto e do algoritmo da decisão mínima, ou minimal decision algorithm (MDA). A partir do Sistema de Informação reduzido, derivou-se um conjunto de regras as quais representaram perfis de comportamento de clientes. Tomando-se os perfis de comportamento fraudulento, consolidou-se um sistema de regras de classificação, o qual alcançou uma taxa de acertos de fraude de 20%.

1.3 Problemas associados às perdas comerciais

Como em outros ramos de negócios, as concessionárias de distribuição de energia elétrica também podem ser alvos de fraudes por parte de seus clientes ou ainda sofrerem por falhas em seus processos. No Brasil, as perdas de receita de algumas empresas podem chegar a margens superiores a 10%.

Uma das formas de combater estas perdas é a execução de inspeções nas unidades consumidoras que muitas vezes devido à sua baixa eficiência, podem se tornar uma atividade de alto custo, demandando grandes disponibilidades de recursos.

A caracterização desta baixa eficiência é a constatação de que a razão entre fraudes detectadas e número de inspeções realizadas fica inferior a 10%, inviabilizando tal processo.

A gravidade do assunto não limita somente na questão citada. Estende-se ainda para situações relativas a imagem da empresa que muitas vezes inspeciona unidades consumidoras que de certa forma sentem-se desconfortáveis com a presença da concessionária de energia.

Pois o consumidor subentende que a visita na sua unidade objetiva-se encontrar irregularidades em sua medição, e na maioria das vezes são inspeções equivocadas.

Outro fato relevante é a situação que em muitos casos o cliente que elabora uma irregularidade nas suas instalações e não é constatada rapidamente pela concessionária gera o incentivo de outros consumidores para a mesma prática.

Assim, fazendo avaliação do processo de perdas conclui-se que quanto maiores forem as perdas, conseqüentemente haverá tarifas de energia com valores maiores.

Partindo-se do princípio que quanto maior a tarifa, maior o número de fraudes a situação torna-se um círculo vicioso caso não seja tomadas providências para a solução do problema.

1.4 Disponibilidade de dados

Houve nos últimos tempos um elevado aumento na quantidade de informações disponíveis em bancos de dados nas empresas das mais diversas áreas. Esse aumento de dados no formato eletrônico é uma conseqüência natural dos avanços tecnológicos e do valor associado a informação no mundo globalizado.

Em função da maneira em que os dados são armazenados, e principalmente pelo grande volume existentes para análise, a interpretação dos dados a cada dia torna-se mais difícil.

Como a disponibilidade para armazenamento se tornou financeiramente acessível e barato, tem sido uma ação maciça das empresas em geral, a prática de guardar essas informações em bancos de dados.

Dentro deste contexto, as concessionárias de energia não fizeram diferente, utilizaram também dessa prática nos últimos anos e armazenaram um volume significativo de informações de seus consumidores.

Um dos seus objetivos foi para atender as necessidades de fiscalização, mas principalmente teve-se a intenção que desta coleta intensiva de dados obtivesse informações para atingir metas e vantagens competitivas.

A recuperação de perdas de receitas ocasionadas por irregularidades é um fator bastante relevante para as distribuidoras de energia elétrica. Porém, a identificação das unidades consumidoras com comportamento fraudulento ou problemas em medição é uma tarefa complexa. Normalmente, esta tarefa envolve inspeção *in loco*, onde geralmente tais inspeções são feitas aleatoriamente, ou a partir da experiência do responsável.

A quantidade de fraudes detectadas nas inspeções é muito baixa comparado com o número total de inspeções. A relação percentual do número de inspeções totais e o número de fraudes efetivamente detectadas em campo é da ordem de 5 %.

Os métodos manuais ocasionam o aumento na possibilidade de erros nas análises e interpretação das informações, e conseqüentemente nas tomadas de decisões pela empresa. Nestes métodos, o especialista compara suas hipóteses com os dados existentes, porém, quanto maior for a quantidade de dados agrupados, maior o tempo necessário para as análises, em decorrência disso, às vezes tal alternativa é inviável.

Este trabalho visa abordar o problema da necessidade do manuseio de uma quantidade cada vez maior de informações. O processo decisório, uma tarefa humana por excelência, depara-se então com uma grande quantidade de variáveis que influenciam na tomada de decisão.

Os sistemas de gerenciamento de banco de dados na maioria das implementações utilizam somente para executar consultas que são disparadas e processadas por uma máquina.

Também podem ser utilizados, outros sistemas que fazem a utilização de um sistema específico de gerenciamento de banco de dados relacional para arquivamento, modificação e gerenciamento de dados.

Dessa forma, a maioria dessas aplicações, possuem um fraco acoplamento com banco de dados, resultando em problemas de desempenho e limitações quanto a memória disponível.

A mineração de dados é uma ferramenta bastante utilizada para descobrir novas correlações de padrões e tendências. A utilização de técnicas de inteligência artificial é bastante eficiente na análise de grandes quantidades de dados armazenados.

Em resumo pode-se concordar que “mineração de dados é um processo de descoberta do conhecimento que consiste na aplicação de algoritmos específicos, sob alguma limitação aceitável de eficiência computacional, para produzir uma enumeração particular de padrões” [Fayyad, 1996].

1.5 Objetivo da dissertação

O objetivo deste trabalho é desenvolver um sistema de auxílio à tarefa de detecção de fraudes em unidades consumidoras e identificação de medidores de energia com problemas em uma concessionária distribuidora de energia.

O sistema proposto identificará consumidores com comportamento de consumo suspeito, os quais devem ser alvos de inspeção *in loco* e ainda valores incompatíveis de consumos registrados em medidores.

O trabalho avaliará uma técnica de Inteligência Artificial chamada Árvore de Decisão. Essa técnica será aplicada ao banco de dados da concessionária de distribuição de energia de elétrica.

Para atingir os objetivos citados, este trabalho utilizou uma metodologia organizada nas seguintes etapas:

- Etapa I - Identificação e análise de atributos relevantes para a determinação de fraudes e/ou problemas em medição.
- Etapa II - Levantamento do histórico de clientes com ocorrências de fraudes e/ou falhas de medições para criação do banco de dados a ser usado pelo sistema desenvolvido, o qual foi chamado: SIFEM – Sistema de Identificação de Fraudes e Erros de Medição.
- Etapa III - Realização de análise dos dados de treinamento com o propósito de adequá-los a ferramenta de Árvore de Decisão.
- Etapa IV - Desenvolvimento do sistema.
- Etapa V - Avaliação do SIFEM utilizando conjunto de dados de teste.

O sistema escolhido de classificação será baseado em Árvore de Decisão e foi desenvolvido, utilizando o software MATLAB.

1.6 Organização do trabalho

Capítulo I

No primeiro capítulo é feita uma introdução básica do setor elétrico, disponibilidade de dados e objetivos a serem alcançados.

Capítulo II

Na seqüência, o segundo capítulo apresenta uma abordagem da história das perdas de energia elétrica nas concessionárias, retratando sucintamente as perdas na geração e transmissão, dando ênfase na área de distribuição, destacando-se os aspectos técnicos e comerciais.

Capítulo III

No terceiro capítulo, é feita uma apresentação teórica de banco de dados e do processo de DCBD (Descobrimento de Conhecimento em Banco de Dados).

Capítulo IV

Já no quarto capítulo é descrita a metodologia, com especificação dos objetivos, das hipóteses, do contexto em que desenvolveu o estudo. Nesse capítulo são relatadas as atividades desde a descrição da origem dos dados, preparação, a aplicação da técnica de Árvore de Decisão com todo o desenvolvimento do sistema, e ainda o seu treinamento e teste.

Capítulo V

No quinto capítulo, são apresentados e discutidos os resultados obtidos no trabalho, nesta última parte da dissertação é também abordada às conclusões obtidas no trabalho e apresentadas propostas de trabalhos futuros.

CAPÍTULO II

PERDAS NO SETOR ELÉTRICO

2.1 Introdução

As perdas de energia nas concessionárias do setor elétrico ganharam destaque a partir de 1994, quando as empresas associadas a ABRADEE (Associação Brasileira de Distribuidores de Energia Elétrica), passavam a se interessar de forma mais concreta pelo tema.

Foi estabelecida através da edição de resolução (CODI 19-34), critérios e padrões para contabilização das perdas de energia elétrica.

A forma anterior a essa definição, não possibilitava de maneira clara uma visão real para identificar e mensurar as perdas de energia ocorridas nos sistemas elétricos das concessionárias.

Estas perdas podem ser classificadas de várias formas, pelo seu efeito, de acordo com o componente do sistema ou ainda pela causa, que podem ser desmembradas em duas categorias, perdas técnicas e não técnicas.

As perdas técnicas são aquelas intrínsecas ao sistema elétrico incluindo-se as perdas por efeito Joule, por efeito corona, por correntes de Foucault, por correntes de fuga, e outras. Podem ocorrer em condutores, nos dielétricos de capacitores, em equipamentos de proteção e controle, em dispositivos de medição, dentre outros.

As perdas não técnicas resultam de erro e/ou da não medição de consumo de energia, e são conseqüência da existência de consumidores clandestinos, medidores defeituosos, erros de

leituras, falta de atualização das informações, cargas sem medição e principalmente furto de energia elétrica.

As perdas foram divididas em três níveis de segmentos: Sistema global, Sistema de Transmissão e Sistema de Distribuição.

Com a estratificação as comparações entre as empresas puderam ser efetuadas com maior precisão e também obter credibilidade nas correlações dos resultados.

A resolução da ABRADDE 0001/26 aprovada em 17.11.94 cita ainda, algumas considerações relevantes que podemos entender como um marco na área de perdas.

A criação do indicador gerencial de perdas de energia nos sistemas das concessionárias possibilitou apuração e divulgação sistemáticas em bases homogêneas, e seu objetivo seria as ações contínuas voltadas para a otimização dessas perdas de energia.

A busca do conhecimento do nível de perdas de energia de forma estratificada por segmento e ainda segundo suas origens técnicas e comerciais, seria fundamental para evolução de técnicas para alavancar os estudos nessa área.

O valor percentual de perdas a ser apurada foi então definida pela seguinte equação (2.1):

$$P(\%) = \left(\frac{Ee - Es}{Ee} \right) \times 100 \qquad (2.1)$$

Onde:

P(%) = Perdas percentuais

Ee = Energia de entrada

Es = Energia de saída

É fato que os investimentos vinham se reduzindo ao longo dos anos, provocando um gradual incremento do indicador de perdas.

Estas perdas podem ser definidas de maneira geral, como sendo a diferença existente entre a grandeza de entrada (requerida) e a grandeza de saída (vendida).

As perdas podem ser classificadas em duas naturezas: perda de potência/demanda e perda de energia.

A perda de demanda é definida como sendo a diferença existente entre a potência de entrada (requerida) e a potência de saída (vendida), em um determinado período de tempo.

Ressalta-se que a potência é caracterizada como a demanda máxima registrada em um determinado instante.

Assim descreve-se na equação (2.2) a forma de apuração dos valores de perdas referentes as potências/demandas de um determinado sistema em função do tempo.

$$\mathbf{PP(t) = Pe(t) - Ps(t)} \qquad (2.2)$$

Onde:

PP = Perdas de Potência

Pe = Potência de entrada

Ps= Potência de saída

E ainda temos a perda de energia PE(t) que é a diferença existente entre a energia de entrada (requerida) e a energia de saída (vendida) em um determinado período de tempo.

Esta situação pode-se ser verificada através da equação (2.3) que caracteriza as perdas de energia em função do tempo.

$$\mathbf{PE(t) = Ee(t) - Es(t)} \qquad (2.3)$$

Onde:

PE = Perdas de Energia

Ee = Energia de entrada

Es= Energia de saída

Além da identificação das perdas em demanda e energia, essas podem ser ainda desmembradas e segmentadas em dois grupos: perdas técnicas e perdas não técnicas.

Perda técnica, resumidamente pode ser dita como a energia ou demanda perdida no transporte e na transformação. Portanto é inerente ao processo e se caracteriza por ocorrer antes do ponto de entrega.

A outra categoria são as perdas não técnicas, que será chamada a partir de agora de Perda Comercial (PC). Este grupo retrata a energia, ou a demanda efetivamente entregue ao consumidor, ao consumo próprio ou a outra concessionária, mas não são contabilizadas no faturamento.

Esta perda é o principal alvo deste trabalho. Nosso objetivo é contribuir para reduzir ao máximo os valores deste segmento.

2.2 Histórico das perdas

Em contexto geral pode-se afirmar que as perdas localizam-se em diversos segmentos dos sistemas elétricos podendo ser encontradas nas áreas de geração, transmissão e distribuição.

Porém em função das estruturas das empresas do setor elétrico, normalmente o sistema de geração e o sistema de transmissão são tratados como um único segmento.

Desta forma definiu-se a classificação das perdas em dois níveis distintos: perdas na transmissão e perdas na distribuição.

Reforçando este conceito foi definido pelo Comitê de Distribuição (CODI) da Associação Brasileira de Distribuição de Energia Elétrica (ABRADEE), para efeito da apuração dos indicadores gerenciais de perdas a seguinte classificação e definições:

Perdas Globais são as perdas totais de energia elétrica e demanda existente, considerando o conjunto dos sistemas de geração, transmissão e distribuição.

Perdas na transmissão são as perdas de energia elétrica e demanda existente, considerando os sistemas de geração e transmissão.

Perdas na distribuição são as perdas de energia elétrica e demanda existente, considerando apenas o sistema de distribuição.

As perdas na transmissão que contemplam também a parte da geração, conforme já comentado, é o segmento com a menor dificuldade para administração e controle. Isto em função de suas próprias características físicas e um número reduzido de itens de verificação para atuação.

O percentual dessas perdas é razoavelmente pequeno, mas não deixa de ser também um dos contribuintes quando na apuração final das perdas globais do sistema. Estas perdas ocorrem principalmente em função de características de materiais condutores que são utilizados na fabricação dos diversos itens que formam o sistema. Materiais estes que de certa forma são considerados ideais em função do custo benefício. Investimentos em materiais com características de perdas menores acarretariam um aumento significativo no modelo tarifário existente.

A título de ilustração pode-se verificar alguns gráficos referentes a dados históricos das perdas registradas em uma concessionária.

A figura 2.1 mostra os índices das perdas globais registradas no período do ano de 1997 ao ano de 2004, percebe-se que em 2001, ano do racionamento as perdas tiveram uma

redução, no entanto foi uma correlação direta com a redução do consumo e não melhorias de performance no processo.

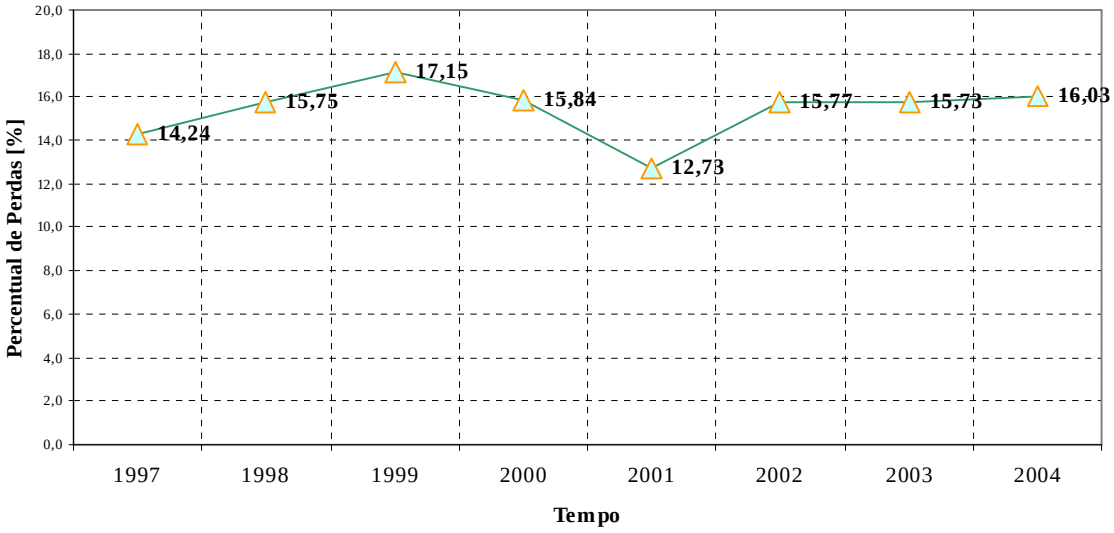

Figura 2.1 – Perdas anuais globais

A figura 2.2 possibilita uma visualização da evolução das perdas globais durante o ano de 2004, nota-se que a variação durante todo o período não ultrapassou a um ponto percentual.

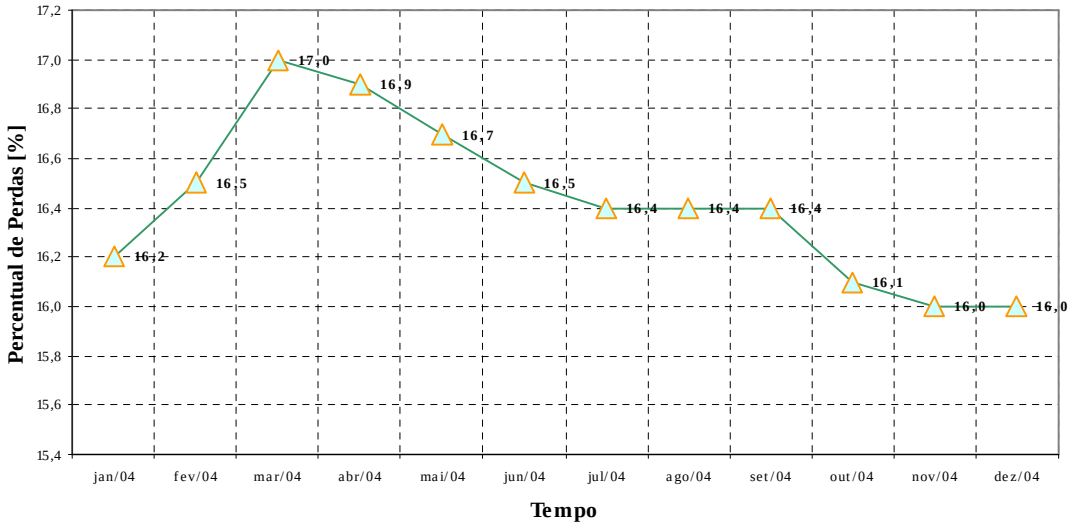

Figura 2.2 – Perdas globais 2004

A partir do conhecimento dos índices globais das perdas, pode-se agora iniciar a estratificação por suas categorias conforme é demonstrado através da figura 2.3. Trata-se de um desmembramento possibilitando um melhor entendimento na mensuração dos índices das perdas nos seus aspectos técnicos e comerciais.

Observa-se que os valores das perdas técnicas são superiores às perdas comerciais, no entanto, estas por sua vez estão diretamente correlacionados a aspectos construtivos, características de materiais, onde o investimento já foi calculado considerando esta situação. Já para as perdas comerciais sugere-se que seus valores deveriam situar próximos a zero, pois correspondem exclusivamente em ações administrativas.

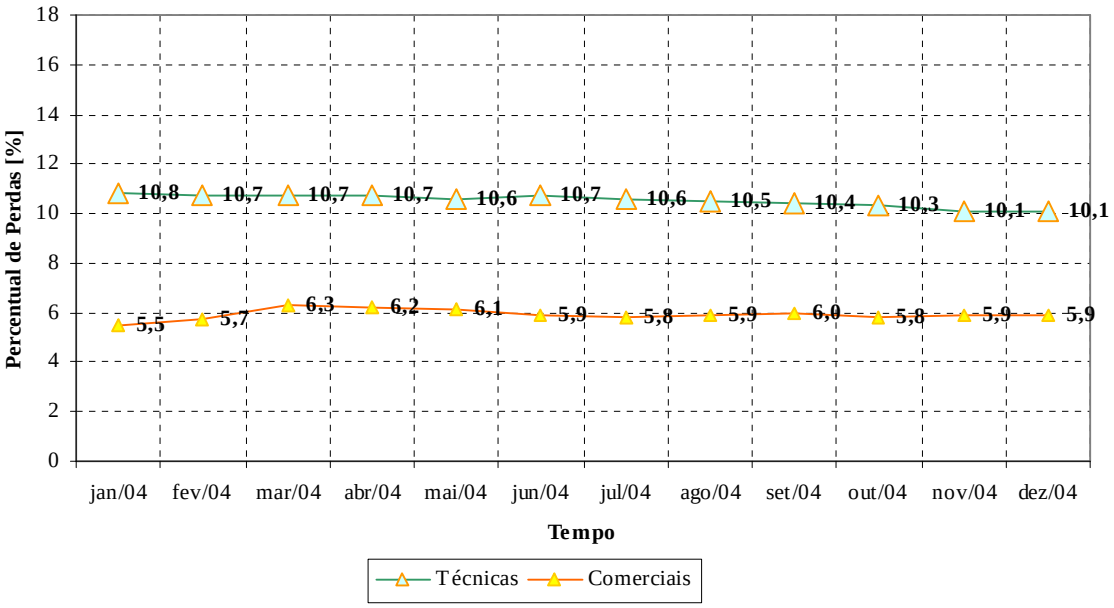

Figura 2.3 – Perdas técnicas e comerciais – global 2004

O grande problema das perdas encontra-se no segmento da distribuição, pois tanto nos aspectos das perdas técnicas como também as perdas comerciais os seus índices são extremamente elevados conforme mostra a figura 2.4.

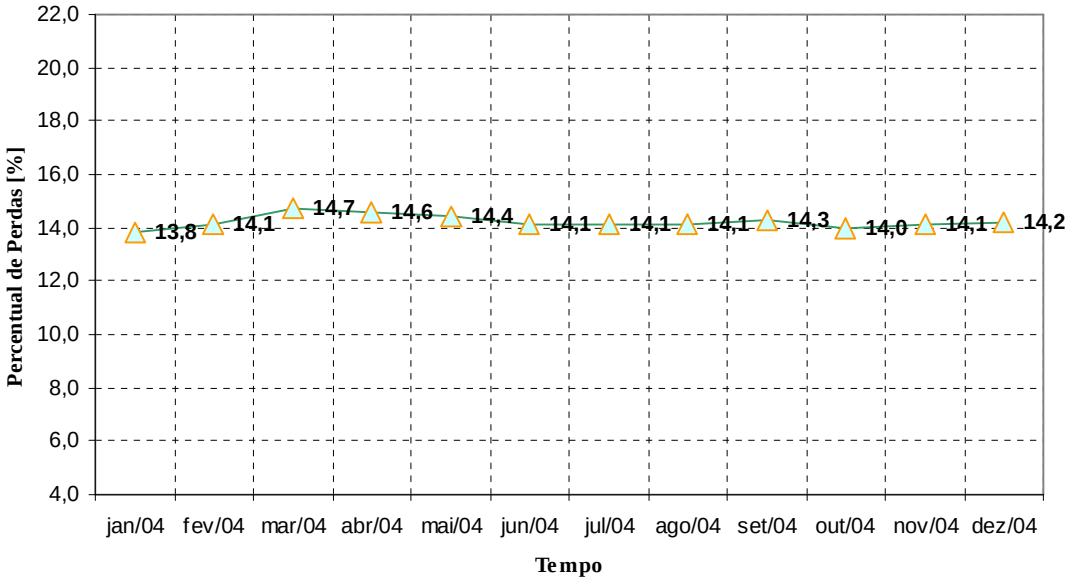

Figura 2.4 – Perdas na distribuição 2004

A figura 2.5 mostra as perdas no seguimento da distribuição separadamente técnicas e comerciais respectivamente, realizadas no ano de 2004.

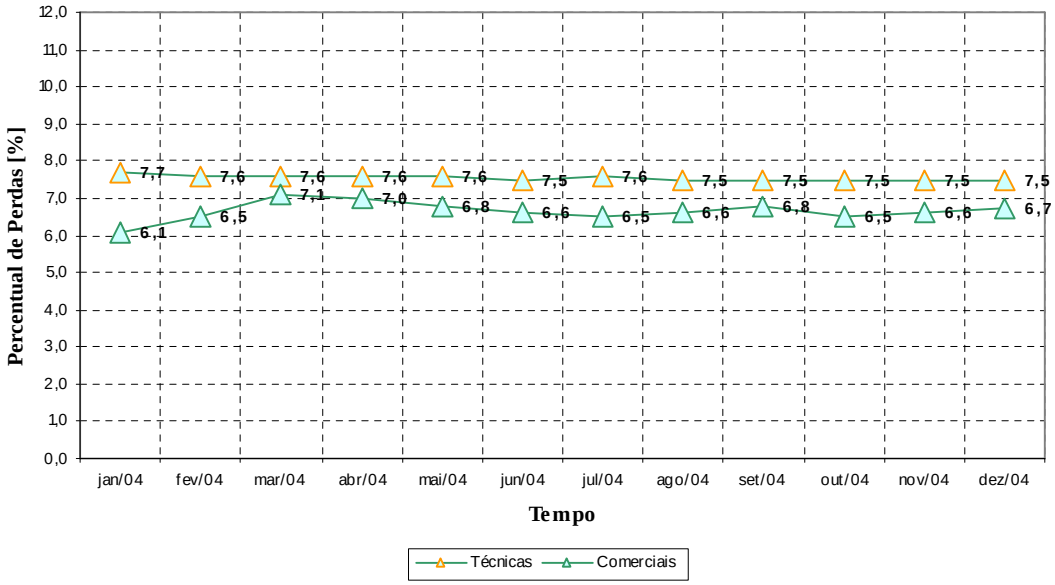

Figura 2.5 – Perdas técnicas e comerciais na distribuição

Depois de apresentado os índices históricos referentes às perdas de uma concessionária de energia percebe-se a representatividade do tema no cenário do setor elétrico nacional.

Na seção seguinte serão discutidas as perdas técnicas e comerciais, abordando ainda seus principais componentes.

2.3 Perdas técnicas na distribuição

Como mencionado anteriormente, as perdas na distribuição se caracterizam como um dos grandes colaboradores no resultado final das perdas globais.

No ambiente de constantes mudanças no setor elétrico, a maioria das distribuidoras tem buscado constantemente o conhecimento das características operacionais dos seus sistemas.

A figura 2.6 mostra um diagrama unifilar de um sistema de distribuição, onde podem ocorrer as perdas técnicas.

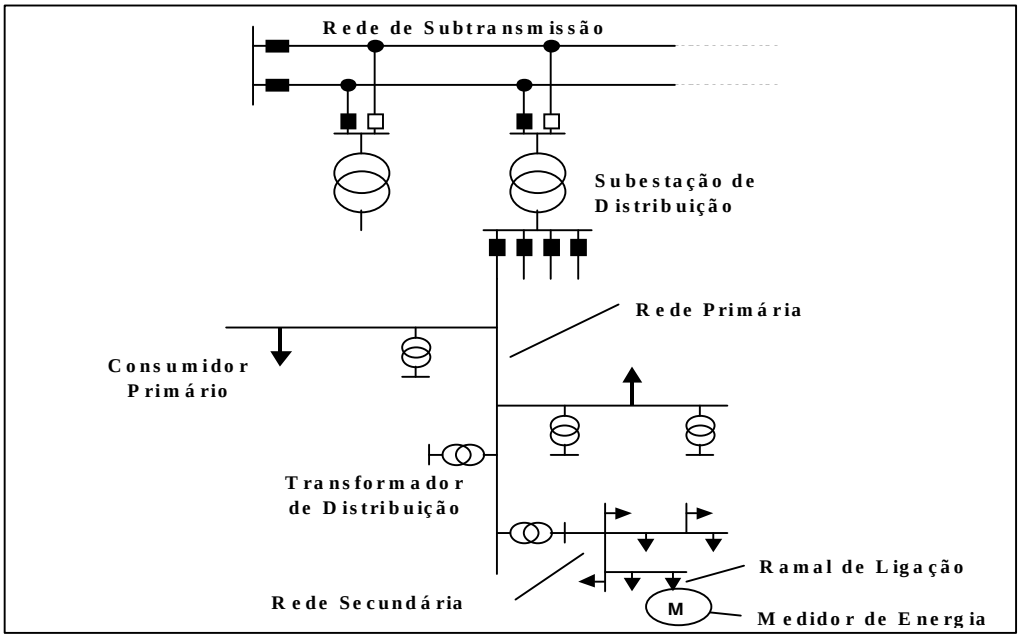

Figura 2.6 - Diagrama unifilar de sistema de distribuição

A cada instante o sistema recebe das subestações responsáveis para seu atendimento um valor de demanda, e fornece aos consumidores um valor de demanda inferior. Essa diferença, entre as duas grandezas, também variável ao longo do tempo, representa perda técnica de demanda no instante (t), conforme pode-se verificar através da equação 2.4

$$\mathbf{PT(t) = D1(t) - D2(t)} \qquad (2.4)$$

Onde:

PT(t) = Perda técnica

D1 = Demanda solicitada

D2 = Demanda fornecida

Estas demandas (D1 e D2) podem ser integralizadas e representadas por energia solicitada e fornecida ao sistema durante determinado período.

A relação das duas grandezas demanda e energia, pode ser obtida através de leituras efetuadas em subestações, pontos de fronteiras e também nos pontos de fornecimento dos consumidores de maior relevância.

A partir dessas leituras é possível identificar o fator de carga, este sendo muito importante para avaliar, no aspecto econômico, o fornecimento e atendimento de cargas, uma vez, que o sistema é dimensionado em função da demanda máxima requisitada.

É muito importante administrar para que o fator de carga se mantenha sempre em índices elevados, pois ele caracteriza o grau de utilização e, portanto, custos unitários (por unidade de energia fornecida).

Otimizar as perdas não só representa liberar investimentos, mas também, melhorar a qualidade do produto oferecido aos consumidores.

Juntamente com o aspecto de credibilidade, no sentido de busca de melhorias no sistema, conseqüentemente resulta em uma tarifa mais justa e adequada aos consumidores.

A avaliação, no que diz respeito às perdas técnicas nos sistemas de distribuição, é bastante complexa. Os principais fatores são decorrentes da grande quantidade de elementos que constituem o sistema, o regime diversificado e aleatório do comportamento das cargas e ainda o contínuo processo de expansão.

A manipulação de todos esses dados e informações para compilação depende de tempo e recursos que podem ser tanto maiores quanto maior os sistemas.

As concessionárias de distribuição utilizam suas bases cadastrais para elaboração e apuração dos resultados. Desta forma essas bases devem estar em condições de fornecer os dados necessários para a avaliação das perdas técnicas, a partir dos diferentes elementos que forma o sistema de distribuição.

É muito importante, e necessário, usar metodologias que utilizem dados e informações que estejam dentro do banco de dados de maneira confiável. O objetivo é atingir um nível de precisão ideal que retrate a realidade do sistema elétrico de distribuição de energia.

Dentre as metodologias utilizadas para o cálculo das perdas técnicas, existem a elaborada e as simplificadas. As elaboradas, como as de gerência de redes e fluxo de carga, apresentam características de resultados que devem se aproximar mais da realidade. Podem ser utilizados para análises individuais e localizadas, exigem uma extensa base de dados e cadastro permanentemente atualizado.

Já as metodologias simplificadas utilizam, na maioria dos casos, de processos estatísticos. Esses processos trabalham com um volume reduzido de dados e permitem a estimativa das perdas de forma aceitável. Essa metodologia é aplicada principalmente aos cálculos das perdas nos condutores da rede primária de distribuição e também, para os transformadores de

distribuição de energia. No caso dos cálculos das perdas nos condutores da rede secundária, é mais adequado o uso de metodologias mais elaboradas.

Nos demais componentes, devido a suas pequenas contribuições no valor total e, em alguns casos, devida a extrema dificuldade para efetuar os cálculos e apurar os resultados, os cálculos deverão ser feitos através de processos simplificados.

O objetivo principal de qualquer sistema que venha a ser utilizado para apuração do cálculo de perdas, deve ter o compromisso entre a precisão dos resultados dos cálculos e o dispêndio de recursos para a obtenção e processamento dos dados necessários. Uma vez obtidos os resultados, deve-se buscar a melhoria, para conseguir a redução das perdas técnicas, ao seu nível ótimo, isto é, aquele do qual nenhum investimento adicional se justifica economicamente, para reduzi-las ainda mais.

As distribuidoras de energia de uma maneira geral classificam as suas perdas técnicas de acordo com o componente elétrico e onde elas ocorrem em função do tempo. A tabela 2.1 apresenta os principais componentes de um sistema elétrico, especificamente de redes de distribuição que serão comentadas detalhadamente a seguir.

Tabela 2.1 - Perdas nos componentes do sistema de distribuição

PERDAS TÉCNICAS NA DISTRIBUIÇÃO	• Condutores da Rede Primária • Transformadores de Distribuição • Condutores da Rede Secundária • Ramais de Ligação • Medidores • Conectores • Equipamentos (capacitores,reguladores de tensão, etc) • Diversas (isoladores, corona, conexões, etc)

2.3.1 Condutores de rede primária de distribuição

Os condutores de rede primária de distribuição, é o meio condutor de energia até as cargas, podendo ser classificados em função da composição do seu material. Na maioria das redes, principalmente devido ao custo benefício, os condutores mais utilizados são de alumínio. Porém pode-se afirmar que existem outros materiais condutores que poderiam ser utilizados. Um desses materiais é o cobre, material cuja composição química permite uma menor perda em relação ao alumínio, mas, por outro lado existem outras variáveis além do seu alto custo, que faz com que o alumínio torna-se o mais recomendado. Outro material muito utilizado é o condutor de alumínio com alma de aço (CAA). Esse tem grandes aplicações em áreas rurais e também em vãos de redes com maiores distanciamentos ou ainda que se tenha que aplicar um maior tracionamento.

Esses circuitos das redes de distribuição são caracterizados e classificados segundo alguns atributos tais como: nível de tensão nominal, resistência do condutor, densidade de carga, temperatura, etc. Tudo isso reflete diretamente no resultado final das perdas até no ponto de conexão com os transformadores de distribuição.

2.3.2 Transformadores

As ocorrências das perdas em transformadores são na maioria das vezes constituídas por perdas no ferro, que dependem da tensão e freqüência, sendo praticamente constantes. No entanto, ocorrem também as perdas no cobre, que estão vinculadas e dependem da carga do transformador. Este equipamento trabalha sujeito a variações constantes em função do aspecto temporal das cargas. O fator de utilização dos transformadores pode ser definido como a potência máxima exigida dos equipamentos em relação ao valor de sua potência nominal.

Estes valores podem ser trabalhados de forma modelada, em que o fator de utilização médio dos transformadores pode ser obtido através da relação entre a potência máxima e o somatório da potência instalada, levando em consideração um erro percentual de variação.

O fator de utilização permite ter uma visão do aspecto carregamento dos transformadores, e conseqüentemente avaliar o resultado das perdas. Existe também a situação do funcionamento dos mesmos, caso estejam trabalhando em regime de sobrecarga, ocasionando aquecimento nos seus enrolamentos.

No caso de transformadores particulares, ou seja, aqueles que são de propriedade de clientes, também são considerados para o efeito de cálculo e apuração das perdas técnicas. Tendo da mesma maneira dos transformadores da concessionária de distribuição a sua contribuição no aumento dos resultados finais apurados das perdas técnicas na distribuição.

Esses transformadores, em caso de algumas concessionárias, representam um número percentual significativo, o qual merece atenção e acompanhamento dos seus carregamentos e também da condição do estado de conservação.

2.3.3 Condutores de rede secundária

Os circuitos secundários, também têm a concentração das perdas em seus condutores, a maioria das cargas/ligações é de características monofásicas, onde a tentativa para o equilíbrio de carregamento é efetuada através da distribuição de ramais de ligações por fase.

Entretanto, cada consumidor detém hábitos e horários de consumo diferenciados no decorrer do dia, tornando-se praticamente impossível garantir o equilíbrio permanente das cargas ao longo de todos os condutores existentes no circuito.

A persistência para a otimização do balanceamento das fases deve sempre ser mantida e afinada através de critérios para ligação de novos clientes para todos os tipos de ligações.

Todas as ligações devem ser adequadamente avaliadas, levando em conta alguns parâmetros, como a sua demanda máxima solicitada, classe do consumidor, número de fios e outros mais.

Também deve ser mantido um acompanhamento mesmo após a ligação, efetuando permanentemente monitoramento dos carregamentos das fases do circuito secundário.

Uma maior credibilidade nos resultados obtidos é possível através de gerência computacional, desde que os cadastros dos dados estejam confiáveis.

É importante também salientar que a redução do nível de desequilíbrio na rede secundária, impacta diretamente na redução das perdas. A melhor distribuição das correntes entre as fases reduz também a queda de tensão na rede de distribuição secundária.

Para amenizar esta situação as concessionárias, sempre que possível, devem adotar a instalação do transformador no centro de carga do circuito secundário.

Como é praticamente impossível afirmar que o transformador esteja, a todo instante, no centro de carga, em função da grande variação temporal das mesmas, a condição do fechamento em anel dos circuitos é uma alternativa que também pode resultar na diminuição das perdas.

2.3.4 Ramais de ligação ou ramais de serviço

Os ramais são os condutores que fazem o transporte de energia elétrica a partir do poste da rede secundária de distribuição até o padrão de entrada da unidade consumidora.

Normalmente, as medidas máximas em extensão chega a 30 metros de comprimento e seus materiais mais utilizados são: alumínio, WPP, cabo multiplexado e raramente o cobre.

A grande quantidade dos ramais de ligação e ainda a dificuldade em levantar dados para detalhar uma avaliação são fatores que dificultam a apuração dos resultados.

Na maioria das situações, utiliza-se calcular as perdas nos ramais, considerando médias de alguns parâmetros tais como: resistência das fases dos ramais, corrente média que circulam por eles e ainda as cargas em função do tipo de ligação, monofásicos, bifásicos e trifásicos e ainda suas classes: residenciais, comerciais, industriais, etc.

Os ramais de ligação devem ser adequadamente dimensionados em função da carga e conseqüentemente da corrente que irá percorrer esse meio condutor. O objetivo desta ação é evitar o sobre-aquecimento, propiciando assim, uma redução nas perdas.

2.3.5 Medidores

Os medidores de energia elétrica têm a finalidade básica de registrar os consumos de energia elétrica ativa e também reativa, no caso dos consumidores de média tensão.

Além disso, no caso dos medidores eletrônicos, eles registram uma diversidade de variáveis que podem ser obtidos através de leitora, e descarregados em micro-computadores.

A perda gerada pelos medidores convencionais é normalmente definida pela potência absorvida por suas bobinas.

Por outro lado, os medidores eletrônicos, que inclusive já estão sendo fabricados para atender o mercado de baixa tensão, têm características diferentes em relação aos convencionais, possibilitando uma redução significativa dentro do segmento de perdas técnicas em medidores.

O valor de perdas abordadas nesta categoria em geral tem participação bastante discreta no resultado final da apuração das perdas na distribuição. No entanto se considerarmos o grande número de medidores existentes nas diversas concessionárias de distribuição do país pode-se chegar à conclusão que ações devem ser tomadas no intuito de minimizar as perdas nesses equipamentos.

Sugere-se que dentre algumas ações, a substituição gradativamente dos equipamentos de medição antigos por medidores com maior eficiência, conseqüentemente proporcionando menores perdas.

2.3.6 Conectores

Os conectores são responsáveis pelas diversas conexões existentes nas redes de distribuição das concessionárias de energia.

Essas diversas conexões cujas resistências nominais representam uma outra parcela que certamente contribui para o aumento das perdas técnicas, com o passar do tempo ficam velhas, desgastam, causam oxidação, ocasionando um acréscimo de resistência elétrica e proporcionando uma parcela adicional nas perdas. Este acréscimo está diretamente relacionado ao desgaste natural em decorrência do tipo de material, à temperatura ambiente, à oxidação, à tensão de compressão aplicada, à expansão térmica do material, entre outros.

Alguns tipos de conectores são bastante utilizados nas redes de distribuição, entre eles estão: conector cunha, conector compressão, conector paralelo, conector parafuso fundido etc. Todas as categorias possuem características particulares que estão correlacionadas com a sua resistência oferecida ao sistema e com a sua vida útil, impactando nas perdas neste segmento.

Porém, alguns fatores podem contribuir para que a evolução do desgaste natural seja reduzida ou pelo menos estabilizada. A questão da mão-de-obra é o principal fator, pois o profissional que utiliza os conectores deve ser capacitado para a tarefa propriamente dita.

A falta de capacitação profissional pode acarretar aumento das perdas em função da diversidade de características que se devem observar no momento da conexão, pois o conector é o elo de ligação entre materiais com composições químicas diferentes.

As principais observações são com relação às bitolas diferentes, existindo a situação da compressão ou aperto adequado, ou ainda a questão de espaçadores/separadores a serem utilizados no local correto, quando do uso de conectores.

Tudo isso pode vir a acarretar oxidação, aquecimento, fuga de corrente, com conseqüências diretas, no aumento das perdas.

2.3.7 Equipamentos

Alguns equipamentos como capacitores e reguladores de tensão, instalados ao longo das redes de distribuição, contribuem de forma discreta nas perdas dos sistemas.

O número de reguladores de tensão instalados é pequeno. São equipamentos de porte robusto e utilizados onde se verificam problemas de níveis de tensão, ou ainda em subestações antes da saída dos alimentadores.

O custo do equipamento é considerado alto, acarretando para algumas situações soluções mais comuns, como a utilização de condutores de maiores bitolas.

Já os bancos de capacitores são equipamentos com grande utilização em todo o sistema, com custo mais acessível, e sua principal função é reduzir o efeito reativo do sistema, elevando o fator de potência.

As perdas de energia elétrica ocorrem em forma de calor e são proporcionais ao quadrado da corrente total. Como essa corrente cresce com o excesso de energia reativa, pode-se estabelecer uma relação direta entre o incremento das perdas e o valor do fator de potência, entretanto, não causa impacto significativo nas perdas.

Mesmo com a pequena interferência pode-se atuar nos bancos capacitores efetuando alguns arranjos em seus esquemas de ligação, permitindo resultados diferentes no fator de

potência do sistema e por conseqüência nas perdas, respeitando sempre as características técnicas das redes.

2.3.8 Perdas diversas

Entre a categoria de perdas diversas pode-se mencionar aquelas decorrentes das perdas por corrente de fuga, por efeito corona, ou ainda por fenômenos transitórios que podem ocorrer em qualquer componente do sistema.

A avaliação dos referidos tipos de perdas fica vinculada a outras situações e variáveis como a qualidade da manutenção, características do equipamento, qualidade do ar ambiente em função de tipos de poluição presentes na atmosfera, tensão, projetos etc.

Desta forma, todas essas variáveis tornam o cálculo das perdas muito complexo, então utiliza-se para esse segmento metodologias mais simples no processo do cálculo, devido à pequena participação no valor total das perdas.

Processos simples podem ser de forma estimativa, a partir de instrumentos que permitam a gerência dos sistemas de redes. Esses instrumentos devem estar constantemente em fase de melhoria e refino, possibilitando assim, uma proximidade maior de resultados obtidos com a situação real.

Ainda dentro do contexto dessas perdas podemos mencionar aquelas provenientes de desequilíbrios e harmônicos, que trata-se de fenômenos que podem gerar acréscimo e ou variações de corrente e tensão nos sistemas.

Esses eventos podem ocasionar também aumento das perdas técnicas, contribuindo para uma majoração nos resultados globais.

2.4 Perdas comerciais na distribuição

As perdas comerciais são aquelas oriundas da energia efetivamente entregue aos consumidores finais ou a outras concessionárias, mas não computadas nas vendas da empresa.

Essas perdas são contabilizadas a partir do resultado da diferença da energia faturada e registrada, descontada a perda técnica. Observa-se através da equação 2.5 as componentes para cálculo das referidas perdas.

$$PC = Er - Ef - PT \tag{2.5}$$

onde:

PC = Perda comercial

Er = Montante de energia registrada pela empresa (geração própria + compra)

Ef = Total energia faturado pela empresa

PT = Perda técnica

As perdas comerciais são merecedoras de uma atenção especial, em função de aspectos que serão comentados posteriormente em seus respectivos grupos.

Do ponto de vista empresarial essa categoria de perdas não significa somente perdas de receitas, mas elas têm conseqüências de maior gravidade, pois geram tarifas maiores, que automaticamente, também acarretam um aumento da inadimplência e atrasos em entradas de receitas no caixa da empresa.

Tudo isso em época anterior, não despertava grandes preocupações, pois no passado havia o ressarcimento de todos os prejuízos acarretados por perdas, em favor das concessionárias, através da reposição na tarifa.

Com o início das privatizações as concessionárias do Setor Elétrico Brasileiro começaram a se preocupar com a otimização de suas receitas e melhorar a qualidade dos serviços prestados aos seus clientes.

Nesse contexto, estabeleceram algumas prioridades, dentre elas, a apuração dos desperdícios que ocorrem nos seus respectivos sistemas elétricos visando promover ações sistemáticas que permitam sempre a sua redução.

As perdas comerciais podem ser originadas e classificadas em diversos segmentos conforme será descrito no sub-item a seguir, o conhecimento da origem e causas das perdas é fator essencial, pois permite uma análise do investimento necessário para sua redução bem como do retorno a ser obtido.

2.4.1 Ligações clandestinas

As ligações clandestinas são aquelas realizadas a revelia da empresa responsável pelo fornecimento de energia elétrica, ou seja, sem autorização e nem aprovação da concessionária.

Essas ligações devem merecer ações de combate específicas, pois são encontrados diferentes tipos de casos. O infrator executa a conexão de fios e ou cabos à rede da concessionária de forma clandestina e irregular, e a energia é utilizada sem nenhum registro pela concessionária. Estes casos na maioria das vezes são os que provocam os maiores números de acidentes aos infratores, pois quase sempre sobem nos postes para efetuar a ligação dos fios à rede.

Mas independente de qualquer situação sempre existirá o risco que na maioria dos casos é sucumbido pelo infrator em função das necessidades e dificuldades financeiras para o pagamento dos ônus e taxas pertinentes à ligação e ao consumo de energia.

Restam à concessionária, como meio para a localização dessas ligações, as informações dos empregados que atuam na área e equipes de manutenção/plantão de atendimento de emergência. Também as denúncias da população em geral podem contribuir, podendo ser incentivadas através da divulgação e conscientização dos consumidores, que as perdas acabam incidindo na tarifa de energia na medida que onera seus custos operacionais.

Constatada esses tipos de ligações, essas devem ser desfeitas e os condutores recolhidos. A retirada deve ser acompanhada de intimação do responsável pelo ressarcimento dos prejuízos causados, estimados através do levantamento da carga existente no momento da inspeção, aplicando-se a legislação vigente. O responsável deve ser alertado quanto à gravidade do ato, que é um crime previsto no código penal e passível de prisão.

Uma das alternativas utilizadas atualmente na regularização dessas situações é a possibilidade de financiamento do padrão de medição, de forma a facilitar a regularização dessas ligações clandestinas, permitindo o benefício de energia a toda a sociedade.

Essas atuações devem ocorrer principalmente nas áreas com alta incidência de ocorrências de ligações irregulares. Área onde normalmente os problemas econômico-social são predominantes deve ser desenvolvida a conscientização da comunidade quanto a necessidade de que todos paguem a energia consumida e utilizem esta de forma segura.

A figura 2.7 mostra casos reais de ligações clandestinas em rede e unidade consumidora.

Figura 2.7 – Ilustrações de ligações clandestinas

2.4.2 Intervenções indevidas no padrão e na medição

A intervenção indevida no padrão de entrada e no seu respectivo sistema de medição tem a finalidade de alterar e ludibriar o registro da energia efetivamente utilizada.

A redução do montante de energia não faturada devido às intervenções ilícitas do consumidor é um dos fatores importantes para a minimização das perdas comerciais das concessionárias de energia elétrica.

Porém, existem dificuldades para a solução do problema em função das diversas técnicas utilizadas. Observa-se a seguir um relato sucinto do que vem a ser algumas delas.

2.4.2.1 Irregularidade no ramal de ligação

Derivação de energia, efetuada nos condutores que interligam o poste da concessionária ao padrão de entrada de serviço da unidade consumidora.

A figura 2.8 mostra casos reais de irregularidades executadas no ramal de ligação de unidades consumidoras.

Figura 2.8 – Ilustrações de irregularidade no ramal de ligação.

2.4.2.2 Irregularidade no ramal de entrada

Derivação de energia efetuada na fiação compreendida, entre o eletroduto de entrada do padrão e a caixa de medição. Esse método de desvio de energia normalmente é efetuado no interior da parede.

A figura 2.9 mostra casos reais de irregularidades executadas no ramal de entrada de unidades consumidoras.

Figura 2.9 – Ilustrações de irregularidade no ramal de entrada.

2.4.2.3 Irregularidade no disjuntor

Derivação de energia, praticada nos bornes do disjuntor, seja na entrada ou na saída do mesmo conectando condutores de derivação clandestina, antes do registro do medidor.

A figura 2.10 mostra casos reais de irregularidades executadas no disjuntor de proteção da unidade consumidora.

Figura 2.10 – Ilustrações de irregularidade no disjuntor.

2.4.2.4 Irregularidade no medidor

As Irregularidades nos medidores podem ser praticadas a partir da violação do lacre e retirada da tampa de vidro do medidor. Mas também ocorrem através de outros artifícios sem qualquer violação. Assim é efetivada a prática de diversas ações provocadas de formas intencionais com o objetivo de alterar o registro da energia efetivamente consumida pela unidade consumidora.

Descrevem-se abaixo as irregularidades mais comuns:

- Manipulação dos ponteiros alterando-se o consumo registrado;
- Introdução de qualquer objeto que altere o giro do disco do medidor;
- Atuação no interior do medidor, fiação, bobinas de corrente e/ou de tensão;
- Atuação na fiação de ligação do medidor etc.

A figura 2.11 mostra casos reais de irregularidades executadas em medidores instalados em unidades consumidoras.

Figura 2.11 – Ilustrações de irregularidade no medidor.

2.4.2.5 Religação à revelia

Religação clandestina de unidade consumidora sem autorização da concessionária, desligada a pedido do consumidor ou por iniciativa da própria concessionária.

2.4.3 Medidores

Os problemas em medidores são de grande relevância. As concessionárias distribuidoras de energia têm grande número de medidores que são antigos e ultrapassados. Estes medidores tendenciam a ter o seu nível de precisão reduzido, fazendo que os seus registro de consumo sejam na maioria das vezes inferiores ao que foi efetivamente realizado.

A experiência tem demonstrado a existência de medidores com altos índices de erros, gerando perdas de energia.

Os testes de normas não conseguem determinar a vida útil precisamente e nem a redução da precisão dos medidores ao longo do tempo. Estima-se que os eletromecânicos devem ser recertificados em 15 anos e os eletrônicos 30 anos.

Essas medições que a variação do erro esteja superior ao permitido e determinado pela legislação que atualmente está definido entre +3,0 % e –3% provocam acréscimo na apuração do resultado das perdas. A substituição desses equipamentos deve ser providenciada gradativamente para evitar impactos nos custos das concessionárias.

Existem também os equipamentos de medição que estejam em situação de defeitos ou propriamente dito desregulados, isso pode ocorrer em função de situações de transporte, defeito de fabricação e aferição em laboratório.

O aumento do grau de exigência dos consumidores e dos órgãos metrológicos já é fato, desta forma, sugere-se que antes de uma possível exigência externa, as empresas busquem se capacitar, adequando seu quadro de profissionais e seus laboratórios, para possibilitar o atendimento dessa demanda.

Mostra-se na tabela 2.2 uma amostragem dos tipos de ocorrências de irregularidades identificadas em medidores de energia.

Tabela 2.2: Irregularidades com perda em medidores

Tipo de ocorrência	Nº	%
Desvio antes do medidor	150	23,59
Medidor com defeito	140	22,01
Circuito de potencial interrompido	114	17,92
Desvio embutido na parede	94	14,78
Medidor com selo violado	40	6,28
Medidor danificado	30	4,72
Medidor com disco parado	24	3,78
Constante errada	22	3,46
Fraude na chave de aferição	22	3,46

As tabelas 2.3, 2.4 e 2.5 indicam uma avaliação percentual de diagnóstico amostral de ocorrências em medidores estratificados por segmentos.

A tabela 2.3 demonstra o percentual das perdas localizadas em medidores monofásicos, bifásicos e trifásicos. Observa-se que a maior perda está localizada no seguimento trifásico como já esperado, pois nessa modalidade esta localizada as maiores cargas.

Tabela 2.3: Perdas estimadas por fases

Fases	Perda Mensal (MWh)	%
Monofásico	5.344	27,46
Bifásico	1.530	7,86
Trifásico	12.584	64,68

A tabela 2.4 estratifica a origem da perda, estas por sua vez ficaram praticamente equiparadas percentualmente, desta forma nota-se que as perdas em medidores podem ocorrer tanto em função de fraude ou ainda problemas em equipamentos.

Tabela 2.4: Perdas estimadas por origem

Origem	Perda Mensal (MWh)	%
Intencional	9.465	48,64
Não intencional	9.993	51,36

A tabela 2.5 categoriza as perdas por seguimentos de classificação, conclui-se que a grande maioria das perdas em medidores de energia está localizado nos seguimentos Residencial e Comercial.

Tabela 2.5: Perdas estimadas por classe

Classe	Perda Mensal (MWh)	%
Residencial	9.579	49,23
Comercial	4.948	25,43
Industrial	546	2,81
Rural	3.785	19,45
Outras	600	3,08

2.4.4 Medições indiretas

As medições indiretas possuem, além dos medidores, equipamentos complementares para efetuar o registro do consumo de energia elétrica e demanda.

Nessas unidades consumidoras são encontrados casos de irregularidades que envolvem grandes técnicas e conhecimento para elaboração.

Quando é identificada situação de irregularidade nessa categoria de unidades consumidoras, percebe-se que na maioria dos casos estão envolvidos quantias financeiras vultosas e grandes valores em energia.

2.4.5 Perdas comerciais de origem administrativas

As perdas advindas das áreas administrativas são decorrentes de procedimentos incorretos ou equivocados. Pode-se afirmar que essa modalidade de perda engloba parcelas de energia que não estão sendo faturadas devido a uma deficiência na gestão administrativa nas distribuidoras.

Descreve-se abaixo algumas das principais atividades executadas por operadores que de alguma forma contribuem e influenciam negativamente nos índices apurados no processo das perdas comerciais.

Erro da leitura do medidor

Trata-se das perdas oriundas de erros de leituras que podem ocorrer devida falha humana, no caso do profissional que efetua a leitura do medidor de energia, ou deficiência técnica das instalações proporcionando uma má visualização e conseqüentemente uma interpretação incorreta dos registradores.

Erros de faturamento

Os erros de faturamento, em algumas situações são compensados no mês posterior a ocorrência, não acarretando assim perdas de energia. No entanto existem situações que envolvem o encerramento das atividades nas instalações, onde é solicitado o consumo final para a unidade consumidora, desta forma os erros não poderão ser compensados devido ao encerramento do fornecimento.

Quando se efetua a troca de medidores, pode ocorrer deficiência de sistema ou de procedimentos gerando uma perda decorrente da não cobrança do resíduo de consumo do medidor retirado.

Existe ainda uma outra situação que contribui para um faturamento inadequado, quando leituras de medidores são faturadas com base em valores de consumo informado e não no valor de leituras registradas, esta situação é predominante na área rural.

Unidades fora de faturamento

A existência de unidades consumidoras ligadas pela empresa em seu sistema de distribuição, no entanto, devido a problemas de procedimentos internos das concessionárias se encontram fora de faturamento.

Constante de faturamento

Os erros referentes a constantes de faturamentos refletem de maneira direta no consumo apurado e conseqüentemente na receita faturada. Tal constante é o fator multiplicador da

diferença entre a leitura anterior e a leitura verificada no momento atual registrada no medidor. Para as ligações diretas, as constantes normalmente são unitárias com exceção de algumas medições trifásicas que tem valores diferentes.

Já para os casos de ligações indiretas os valores das constantes da medição são diversos, em função das próprias características das modalidades tarifárias oferecidas para essa categoria de consumidores.

Desta forma é muito importante a correta informação dos valores das constantes, pois estas podem acarretar perdas significativas no faturamento.

2.4.6 Falta de medição

Dentre as diversas causas das perdas comerciais analisadas, a falta de medidores ocorrem em várias concessionárias sendo este evento responsável por uma parcela das perdas.

Por outro lado, esta causa permite uma análise de custo-benefício mais precisa em relação a outras, facilitando assim uma tomada de decisão.

A definição deverá ser balizada no custo da aquisição de novos medidores ou a repotencialização de medidores defeituosos.

Desta forma as empresas concessionárias devem efetuar a instalação gradativa desses medidores, pois se sabe que a ausência de medição de energia leva a um consumo exagerado, provocando desperdícios de energia.

2.4.7 Cargas especiais sem medição

Algumas unidades consumidoras denominadas como cargas consideradas especiais são instaladas sem medição.

São elas: relógios digitais, semáforos, lombadas eletrônicas e radares, iluminação de out-doors, entre outros. Normalmente os consumos destas cargas, podem ser calculados com uma boa precisão. No entanto, o que se observa na prática é que, com muita freqüência, são realizadas alterações, principalmente acréscimo nas cargas informadas sem o devido conhecimento da concessionária.

Diante desse fato, deve-se buscar a instalação de equipamentos para registro de consumo sempre que possível, mesmo para estas modalidades de ligação que representem pequena monta de consumo.

2.4.8 Perdas na transformação

Além das questões das fraudes existentes nos consumidores do grupo A (alta tensão), há outra preocupação das concessionárias, são as perdas de transformação. A concessionária em algumas situações não instala os equipamentos de medição para apuração dessas perdas, praticando os acréscimos adicionais autorizados pela resolução 456 da ANEEL.

Aplica-se para os consumidores atendidos em tensão igual ou inferior a 44 kV um fator de correção de 2,5 % (dois e meio) e 1% (um) para os consumidores com tensão superior a 44 kV sobre os valores medidos de energia e demanda prevista na legislação.

Na maioria dos casos os valores praticados não são suficiente para cobrir a diferença a maior das perdas originadas pelos equipamentos.

Uma das alternativas para minimizar essas perdas é a exigência da apresentação de laudos de ensaios dos transformadores novos como para reformados ou usados.

2.4.9 Perdas em iluminação pública

As perdas em iluminação pública podem ser verificadas nas lâmpadas, nos reatores, e nos relés fotoelétricos. Essas perdas podem ser significativas, já que os valores apurados são baseados em consumos estimados. Nesta modalidade não são instaladas medições; normalmente é considerada a média do número de horas em funcionamento da potência total das lâmpadas instaladas, mais os acréscimos estimados das perdas. Essa categoria de iluminação é localizada em locais públicos, como logradouros, vias, praças, avenidas, ruas, canteiros centrais, etc.

A seguir estão descritas as perdas em cada um dos elementos que constituem esta categoria:

Lâmpadas

As lâmpadas encontradas na iluminação pública têm uma diversidade de características, podendo ser, incandescente, fluorescente, fluorescente compacta, mistas, vapor metálico, vapor de mercúrio, vapor de sódio etc. As potências também são bastante variadas, sendo as mais usuais 80W, 125W, 150W, 250 W, 300W, 400 W, 500W, 1000W.

Muitas vezes os valores de energia cobrados pelas concessionárias referentes à iluminação pública são sub-dimensionados, devido à falta de controle das alterações introduzidas no sistema de iluminação.

A falta de um programa para manutenção do sistema de iluminação pública, com o objetivo de evitar o consumo não faturado/desperdício advindos de lâmpadas que permanecem acessas durante o dia aumenta também as perdas. Outra questão importante é a correta instalação dos equipamentos auxiliares que serão comentados posteriormente.

Reatores

Para quase todas as categorias de lâmpadas é necessário um reator, que também é um equipamento agregado ao sistema e contribui para o acréscimo do conjunto. As perdas referentes à operação dos reatores de iluminação pública devem ser consideradas quando da composição do consumo mensal, conforme previsto em normas brasileiras reguladoras NBR. Porém, podem não representar as situações reais, tornando-se mais um ponto de erro dentro do contexto de avaliação das perdas.

Estudos efetuados a partir de medições realizadas nos sistemas de iluminação pública em funcionamento, observaram que o valor das médias das perdas nestes reatores é da ordem de 13,5% maior que a potência faturada por reator especificada em norma. Verifica-se também que as perdas no reator são agravadas pelo baixo fator de potência da carga. Para que a diferença, entre o faturamento e a energia consumida, permaneça em níveis razoáveis é fundamental que a tensão permaneça próxima da tensão nominal.

Relés fotoelétricos

Outro ponto de perdas na iluminação pública deve-se à má qualidade e o modo incorreto de instalação dos relés fotoelétricos nos postes de iluminação pública. As células dos relés são componentes extremamente vulneráveis aos surtos de tensão, danificando-se facilmente e fazendo com que as lâmpadas permaneçam acesas durante o dia.

A instalação desse equipamento deve ser efetuada corretamente, observando a posição das células, de forma que a mesma não acione o funcionamento da lâmpada durante o dia.

Outra ação a ser desenvolvida refere-se à instalação de relés de melhor qualidade e que garantam uma melhor sensibilidade aos níveis de iluminação, evitando as antecipações e atrasos no funcionamento da iluminação.

2.4.10 As perdas no ponto de vista jurídico

As perdas comerciais nesse item correspondem ás ações irregulares ou ilícitas que são praticadas pelo consumidor. O objetivo dessas ações é a redução no valor financeiro da fatura de energia, desembolsando valores menores que o efetivamente devido pela energia consumida.

A conseqüência será um valor de energia apurado menor que o devido, uma vez que parte da energia consumida não é registrada pelos medidores. A utilização destes métodos irregulares ou ilícitos traz transtorno para ambas as partes, concessionária e causador da irregularidade.

Para a concessionária cabe a aplicação das sanções previstas e para o causador os custos e o desgaste que os processos judiciais demandam.

A análise jurídica se verifica em dois campos, no intuito de combater tais irregularidades. Sob o aspecto do Direito Penal, segundo o Código Penal Brasileiro, e conforme o descrito em seu Art. 155, *equiparão mesmo a coisa móvel à energia elétrica ou qualquer outra que tenha valor econômico, qualificando-a como objeto do crime de furto, diz ainda: "Subtrair, para si ou para outrem, coisa alheia móvel, cabe pena de reclusão, de um a quatro anos, e multa*", estipulando assim, a pena cominada ao delito, e por conseqüência ao indivíduo que o pratica.

Quanto à natureza administrativa, a Resolução ANEEL 456, de 30/11/2000, no Art. 30, inciso I, permite ao concessionário, a suspensão do fornecimento de energia, no caso comprovado de fraude/furto, sem prejuízo das sanções penais cabíveis e da correspondente responsabilidade civil.

A concessionária faz o uso de alguns artifícios, no âmbito de atuação jurídica, para combater as fraudes/furtos de energia elétrica, estes artifícios são relacionados a seguir.

- Para a situação, na qual são apuradas a materialidade e autoria do delito, ou indícios que possam a isso levar, as distribuidoras repassam para autoridade policial os elementos necessários à instauração de inquérito policial.
- Posteriormente, se for o caso, apresenta por meio do Ministério Público a denúncia, que caso aceita, dará ensejo à ação penal na qual a concessionária lesada poderá figurar como assistente de acusação. A tutela jurídico-penal do estado poderá condenar o autor do furto à pena de restritiva de liberdade e a pagamento de multa.
- A irregularidade também pode ainda ser tratada na esfera administrativa, conforme previsto na Resolução 456/ANEEL/2000 no art. 72, que permite à concessionária que execute a suspensão do fornecimento de energia elétrica ao consumidor irregular.
- É emitido o termo de ocorrência de irregularidade em formulário próprio, contemplando todas as informações necessárias para seguimento do processo.
- E ainda recuperar com essa sanção administrativa, o montante que deixou de faturar, bem como regularizar a medição de consumo.

É prudente notar que esta medida não impede ou desobriga da mesma, eventualmente ser seguida de ação judicial para a cobrança de valores devidos. A preocupação das concessionárias está muito mais voltada para o recebimento dos valores que deixou de faturar em razão da irregularidade, do que com a condenação penal dos consumidores infratores.

Para cada situação deverá ser avaliada a questão do custo benefício do procedimento a ser adotado, levando-se em consideração a realidade policial e da própria concessionária.

Mas sempre que houver condições, deverá acionar as autoridades policiais para conhecimento do fato ilícito para que este seja registrado em boletim de ocorrência, visando a posterior instauração do inquérito policial.

Para as situações, em que o consumidor estiver ciente da irregularidade constatada pela concessionária, na hipótese mais comum, ressarci à concessionária.

A instalação da unidade é regularizada e o caso é encerrado na esfera administrativa, sem discussão sobre a autoria do delito, e não é adotada nenhuma providência de ordem jurídica (cível ou penal). O caso torna-se mais complexo quando não ocorre a confissão, pois a autoria não estará demonstrada, restando à concessionária somente prova testemunhais, periciais e documentais que são de difícil consecução e seu valor como prova é relativo. De qualquer forma, a obrigação de provar a autoria não é responsabilidade da concessionária, mas sim das autoridades policial e judiciária. A concessionária deve prestar o apoio necessário no sentido de auxiliar as autoridades, mesmo porque tem pleno interesse no desfecho do processo e, ainda, a inibição de tais situações. Sabe-se que a impunidade traz como conseqüência mais danosa a repetição da infração.

A sociedade de um modo geral deve ter o conhecimento de que um infrator foi penalizado em função da falta que cometeu, não somente no que se refere a furto de energia, mas sim, em todos os segmentos onde existirem regras e leis a serem cumpridas.

A concessionária deve atuar e dar atenção ao aspecto criminal da subtração de energia elétrica, exatamente com a finalidade de exigir do Poder Público, a aplicação do castigo ao delinqüente para que dessa punição outros tomem como exemplo, e se abstenham de delinqüir.

O retorno da comunidade perante as atitudes tomadas pela concessionária ocorrerá de forma lenta, não surtindo efeitos imediatos, ou seja, é um processo de conscientização que ocorre gradativamente.

No entanto, se alguns casos mais relevantes de irregularidade na medição ou furto de energia elétrica forem submetidos ao procedimento policial, e seguidamente ao processo Judicial Criminal, em médio prazo, o número de infrações será reduzido substancialmente.

Mas para que esta situação seja concretizada é necessário que as concessionárias tomem a devida consciência, de não somente se preocuparem em receber valor que deixou de faturar.

Não deve ser observado exclusivamente o aspecto econômico, mas também se engajar no combate ao crime, buscando adotar medidas capazes de apurar inquestionavelmente a materialidade do delito. E ainda, a concessionária deverá atuar de maneira pró-ativa, mantendo mecanismo interno de controle, bem como mecanismos externos de trocas de informações com outras concessionárias.

A energia elétrica trata de algo invisível, dependendo de amplo conhecimento técnico para especificação de variáveis como valores e quantidades, gerando então margens de dúvida e credibilidade para leigos. Mas, independentemente dessas dificuldades, o fato é real e deverá ser apurado, aplicadas as ações pertinentes e justas, de forma a manter um relacionamento adequado e transparente entre consumidores de energia e concessionária.

2.5 Combate às irregularidades

O combate à irregularidade na medição e ao furto de energia, baseia-se em programas de inspeções em entradas de serviço de unidades consumidoras.

Este combate retrata a efetiva disposição da empresa na detecção de situações irregulares, porém a maneira de definição para inspeção ainda é bastante deficiente.

As perdas comerciais já vem sendo enfrentadas pela empresas distribuidoras de energia de várias formas. Dentre elas, podemos citar a autuação em flagrante, divulgações na mídia, disponibilidade de denuncia em internet e através de call center 0800 e a principal, a inspeção em campo.

As inspeções em campo além de caras, são demoradas e requerem a mobilização de pessoal e recursos para a constatação da fraude e autuação em flagrante do infrator.

Serão descritos a seguir os principais tipos de inspeções realizadas pela empresa nas unidades consumidoras com o objetivo de identificar e combater as perdas.

2.5.1 Inspeções de varredura

Este tipo de inspeção abrange especialmente os consumidores classificados em modalidade tarifária residencial e comercial de pequeno e médio porte. Essas inspeções são direcionadas para áreas na qual a perda comercial é, sabidamente, elevada. Uma vez selecionada a área de atuação, são reunidas equipes de inspeção e autuação, que em conjunto, percorrem todas as unidades consumidoras da área, realizando inspeções individuais e constatando de imediato as fraudes. O tempo necessário para realização de um lote completo de inspeções de varredura varia de um a três meses dependendo da área escolhida.

Nesta modalidade de inspeções, todas as unidades consumidoras são visitadas, sendo necessária uma enorme quantidade de inspeções. Isto é um procedimento caro, demorado e nem sempre tão eficaz. O processo inicia-se em determinada parte da região sendo investigada. Como são áreas extensas (por exemplo, um bairro), a vizinhança percebe que está sob estado de observação e, no decorrer das inspeções, os fraudadores do bairro ainda por serem investigados desfazem as irregularidades de suas instalações.

No momento da inspeção estes fraudadores são então observados como consumidores normais e assim classificados nos sistemas da empresa.

Em varias situações, mesmo sabendo que naquela unidade consumidora havia fraude, as equipes são obrigadas a classificá-las como normais por conta da ausência de flagrante.

A eficácia desse tipo de inspeção fica comprometida principalmente por que os inspetores são facilmente identificáveis, e a divulgação desta informação, logo chega a quem executa fraudes.

2.5.2 Inspeções de consumo zero

Trata-se de inspeções que são definidas a partir de um simples critério, mas que abrangem uma classe considerável.

Define-se um número mínimo de meses que a unidade consumidora realizou consumo zero. Após o critério estabelecido inicia-se o processo de inspeção. Enquanto o critério mantém um nível alto de sucesso na identificação de fraudes, o mesmo é mantido. Quando se observa a queda nos resultados satisfatórios, o período deve ser revisto.

Este tipo de inspeção tem se mostrado ineficaz após algum tempo, sendo sempre necessária uma revisão dos critérios. Na grande maioria, esses casos são imóveis fechados e situações que realmente justificam as medições registradas.

2.5.3 Inspeções de unidades consumidoras inativas

Nesta modalidade, são inspecionadas as unidades consumidoras desligadas. O motivo dessas inspeções é que existem muitas situações que ocorrem as auto-religações, ou seja, o consumidor efetua a sua religação sem autorização da concessionária.

Geralmente o objetivo desses casos é o acompanhamento de unidades consumidoras cujo fornecimento de energia foi suspenso por falta de pagamento.

Esta verificação é importante quando o consumidor por algum tempo não efetua a quitação do débito e nem solicita o pedido de religação para concessionária.

2.5.4 Inspeções a partir de denúncias

As inspeções podem ser executadas a partir de diversas fontes de denúncias. Esta modalidade ocorre pontualmente em instalações que são suspeita de ações irregulares executadas em suas unidades consumidoras.

Normalmente as denúncias de irregularidades à empresa, são por consumidores de outra unidade vizinha ou ainda por funcionários da própria concessionária. Esses grupos de pessoas quando percebem a irregularidade seja de forma direta ou anônima procuram oferecer informações sobre a localização e o tipo de irregularidade que está sendo praticada. Imediatamente a empresa envia uma equipe de inspeção ao local.

As informações de denuncias de irregularidades chegam a concessionária por diversos canais entre eles os principais são as agências de atendimento, call center ou 0800 e internet.

2.6 Procedimentos de inspeção

Na maioria das empresas de distribuição de energia existem normas que definem os procedimentos a serem adotados nas inspeções, bem como procedimentos para a autuação, e o registro cadastral dos resultados.

Pode-se descrever tais procedimentos da seguinte forma:

1) Identificação da região com maior perda

Em princípio efetua-se uma análise dos cálculos das perdas técnicas por alimentador de uma subestação, e posteriormente, estima-se a perda comercial por alimentador com base nas informações de faturamento. A seleção é através da escolha daqueles alimentadores que apresentem uma maior perda comercial.

Aquelas regiões identificadas onde há alto índice de perdas comerciais e que predominam as residências de periferia, ou seja, não existe indústria de grande porte ou centros comerciais, sugerem a utilização de inspeções de varredura. Em regiões industriais e comerciais, deve ser aplicada a inspeção originada a partir de análise mais apurada, uma vez que atinge consumidores de maior relevância.

2) Mobilização de força tarefa para realizar inspeções em campo

A empresa dispõe de um efetivo para realização de inspeções em campo, dependendo da estratégia definida são enviados vários técnicos especialistas para determinada ação. Podendo ser tanto para identificação de condições irregulares, quanto de autuação em flagrante e com poder para renegociação de débitos causados por fraudes identificadas.

Em algumas situações mais complexas acompanha também as equipes prontas para autuação em flagrante, oficial de justiça, e até mesmo força policial.

3) Cadastro de irregularidades

O cadastramento dos resultados das inspeções são registrados nos sistemas da empresa, e estes dados são utilizados para o desenvolvimento deste trabalho.

O sistema utilizado pela empresa permite inserir o registro das situações encontradas em campo, porém a condição para cadastro tem apenas sete opções de entradas, são elas:

- Irregularidade comercial: quando existe alguma alteração cadastral a ser realizada, por exemplo, ao ser encontrada em campo uma unidade comercial, mas que, para os registros da empresa apresenta a classificação residencial. Alguns casos de irregularidade comercial não provocam alteração em valores de tarifas, no entanto, existem muitos casos em que isto acontece.
- Falha na medição: quando existe um problema no valor lido, isto é o valor informado no sistema difere do valor lido. Existem casos de erro na inserção da informação da leitura nos sistemas da empresa. Estes podem ser causados pela leitura inadequada, ou

erro de digitação. Tais erros podem provocar um aumento exagerado no valor da conta ou um valor equivalente a um retrocesso no consumo (com uma conta de valor de consumo negativo).

- Irregularidade técnica: quando é encontrado um problema técnico nas instalações elétricas do consumidor, que, por averiguação pela empresa, não foi provocado artificialmente, isto é, representa um mau funcionamento ou um impedimento no funcionamento adequado do medidor, e, portanto impede a correta aferição do consumo daquela unidade. Estas irregularidades podem provocar prejuízo e perda de arrecadação. A unidade consumidora não é penalizada, pois não é detectada a intenção deliberada de adulteração dos equipamentos.
- Auto-religação: quando houve a suspensão do fornecimento de energia para o consumidor por falta de pagamento, e o mesmo se religa ou conecta novamente à rede elétrica, sem o conhecimento da empresa.
- Impedimento: por algum motivo alheio a vontade da concessionária existe impedimento de acesso ao medidor. Um exemplo seria quando um cão impede a verificação das instalações ou ainda quando o imóvel se encontra fechado.
- Normal: Neste caso a unidade consumidora é encontrada em seu estado normal. Vale ressalvar que esta situação contempla também aqueles consumidores que, durante a inspeção, não foi possível evidenciar o flagrante, porém foi possível verificar fortes evidências de uma fraude. Os atuais sistemas da empresa não são ainda capazes de registrar tal consumidor como suspeito. Ainda para esta situação, encontra-se as casas abandonadas, as casas vazias (casas de veraneio, por exemplo) e situações adversas que configuram um perfil de consumo bastante diferenciado, mas que, para efeito da massa de dados a ser verificada, possuem a mesma classificação.

Ainda existem aquelas situações impossíveis de serem registradas nos sistemas da empresa, que são: ligações completamente clandestinas (cujos usuários não possuem o menor registro na empresa), ligações alternativas que são realizadas diretamente na rede (sem a intermediação de um medidor) antes de o consumidor ser ligado pela primeira vez, e que, somente após sua descoberta podem ser classificadas como fraudes.

- Fraude: Quando é identificada uma violação, ou adulteração de equipamentos de medição com objetivos de redução ou eliminação do consumo da unidade consumidora. Nestes casos, o infrator é autuado em flagrante.

2.7 Comentários finais

Os sistemas de gerenciamento de dados de medição devem estar sempre atualizados e em condições de serem utilizados de maneira prática. Tais sistemas são instrumentos importantes e vitais para que possam ser evitadas as ocorrências de falhas que levem a perda de energia.

Algumas irregularidades são de difícil detecção e exigem das concessionárias, equipamentos especiais ou técnicas refinadas para facilitar a sua identificação. Quando o início da irregularidade ocasiona redução do consumo, pode-se a partir deste indício, iniciar a pesquisa da possibilidade de sua existência. No entanto, muitas vezes, o desvio é executado desde o início da ligação, tornando bem mais difícil a sua identificação.

Em geral, as concessionárias de energia elétrica encontram grandes dificuldades para operacionalizar um programa de inspeções com eficiência, em função da diversidade de dados, informações, análises, escassez de recursos humanos e principalmente a falta de ferramenta adequada na identificação dos clientes potencialmente fraudadores.

A recuperação de receita devida, referentes à defeitos na medição, está definida pela legislação e está limitada a retroação do ciclo de faturamento vigente, ou seja, em torno de 30 dias.

Desta forma, torna-se mais evidente a necessidade das concessionárias encontrarem métodos que agilizem a identificação de tais situações, pois a demora da localização desses casos com certeza vai acarretar prejuízos para a empresa.

Fazendo um resumo geral deste capítulo nota-se a grande importância de um aprofundamento nos problemas relacionados as perdas no setor elétrico nacional. Principalmente referente as perdas comerciais localizadas na distribuição como mostrado anteriormente, pois neste seguimento aparecem os índices mais elevados. Acredita-se que este trabalho poderá corroborar para minimização destas perdas, proporcionando uma condição de melhor utilização da energia elétrica, evitando principalmente investimentos desnecessários.

A utilização racional da energia elétrica é um dos fatores fundamentais para que o setor elétrico nacional mantenha a condição de crescimento do país.

CAPÍTULO III

PROCESSO DE DCBD (DESCOBRIMENTO DE CONHECIMENTO EM BANCO DE DADOS) E MINERAÇÃO DE DADOS

3.1 Introdução

O processo de Descobrimento de Conhecimento em Banco de Dados, mais conhecido pelo seu acrônimo DCBD (Knowledge Discovery in Database – KDD) é o nome do processo composto pelas etapas que produzem conhecimentos a partir de bancos de dados. Entre estas etapas, a mineração de dados é considerada uma das principais. Nela ocorre a identificação dos padrões, os quais podem representar o conhecimento. De modo geral, pode-se afirmar que mineração de dados é uma técnica de se extrair conhecimento de grandes bases de informação não refinadas, através de metodologias de reconhecimento e identificação de padrões. O entendimento destas regras e padrões gera o conhecimento, o qual é a base de um sistema de suporte a tomada de decisão.

Entre as aplicações desta técnica pode-se citar a determinação da estratégia de marketing baseada em padrões de consumo dos clientes, o reconhecimento de fraudes em áreas de telefonia e cartões de crédito, entre outras.

A mineração de dados foi definida por [Fayyad 1996] como sendo um *"processo não-trivial de identificação de padrões válidos, até então desconhecidos, potencialmente úteis e de possível entendimento em grandes bases de dado"*.

O termo *processo* caracteriza que existem diversas etapas a serem desenvolvidas as quais são relacionadas a seguir:

- Entendimento do domínio do problema;
- Transformação e preparação dos dados;
- Identificação e análise de padrões;
- Avaliação do conhecimento;
- Aplicação/utilização do conhecimento extraído.

No caso de um *processo não trivial* é utilizado para enfatizar que mineração de dados busca por padrões ou modelos não convencionais. Na Figura 3.1 são mostradas as etapas do DCBD conforme descrito por [Fayyad 1996].

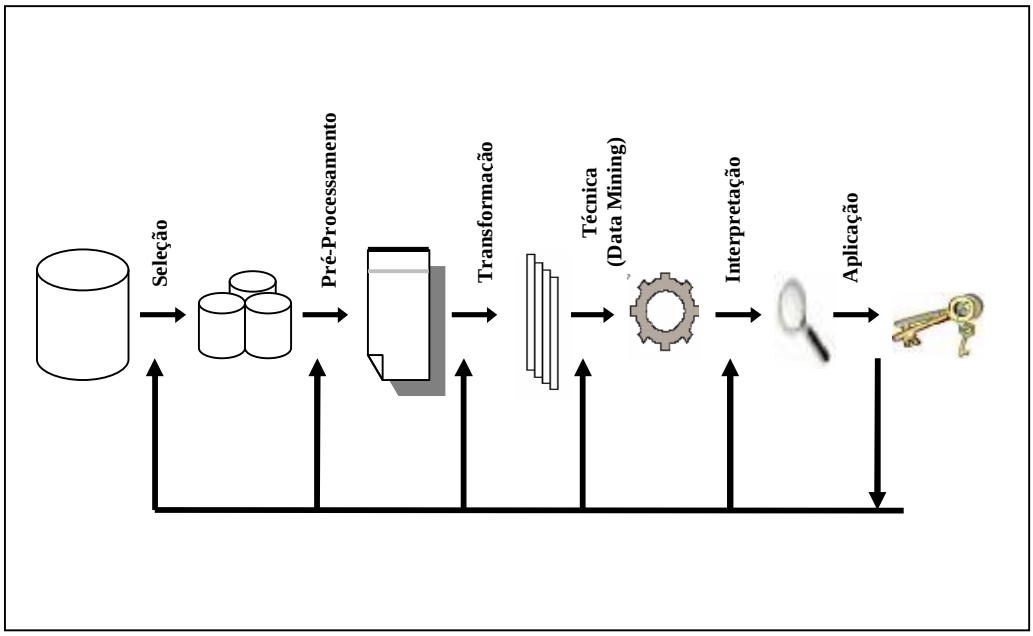

Figura 3.1 Diagrama de Blocos do Processo de DCBD

O processo DCBD é o resultado da fusão de áreas como banco de dados, aprendizagem de máquina (Inteligência Artificial) e estatística. Ele tem o objetivo de encontrar e interpretar padrões nos dados, de modo iterativo e interativo, através da repetição dos algoritmos e das

análises de seus resultados. O objetivo da Inteligência Artificial - IA dentro do processo de DCBD é o desenvolvimento de paradigmas ou algoritmos para que as máquinas realizem tarefas semelhantes às tarefas cognitivas humanas [Sage, 1990]. Ou seja, IA compreende métodos, ferramentas e sistemas para a modelagem de situações que normalmente requerem inteligência humana [Russel e Norvig, 1995].

Para dar maior capacidade a sistemas computacionais, duas estratégias podem ser utilizadas, introduzir no sistema o conhecimento humano ou ainda fazer o sistema extrair conhecimento implícito. Para executar estas ações, um sistema de IA deve ser capaz de: armazenar conhecimento, aplicar o conhecimento armazenado para resolver problemas e adquirir novo conhecimento através da experiência [Sage, 1990].

Neste capítulo são descritas brevemente as etapas do DCBD e a técnica Árvore de Decisão, a qual será utilizada na etapa de mineração de dados.

3.2 Descobrimento de Conhecimento em Banco de Dados

3.2.1 Seleção dos dados

A primeira etapa da descoberta de conhecimento, segundo [Fayyad, 1996], chamado de seleção de dados, requer o conhecimento do domínio do problema.

O domínio da aplicação e os objetivos do problema devem ser bem entendidos para que seja possível a seleção das bases de dados alvo, ou seja, aqueles que possivelmente contemplem informações que possam gerar o conhecimento requerido.

O objetivo é criar um conjunto de dados determinado a focar um sub-conjunto de variáveis ou dados de exemplo visando a utilização do usuário final.

Deve-se ainda ter pleno domínio dos dados que servirão de base para iniciar a descoberta do conhecimento. Este processo iterativo é sempre revisto ao longo de todo o descobrimento e é extremamente importante para o início dos trabalhos. A seleção adequada significa a utilização mais objetiva das informações disponíveis e a eliminação daquelas informações que, sabidamente, não irão ajudar na descoberta dos conhecimentos que são focados como objetivo do processo.

Contudo, a seleção também poderá significar uma redução no alcance dos resultados, caso não se faça uma análise e um planejamento adequado, podendo-se limitar a atuação das técnicas no seu campo de atuação.

Em determinados casos, isto poderá significar uma restrição precoce no processo de se descobrir novos conhecimentos. Mesmo com este risco, é necessária uma seleção prévia das informações que serão analisadas, pois atualmente a quantidade e a diversidade de informações disponíveis para estes tipos de processos tornam impossível a sua total utilização de maneira efetiva.

Como mencionado anteriormente, o conhecimento do domínio do problema é fundamental nesta etapa, e o envolvimento de especialistas no domínio é fortemente recomendado.

3.2.2 Pré-processamento de dados

O pré-processamento dos dados objetiva, de forma geral, a eliminação de registros duplicados, campos com falta de dados e campos com dados errados, etc. Durante o andamento de verificação do pré-processamento é definida a estratégia de suporte a campos que estejam com dados faltantes. Além disto, como os dados podem vir de várias tabelas, com

modelos distintos, é necessária uma integração dos dados, visando uma maior confiança nos valores dos mesmos.

Uma vez selecionadas as informações consideradas mais relevantes, é necessário extrair tais informações dos seus repositórios e tratá-las adequadamente de maneira a prepará-las para serem analisadas.

Os dados usados para a mineração são geralmente extraídos de bases que em sua grande maioria, não foram construídas para este objetivo. Desta forma, eles devem ser limpos e modelados para tornar possível a execução de um processo eficiente. Depois, devem ser transformados para um formato específico para o tipo de algoritmo que se pretende utilizar.

Alguns trabalhos já realizados nesta área de DCBD mostraram que mais de 70% do tempo gasto em um processo completo de descoberta vem sendo usado em pré-processamento e transformação de dados.

A definição da forma do pré-processamento tem influência direta no resultado final do processo [Engels, 1998]. A qualidade da preparação dos dados pode comprometer a etapa de mineração, direcionando-a na indicação de um algoritmo inadequado para solução ideal.

3.2.3 Transformação dos dados

O objetivo da transformação dos dados é a redução do número de variáveis a se considerar. Os dados pré-processados para serem utilizados com eficiência necessitam passar por um processo de redução, pois, a base de dados ainda contempla um volume considerável de informações. Isto pode ser feito através da redução na quantidade de atributos, redução do conjunto de dados usado para treinamento por amostragem (*sampling*) ou outras técnicas. Ao final do processo de redução da base, pode ser necessária ainda uma adaptação dos dados ao algoritmo utilizado na próxima fase.

Em aplicações reais, os dados podem ser incorretos, tornando as informações inconsistentes e incompletas. Estes erros podem ser gerados por instrumentos coletores de dados, falhas humanas nas entradas de dados, problemas de transmissão de dados, dentre outros. Por causa destes problemas, surgem campos com valores desconhecidos (*missing values*) ou com valores discrepantes (*outliers*).

Os campos com valores desconhecidos são aqueles que por alguma razão, não estão preenchidos para um determinado registro. Ignorar este problema pode gerar resultados errados ou conclusões incompletas, enquanto que substituir estes campos por valores pode introduzir inconsistências na base de dados. Desta forma, a substituição de valores desconhecidos deve ser feita de forma criteriosa para não alterar os padrões da base de dados.

A maneira mais simples para resolver este problema é descartar os registros que apresentem um ou mais campos com valores desconhecidos. Isto é possível se a massa de dados for extremamente confiável e abundante. Mesmo neste caso, corre-se o risco de eliminar registros importantes para a mineração.

Um método que tenta melhorar isto o faz através da eliminação de registros que contenham variações percentuais discrepantes na amostra. Apesar de ser uma medida simples, esta técnica vem sendo bastante utilizada e tem retornado resultados confiáveis. De qualquer forma, o risco de eliminação de registros importantes também existe neste método.

Uma outra forma de resolver tal problema é apenas ignorar os campos com valores desconhecidos, substituindo o valor inexistente por uma constante. Com a utilização dessa técnica surgem duas situações: na primeira, o algoritmo de mineração pode identificar estes valores especiais como outros quaisquer e tratar todos os registros com estes valores especiais como se fossem de um mesmo grupo, mesmo que eles pertençam a grupos completamente distintos. Isto torna claramente inadequada a mineração de dados; em uma segunda situação, e mais promissora, o algoritmo pode estar preparado para funcionar com estes valores especiais

e tratá-los de forma adequada, até mesmo adquirindo conhecimento a partir da inexistência de informações. Porém, um problema que surge é a possibilidade destes registros com campos desconhecidos serem decorrentes de um erro. Nesta última situação, mesmo os algoritmos preparados podem retornar conclusões erradas.

Um método mais aprimorado é o de inferir valores para estes campos. Uma estratégia poderia ser o uso da média dos valores daquele atributo para substituir os valores desconhecidos. Uma outra estratégia seria usar a média dos valores de exemplos pertencentes à mesma classe do registro analisado. Outra variante destas regras seria a de se obter o valor mais provável para o valor desconhecido [Han, 2001]. Pode-se também usar regras para inferir os valores de alguns campos a partir de outros. Mesmo nesse caso é possível ocorrer inferências incorretas.

Valores discrepantes podem ser descritos como informações que diferem em um grau tão elevado das normais que despertam suspeitas a respeito de sua correção. Além da identificação de *outliers*, faz-se necessário tomar providências para correção após a sua descoberta. O mais lógico seria eliminar os registros ou fazer uma substituição do valor discrepante. Porém, existe a possibilidade deste valor discrepante não ser uma informação incorreta e sim, um dado valioso. Na maioria dos problemas de detecção de fraudes, os valores discrepantes são os que buscamos. Logo, deveremos analisar criteriosamente a atitude a ser tomada depois da identificação de um *outlier*.

Uma estratégia para identificação é a técnica de agrupamento (*clustering*), de forma que os registros são agrupados de acordo com informações de alguns atributos. Os raros registros que não pertencerem a nenhum dos grupos classificados seriam os *outliers*.

A figura 3.2 ilustra o método. Percebe-se que alguns valores não pertencem a nenhum grupo.

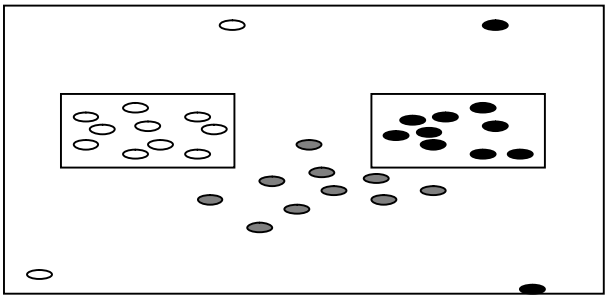

Figura 3.2. Agrupamento para identificação de *outliers*.

Pode-se ainda utilizar na identificação de *outliers*, a técnica de regressão linear, a qual visa aproximar os dados de um campo a partir dos valores de outro(s) campo(s) por meio de uma função. Desta forma, os valores *outliers* seriam os que não pertencessem à função.

A figura 3.3 refere-se a uma função linear. Nota-se que alguns pontos se encontram localizados distantes do eixo da reta, caracterizando valores discrepantes.

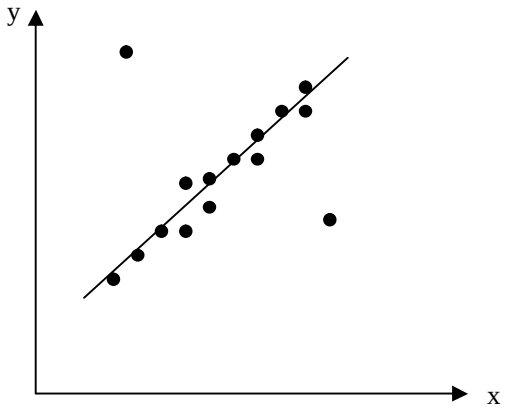

Figura 3.3 - Regressão linear para identificação de *outliers*.

O método mais simples na identificação dos *outliers*, é através da verificação de percentuais de variação da média dos valores de um atributo a ser definido pela equipe de análise do processo.

Ex: A média dos valores de um atributo é 100. Suponha que a porcentagem escolhida seja 30%. Então, valores acima de 130 ou abaixo de 70 são *outliers*. A Figura 3.4 mostra os dados de um exemplo. Para o exemplo mostrado, os valores 59, 140 e 155 são considerados *outliers*.

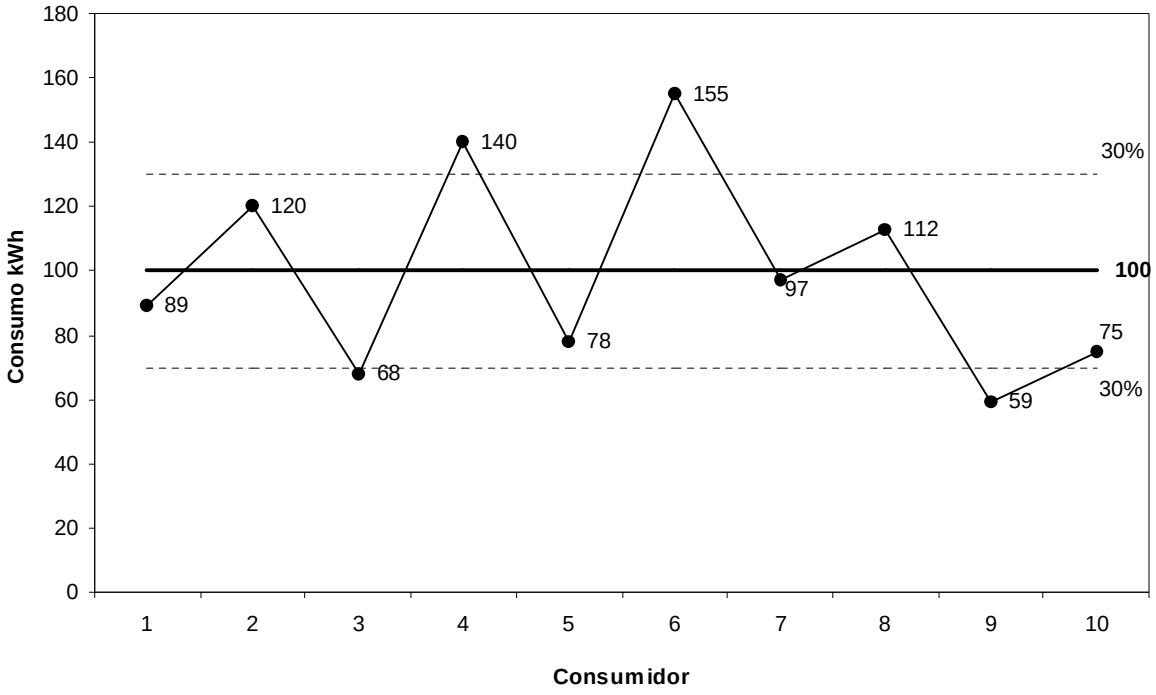

Figura 3.4 – Percentual de variação para identificação de *outliers*.

Após a conclusão e a definição do método a ser aplicado, inicia-se a integração dos dados em suas bases. Este processo consiste na agregação de informações de diferentes fontes de maneira coerente para que sejam examinados pelas várias técnicas de mineração. A coesão no momento da agregação dos dados pode mostrar-se complexa dependendo dos casos.

Existem bases cujos registros são identificados por diferentes chaves e por diferentes objetivos. No entanto, existem bases de dados de interesse que podem estar com informações agrupadas por categoria, classe ou por percentual de representatividade. A integração deste tipo de informação requer uma análise criteriosa.

Outra questão associada à integração é a redundância de valores, ou seja, informações que podem ser derivadas de outras informações. Por exemplo, eliminar atributos que coincidem com combinações específicas de outros atributos. Esta redundância pode ser eliminada através da análise de correlação entre as informações [Han, 2001].

A parte final deste processo corresponde à transformação dos dados, uma vez que os mesmos foram corrigidos e integrados.

A transformação de dados pode envolver, limpeza, generalização, normalização, discretização, transformações específicas ou construção de atributos (através da construção de novos valores derivados dos valores básicos para o auxílio da mineração) [Han 2001].

A normalização é uma importante técnica para a análise de dados numéricos de característica contínua. Nela são eliminados os efeitos de valores muito diferentes em escala, mas que potencialmente dizem respeito ao comportamento básico da característica que está sendo analisada.

Um algoritmo pode precisar de dados num formato específico. Além disto, algoritmos diferentes necessitam de transformações específicas para poderem trabalhar corretamente.

A discretização de dados contínuos é usada para reduzir o número de possíveis valores para um determinado atributo, através da divisão da faixa destes valores em intervalos. Para cada intervalo é escolhido um valor representante que substituirá o valor que realmente consta na base de dados.

Um método de discretização é a hierarquia de conceitos. Neste método, há uma substituição de valores. É possível substituir o valor numérico do consumo por valores que descrevam a sua intensidade.

Na tabela 3.1 é mostrado um exemplo de discretização de valores de consumo de energia elétrica.

Tabela 3.1- Discretização de consumo de energia elétrica.

Categoria	Faixa de Consumo
Consumo muito baixo	$0 \leq 30$ kWh
Consumo baixo	$> 30 \leq 100$ kWh
Consumo médio	$> 100 \leq 200$ kWh
Consumo alto	$> 200 \leq 300$ kWh

Na seqüência é feita a redução final do volume de dados, pois a base ainda pode estar muito grande, o que pode comprometer a eficiência do sistema. Podem-se destacar dois tipos de redução.

- Redução vertical (redução das dimensões dos dados);
- Redução horizontal (redução do número de exemplos).

A redução vertical consiste em diminuir o número de atributos usados na mineração. Desta forma, objetiva-se encontrar o menor número de atributos que tenha performance equivalente a de todos eles. Se forem testados todos os atributos e suas combinações em busca de um número ótimo, o problema tem complexidade exponencial. Assim, são utilizados métodos específicos para realizar este processo.

Uma estratégia utilizada para redução são as árvores de decisão. Tenta-se dividir a base pela classificação de um atributo. O atributo que dividir melhor, gerando a menor entropia, é usado para fazer esta primeira divisão. Depois desta fase, tenta-se dividir as duas bases seguintes por outros atributos quaisquer. Este processo é repetido até que a classificação chegue a um estágio suficiente.

Outro método de redução vertical é através de regras de associação, o qual elimina as redundâncias entre atributos. Baseia-se no fato de que, se for possível inferir o valor de um atributo X1 através dos valores de um ou mais atributos da tabela (X2, X3, ..., Xn), então o valor X1 é redundante e pode ser descartado.

A redução horizontal consiste em diminuir o número de registros utilizados no processo de mineração. Isto porque a base de dados neste momento pode ainda estar muito grande, tornando a aplicação da técnica de mineração de dados ineficiente. Esta redução deve ser feita de modo criterioso para que o conjunto escolhido seja representativo, ou seja, equivalente com a situação da base de dados completa.

A amostragem estratificada dos dados é uma técnica de redução horizontal dos dados. Consiste em usar um método de agrupamento para os registros e, posteriormente, escolher randomicamente um número de registros de cada grupo, de forma que cada um mantenha sua porcentagem de elementos no conjunto de treinamento igual à de elementos na base total.

Mostra-se abaixo a figura 3.5 que retrata a técnica horizontal de redução de dimensões de dados:

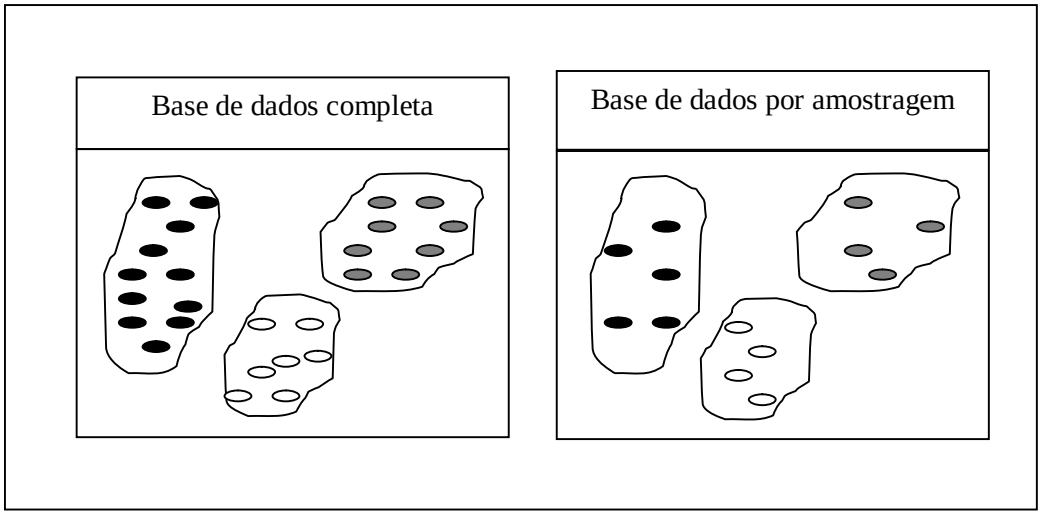

Figura 3.5 Redução de dados por amostragem estratificada

Após a aplicação das técnicas pertinentes e a identificação dos dados a serem trabalhados finda o pré-processamento e as transformações necessárias, o próximo passo é a aplicação da

técnica de mineração de dados, na base de dados selecionada após as transformações necessárias.

3.2.4 Mineração de dados

Esta etapa é uma das principais fases do processo de DCBD, pois aqui se define o algoritmo utilizado para fazer a identificação dos padrões nos dados. Esta etapa tem o objetivo de “minerar” os dados, procurando por padrões de interesse em uma forma de representação particular ou uma representação qualquer. É essencial que seja escolhida uma técnica que se adapte às características do problema em questão, mesmo que este processo de escolha demande um longo tempo em procedimentos de testes. Uma outra possibilidade é a integração de duas ou mais técnicas de forma a aumentar a confiabilidade do método.

O processo de aplicação de uma técnica de mineração de dados é conhecido como aprendizagem. Os tipos de aprendizagem podem ser supervisionadas ou não supervisionadas. Na supervisionada, é apresentado ao algoritmo exemplos contendo tanto os atributos de entrada quanto a saída, isto é, a categoria a qual o exemplo pertence. Assim, após o treinamento, o sistema tem a capacidade de classificar um novo exemplo, o qual ainda não o foi apresentado. Na aprendizagem não supervisionada, somente é apresentado ao sistema exemplos contendo atributos de entrada. Assim, o mesmo apenas classifica os mesmos em grupos, os quais possuem similaridades do ponto de vista de padrões de entrada, ficando a critério do especialista identificar quais as características de saída que tais grupos podem ter.

No caso de aprendizagem supervisionada, o classificador não é capaz de identificar novas classes, mas apenas se um novo caso pertence a uma das classes conhecidas. No aprendizado não supervisionado, o classificador seleciona aquelas ocorrências que mais se assemelham e

as agrupa na mesma classe. Neste caso, o classificador é capaz de criar um novo grupo de indivíduos que, após uma análise poderá se configurar em uma nova classe.

No aprendizado supervisionado, os classificadores podem apresentar o que chamamos de especialização ou sobre-ajuste (Over-fiting). Esta característica pode ocorrer quando o classificador se adapta aos dados de treinamento e sua capacidade de generalização fica limitada, isto é, ele classifica a maioria dos novos exemplos de forma errada.

A literatura tem relatado trabalhos utilizando tanto técnicas de computação convencional, como técnicas de computação flexível, mais conhecidas como técnicas Inteligência Artificial, para uma diversidade de áreas e muitos seguimentos de atuação. No entanto, para identificação de fraudes, é conhecida somente aplicação de tais ferramentas em algumas áreas específicas como: cartões de crédito, telefonia celulares e seguros etc. A seguir são brevemente apresentadas três técnicas de inteligência artificial, Redes Neurais Artificiais, Lógica Nebulosa e Conjuntos Imprecisos (Rough Sets), as quais têm sido muito utilizadas como técnicas de mineração de dados no problema de identificação de fraude. Na seção seguinte, será então apresentada a técnica Árvore de Decisão, a qual foi a técnica utilizada neste trabalho.

Redes Neurais Artificiais ou simplesmente Redes Neurais [Haykin, 2001] é um método eficiente para aproximação de funções reais, discretas ou para solução de problemas de agrupamento. Seu funcionamento é baseado em células independentes de processamento (neurônios), que podem estar conectadas aos dados de entrada e a outras células [Braga, 1998]. A cada uma destas conexões, é atribuído um peso que define qual será o comportamento da rede para determinado padrão de entrada e o método de defini-los consiste no algoritmo de aprendizado da rede. Inicialmente idealizadas para reproduzir e se beneficiar do comportamento conhecido dos neurônios biológicos, as redes neurais baseiam-se no

processamento de uma tarefa complexa. Os resultados das operações influenciam as próximas unidades de processamento [Wermter, 2000].

As redes neurais podem ser classificadas quanto a sua arquitetura básica ou ainda quanto a sua arquitetura de uso. Na arquitetura básica as redes neurais podem ser classificadas como redes de alimentação direta, mais conhecidas como *feedforwad networks,* onde a saída de um neurônio só influencia as camadas posteriores e redes recorrente, também conhecidas como *feedback networks,* onde a saída de um neurônio pode influenciar as camadas anteriores. Quanto à arquitetura de uso existem alguns tipos de redes neurais associadas à sua categoria de uso e de possíveis aplicações, como exemplos existem: redes perceptrons, redes perceptrons de multi-camadas, redes lineares, redes de base radiais, redes de Elmann, entre outras..

Redes neurais têm sido aplicadas para realizar tarefas de previsão, classificação, associação, conceituação e filtragem de dados [Anderson, 1992]. Por serem baseadas no conceito de neurônios, as várias arquiteturas de redes neurais possuem várias similaridades. A maioria das diferenças reside nas várias regras de aprendizagem e como elas podem modificar a topologia típica da rede.

As Redes Neurais podem ter dois tipos de aprendizado, o supervisionado e o não supervisionado. No aprendizado supervisionado, a rede é treinada para determinar qual a saída, para determinado conjunto de atributos como entrada. O grau de aprendizado é mensurado de acordo com o índice de acertos da rede através dos resultados dos testes.

O algoritmo de aprendizagem consiste no método a ser utilizado para determinação dos diversos pesos atribuídos à entrada de dados na rede. Existem vários algoritmos de aprendizado para redes neurais. Em se tratando de aprendizado supervisionado, os mais tradicionais se propõem a minimizar o erro médio quadrado.

Para o aprendizado não supervisionado, é fornecido à rede um conjunto de atributos, com base nos valores destes, ela deve ser capaz de construir diferentes classes. Ao ser apresentada alguma excitação na entrada da rede, esta deve ser capaz de associá-la a alguma classe, de acordo com sua semelhança com os exemplos usados durante a fase de treinamento.

Do ponto de vista de treinamento, a rede pode ser treinada de forma estática, de tal forma que use a estrutura da rede de forma fixa, ou ainda de forma dinâmica em que o número de elementos da rede pode variar durante o processo.

Conforme comentado anteriormente, a junção de algumas técnicas podem ser utilizados para se obter melhores resultados.

Por sua vez, Lógica Nebulosa oferece um ambiente muito poderoso para aproximar o raciocínio, num esforço para modelar o pensamento humano. Sistemas nebulosos adquirem o conhecimento de especialistas e o codificam em termos e regras se/então. Estes sistemas empregam tais regras num método de interpolação, simulando o raciocínio, para responder a novas questões. Em contraste, as redes neurais oferecem uma arquitetura altamente estruturada, com capacidade de aprendizado e generalização. Uma junção entre estas técnicas dá origem a uma poderosa técnica híbrida, chamada neuro-fuzzy. Ao se projetar um sistema neuro-fuzzy, agregam-se as características de transparência de raciocínio da lógica nebulosa à capacidade de aprendizado e generalização das redes neurais.

Os conceitos dos Conjuntos Imprecisos, ou Rough Sets, são de fácil compreensão prática e aplicação. A despeito de sua utilização direta como técnica de Inteligência Artificial, Rough Sets possui uma fundamentação teórica bem consolidada. A Abordagem desta técnica será realizada de forma sucinta com base em conceitos genéricos de "*Rough Sets: Theoretical Aspects of Reasoning about Data*" (Pawlak, 1991).

A teoria de Rough Sets está calcada em dois elementos: **objetos** e o **conhecimento** acerca dos mesmos. Os objetos são instâncias (ou exemplos, registros) de qualquer elemento real ou

imaginário. Ou seja, objetos podem representar seres humanos, objetos concretos, medidas de algum fenômeno amostradas no tempo, ou qualquer outra entidade que se possa imaginar.

A um conjunto de objetos, doravante denominado *universo de discurso* (ou simplesmente *universo*), é possível aplicar uma ou mais características, definindo uma classificação de objetos. A estas características dá-se o nome de conhecimento. Portanto, dado um universo e o conhecimento disponível sobre o mesmo, é possível realizar classificações ou partições neste universo. Por exemplo, dado um conjunto de pessoas (universo) e seus respectivos sexos (conhecimento), é possível encontrar uma partição (classificação) deste conjunto: o subconjunto de homens e o subconjunto de mulheres.

Um *conceito* pode ser entendido como uma classificação, uma partição de objetos do universo, porém, nem sempre, um *conceito* é *definível* para a base de conhecimento considerada. Em outras palavras, muitas vezes não é possível definir uma classificação exata dos objetos a partir das relações de equivalência encontradas em uma base de conhecimento. Uma alternativa a este problema, o qual ficará mais evidente à seguir, é proposta por Rough Sets: encontrar *conceitos* (ou classificações) aproximados em uma base de conhecimento.

O conhecimento existente sobre um universo de objetos pode ser insuficiente ou mesmo excessivo. Quando insuficiente, leva a formação de *conceitos* indefiníveis e baixas medidas de precisão. Já quando é exagerado, é conveniente identificar aqueles conhecimentos que podem ser desconsiderados sem promover mudanças nos *conceitos*. Esta *redução de conhecimento* torna-se mais relevante quando o tamanho da base de conhecimento é limitado como uma forma de classificação (ou partição) através de *conceitos*. Para uma melhor manipulação dos objetos e do conhecimento, utiliza-se um *Sistema de Representação do Conhecimento*, normalmente chamado de *Sistema de Informação (SI)*. Um SI é uma representação sintática do conhecimento sobre um conjunto de objetos e consiste de uma tabela de dados, onde as colunas são nomeadas como *atributos* e as linhas como *objetos*. Cada

coluna representa uma relação de equivalência e cada linha armazena as classes de equivalência na qual o objeto desta linha está inserido. Um SI normalmente é acrescido de pelo menos um atributo, o qual realiza uma classificação sobre os objetos, levando à tomada de decisões. Os SI incrementados por atributos de decisão são chamados *Tabelas de Decisão*. Tais tabelas permitem que objetos dêem origem à regras de decisão, possibilitando a aplicação do conhecimento dos objetos existentes na classificação de novos objetos.

Tabelas de Decisão são utilizadas em várias aplicações, envolvendo problemas de classificação, reconhecimento de padrão, sistemas especialistas, etc. Normalmente, estas tabelas são submetidas a processos de redução ou simplificação, dentre eles:

1. Redução de atributos condicionais: obtida através do cômputo do reduto, permitindo que atributos dispensáveis sejam removidos;

2. Eliminação de regras duplicadas: após selecionar os atributos condicionais de um reduto, linhas ou regras de decisão podem tornar-se idênticas, sendo suficiente manter apenas uma regra representante;

3. Redução de valores de atributos condicionais: é possível que uma regra seja simplificada através da eliminação de restrições condicionais, visto que eventualmente nem todas condições de umas regras necessitam ser testadas para realizar-se uma decisão.

Para realizar-se uma redução em Tabelas de Decisão por eliminação de valores de atributos condicionais, utiliza-se um método semelhante àquele empregado na identificação de redutos em SI.

3.2.5 Interpretação do conhecimento descoberto

Com o término da etapa de mineração de dados, pode-se analisar os resultados alcançados. O conhecimento adquirido nos padrões obtidos é interpretado e analisado e

testado para avaliação de sua performance. É verificado se o resultado é satisfatório ou se há necessidade de retornar as etapas anteriores para reformulá-las.

A presença dos especialistas para avaliação da interpretação dos resultados conquistados é fator preponderante para que os resultados sejam validados como nova descoberta. Esta interpretação pode ser feita de várias formas, desde a simples revisão dos resultados até a sua comprovação em campo. Os resultados podem ser fornecidos de forma probabilística, na forma simbólica, ou simplesmente classificatória, isto é, pertencente ou não a uma determinada classe.

3.2.6 Consolidação do conhecimento descoberto

Nesta etapa do DCBD, é consolidado o conhecimento obtido incorporando-o ao processo ao sistema de suporte a tomada de decisão, também conhecido como SSTD. Neste ponto, pode-se utilizar o conhecimento obtido pelo método nas tomadas de decisões gerenciais.

3.3 Árvore de Decisão

Árvore de Decisão é uma técnica que tem sido intensivamente explorada para problemas de classificação. Uma das principais características desta técnica é a sua forma simples de representar o conhecimento e sua facilidade de implementação, baseando-se em treinamento por casos.

A árvore de decisão dá uma visão gráfica da tomada de decisão por regras se/então. Nesta técnica, um problema complexo é decomposto em sub-problemas mais simples para fazer a classificação. Para tal, é utilizado um algoritmo que subdivide o conjunto de treinamento

repetidas vezes até alcançar uma partição que represente casos pertencentes à mesma classe, ou até que um pré-definido critério de parada seja alcançado.

Sua estrutura é um diagrama de fluxo em formato de árvore, em que cada nó interno indica um teste em um atributo, cada ramificação representa um resultado de um teste e os nós folha representam classes ou distribuições de classes.

Na figura 3.6 é mostrado um modelo genérico do formato de uma árvore de decisão.

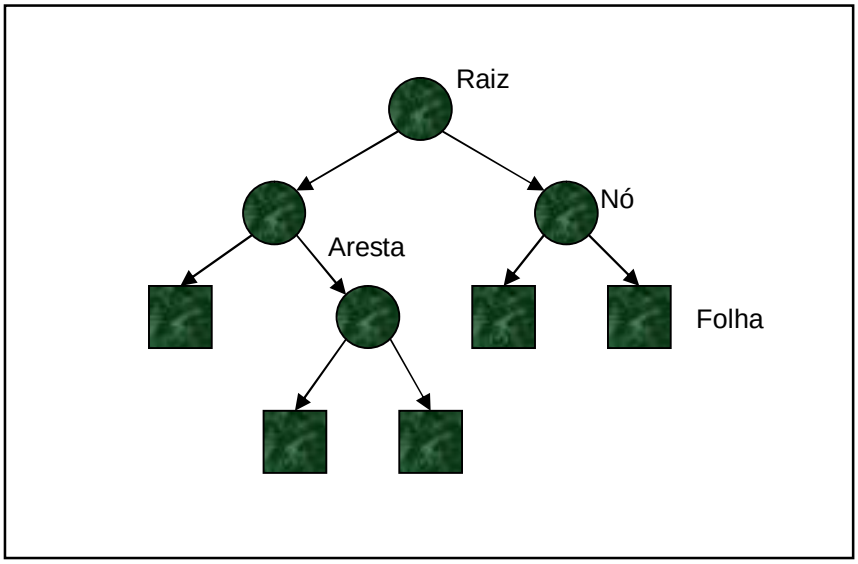

Figura 3.6 – Modelo de árvore de decisão

Na tabela 3.2 é apresenta-se as componentes de uma árvore de decisão com suas respectivas especificações.

Tabela 3.2 – Componentes da árvore de decisão

Raiz	Primeiro nó da árvore.
Nó	Representa uma pergunta, ou seja, o teste de um atributo (característica);
Folha	Classificação final para o exemplo (sim/não, 0/1);
Aresta	Ligação (ou caminho) entre nós ou entre um nó e uma folha.

A figura 3.7 mostra outra representação de uma árvore de decisão. Na figura, X1 é um nó raiz, enquanto Classe A e Classe B são nós folhas. Na figura pode-se ainda perceber o seu funcionamento, onde em cada nó que não seja um nó folha, um teste é feito aos exemplos (?), e ele é dividido de acordo com as respostas sim (S) ou não (N). Os nós folhas representam a classe a qual pertence aquele exemplo.

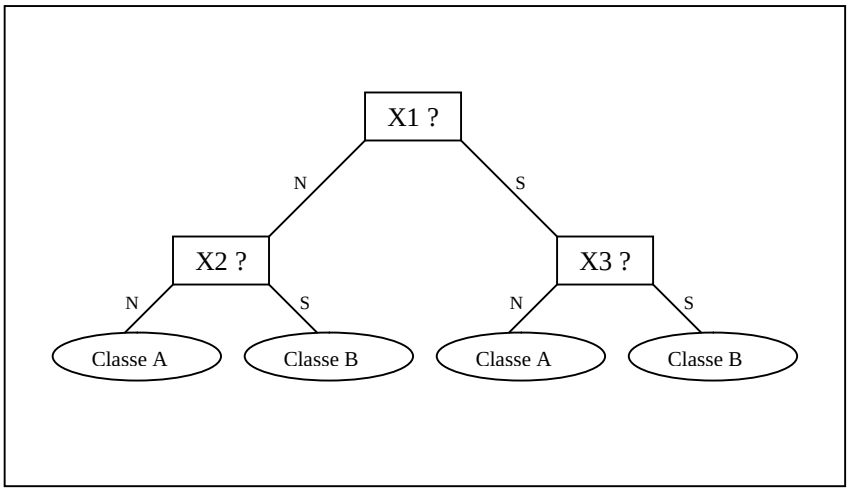

Figura 3.7- Redução de dados com uso de árvore de decisão

Algoritmos do tipo TDIDT (*Top Down Induction of Decision Trees*) [Quinlan, 1985] geram modelos no formato de árvores de decisão. Métodos de indução de árvores de decisão usam um algoritmo guloso que subdivide recursivamente o conjunto de treinamento até encontrar uma partição que represente os casos que pertencem a uma mesma classe. A cada partição, é realizado um teste estatístico para decidir qual atributo vai ser usado por cada subdivisão.

Sendo o conjunto C - Clientes dados de treinamento e {CR-Residencial, CC- Comercial, CI- Industrial,..., Cn} as classes, temos o seguinte método para a construção de uma árvore de decisão:

a) Se C contém exemplos que pertencem a várias classes a idéia é refinar C em subconjuntos de exemplos que são (ou aparentam ser) conjuntos de exemplos pertencentes a uma única classe.

b) Um teste é escolhido baseado em um atributo com os resultados mutuamente exclusivos. Cada possível resultado do teste gera um subconjunto de C.

Os passos "a", "b" são aplicados recursivamente para cada subconjunto de C. Em cada nó, as arestas levam para as sub-árvores construídas a partir do subconjunto de C.

O critério utilizado para escolher o atributo que particiona o conjunto de exemplos em cada iteração depende do indutor. Uma abordagem é a escolha aleatória do atributo. Devido à possibilidade de selecionar atributos com baixo poder preditivo, esta escolha pode levar à indução de árvores com baixo poder de predição e generalização. Uma abordagem mais adequada é utilizar alguma medida de avaliação dos atributos para selecionar aquele que tenha maior probabilidade de melhorar o desempenho de predição da árvore. Esta abordagem tende a gerar árvores menores com maior poder de predição. Os conceitos de entropia e ganho de informação são utilizados pelos algoritmos ID3 [Quinlan, 1986] e C4.5 [Quinlan, 1993] para avaliar se um atributo aumentará o desempenho da árvore.

Um dos algoritmos de aprendizagem para treinamento de árvore de decisão mais popular é o C4.5. Ele constrói uma árvore de decisão a partir dos dados de treinamento de maneira direta, isto é, do nó raiz para o nó folha. A seleção do atributo a ser testado em um nó é baseada na medida de razão de ganho. A razão de ganho mede a quantidade de informação obtida com aquele atributo na classificação, durante o treinamento. O atributo com maior razão de ganho é selecionado para ser o nó raiz, e os exemplos são particionados, ou separados, de acordo com seus valores daquele atributo. Para cada partição de exemplos, o próximo atributo com maior razão de ganho é selecionado para ser o nó da árvore. Cada partição é então dividida novamente em menores partições de acordo com os valores dos

atributos selecionados. Este processo continua até que as partições finais tenham exemplos pertencentes a uma mesma classe. Os últimos nós são chamados nós folhas, e representam partições para as classes. A razão de ganho é definida conforme a equação 3.1. [T. Michell, 1997]:

$$RazãodeGanho(S,A) \equiv \frac{Ganho(S,A)}{DivisãodeInformação(S,A)} \tag{3.1}$$

Onde

$$DivisãodeInformação(S,A) \equiv -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \tag{3.2}$$

E

$$Ganho(S,A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v) \tag{3.3}$$

E

$$Entropia(S) \equiv \sum_{i=1}^{c} - p_i \log_2 p_i \tag{3.4}$$

Onde p_i é a proporção dos exemplos pertencentes à classe i. A Entropia, conforme definida na equação (3.4), mede a uniformidade da distribuição das classes dos exemplos de treinamento. Basicamente ela é o número de bits necessários para codificar a classificação dos exemplos de treinamento, e seu valor pode variar de 0 até 1. Para uma classificação binária, uma entropia de 0 indica que todos os exemplos pertencem a uma mesma classe. Por outro lado, uma entropia de 1 indica que metade de todos os exemplos (½) pertence para uma classe, enquanto a outra metade (½) pertence para a outra classe. O ganho de informação, definido na Equação (3.3), mede a expectativa da redução na entropia causada pelo particionamento dos exemplos de treinamento usando os valores de um atributo. O ganho de informação é alto se o decréscimo da entropia é alto. Isto implica que a entropia de cada S_v (cada subconjunto dos exemplos particionados usando o atributo) é relativamente pequena e

uma alta proporção de exemplos de cada S_v pertencem para apenas uma classe. Portanto, alto ganho de informação significa que muitos exemplos podem ser selecionados para uma classe correta usando os valores dos atributos. A divisão de informação, definida na equação (3.2), mede a distribuição dos valores dos atributos dos exemplos, usando um determinado atributo. Para um atributo com dois valores, a divisão de informação de 0 indica que todos os exemplos têm o mesmo valor de atributo. Uma divisão de informação de 1 indica que metade (½) dos exemplos tem um valor, e a outra metade (½) deles tem o outro valor. Então, a divisão de informação dá o conhecimento da uniformidade da divisão dos exemplos de treinamento para um dado atributo. Atributos significativos são aqueles com número pequeno de valores e pequenos valores de entropias para os S_v*'s*. Entretanto, se atributos com um número de valores muito altos e atributos com um número de valores baixos coexistirem no conjunto de dados, o ganho de informação do primeiro será maior que o do segundo. A divisão de informação é incorporada na razão de ganho, conforme mostra a equação (3.1) para penalizar os atributos com muito alto número de valores, o que assegura que os atributos significativos são selecionados durante o treinamento da árvore de decisão.

Após a construção da árvore de decisão, é possível que o classificador induzido seja muito específico para o conjunto de treinamento. Isto torna a precisão do classificador alta para o conjunto de treinamento, mas piora o desempenho em um conjunto de teste. Neste caso, diz-se que o classificador superajustou os dados de treinamento, ou seja, ocorreu uma especialização (*overfitting)*. Entre algumas ações para resolver este problema pode-se aplicar o procedimento chamado "poda" (*pruning*), o qual desempenha um papel importante, efetuando a redução da mesma e produzindo árvores menores com potencial de precisão mais eficaz para novos casos considerados. Este procedimento consiste da remoção de alguns nós internos, reduzindo a complexidade da árvore, mas melhorando o seu desempenho e capacidade de generalização.

O processo de pré-poda que é efetivada durante a indução da árvore de decisão. Define um critério de parada, assumindo que um nó pode tornar-se folha sempre que certa porcentagem dos exemplos pertencerem a uma mesma classe.

Outra modalidade é a pós-poda, que é aplicada após a indução da árvore de decisão. Um conjunto de exemplos de teste é fornecido à árvore para ser classificado. Se a taxa de erro for menor pela substituição de uma sub-árvore (formada por um nó não terminal ligado diretamente a nós folha) por um nó folha, então é realizada a substituição, ou seja, a árvore é podada. Este processo é realizado até que nenhuma melhora possa ser feita nas sub-árvores.

Uma árvore de decisão pode ser utilizada para classificar novos exemplos, iniciando-se pela raiz da árvore e caminhando através de cada nó de decisão até que uma folha seja encontrada, e então a classe do novo exemplo é dada pela classe daquela folha. Cabe ressaltar que existe somente um caminho que pode ser percorrido por um exemplo, já que a árvore divide o espaço de descrição do problema em regiões disjuntas.

O principal problema com árvores de decisão é que elas necessitam de grandes quantidades de dados para descobrir estruturas complexas. Porém, elas podem ser construídas consideravelmente mais rápido que outros algoritmos de aprendizagem de máquina produzindo resultados com precisão similar [Sousa, 1998]. Além disto, as árvores de decisão são de mais fácil entendimento pelo ser humano se comparadas a algumas das outras técnicas.

3.4 Comentários finais

Na literatura encontra-se um grande número de técnicas para a identificação de fraudes. Dentre as mais utilizadas, estatística, redes neurais, árvores de decisão e combinações destas, entre outras, se mostraram muito eficazes.

A detecção de fraudes é um grande desafio para empresas e especialistas, principalmente em função de sua natureza mutável ao longo dos tempos. Assim que algum tipo de fraude passe a ser conhecida, naturalmente estratégias para o seu combate são rapidamente construídas. No entanto novos tipos de fraudes são então criadas, de maneira que sempre existirão casos a serem descobertos.

Outro fator que dificulta a identificação do perfil ou das características de um fraudador é o fato de existir uma enorme quantidade de casos normais dentro dos dados analisados, ou seja, a relação número de não fraudadores por fraudadores é muito grande, sendo que os fraudadores aparecem apenas como ruído nos dados, portanto a sua descoberta é praticamente impossível se realizada de forma artesanal.

Para o aprendizado através de técnicas de mineração para a detecção de fraudes, é necessário que o algoritmo de aprendizado seja treinado utilizando uma base de exemplos. Tais exemplos são de suma importância para o bom desempenho do algoritmo. Isto significa que, bases de dados de treinamento geralmente podem trazer resultados iniciais extremamente frustrantes, já que todos os exemplos, ou a maioria deles, serão então classificados como normal. Porém, superado a questão da classificação inadequada, o algoritmo escolhido será capaz de determinar apenas aqueles fraudadores cujas fraudes são conhecidas. Obviamente, não haverá como descobrir outros tipos de fraude, uma vez que o algoritmo não foi treinado para tais.

Desta forma concluí-se que são vários os fatores que dificultam no processo de identificação dos fraudadores. Podemos citar como exemplos algumas dessas dificuldades: quantidade de dados/informações, alterações de padrões de comportamento, similaridade dos dados entre fraudadores e não fraudadores, alto custo para identificação de fraudes, questões sócio-econômicas etc.

CAPÍTULO IV

DESENVOLVIMENTO DO SISTEMA DE IDENTIFICAÇÃO DE FRAUDES E ERROS DE MEDIÇÃO

4.1 Introdução

Observou-se no capítulo anterior que a identificação de perdas comerciais não é um processo trivial, automatizá-lo torna-se ainda mais complexo. Neste trabalho foi preciso adequar-se dentro de algumas condições, que de certa forma comprometeram o desenvolvimento da pesquisa, são elas: a dificuldade no acesso e a confidencialidade dos dados. Nenhuma empresa quer tornar públicas suas vulnerabilidades.

Estes dois aspectos não foram citados em outros trabalhos cuja bibliografia foram estudadas para o desenvolvimento deste. Encontramos apenas referência à dificuldade em se trocar experiências a respeito da busca por fraudes [Kou, 2004].

Assim o trabalho se deparou com a impossibilidade de se ter acesso a todas as informações disponíveis. Esta característica tornou a pesquisa ainda mais desafiadora. O Data Warehouse (DWH), software utilizada pela concessionária cuja função é o armazenamento e extração de dados possui cerca de 3000 tabelas e 15.000 colunas diferentes. Certamente nem todas seriam relevantes ao problema, porém, ficou-se limitado a trabalhar inicialmente com um número bastante reduzido destas.

A outra dificuldade se mostrou em relação a confidencialidade das informações. Não é possível conhecer as informações diretamente por estar se tratando de informações de consumidores. Portanto, o trabalho foi todo desenvolvido em cima de códigos de

identificação. Foram então utilizados identificadores, para os resultados de inspeção e tipos de atividades dos consumidores. Apesar das dificuldades, este trabalho teve por objetivo melhorar o processo de seleção de consumidores de baixa tensão (BT) a serem inspecionados por uma concessionária de distribuição de energia.

Este trabalho faz parte de um conjunto de ações no combate a perdas comerciais. A intenção é fornecer aos usuários uma lista de consumidores a serem visitados em campo, e que, em sua maioria de consumidores selecionados seja constituída por verdadeiros fraudadores.

Neste capítulo será descrita a aplicação do processo de DCBD abordado no capítulo 3, utilizando dados de consumidores de uma concessionária de energia elétrica.

4.2 Processo de seleção de dados

O banco de dados de uma empresa de distribuição de energia elétrica contém inúmeras informações, desde o histórico de consumo dos clientes a dados técnicos dos dispositivos de distribuição. Enfim, uma grande quantidade de dados que requer segurança e confiabilidade, tanto no acesso como no armazenamento e recuperação de informações.

A tarefa de selecionar tabelas, registros e atributos do Banco de Dados que serão estudados é fundamental no processo de descoberta de conhecimento. Principalmente porque, nas etapas iniciais, não se sabe exatamente quais informações são excessivas e quais são imprescindíveis.

Inicialmente, foram levantados todos os atributos existentes relacionados às unidades consumidoras conforme tabela 4.1. Entrevistas e discussões com especialistas da empresa foram realizadas com o intuito de compreender as informações que cada atributo contém.

Tabela: 4.1: Lista de atributos

Relação de atributos obtidos no sistema de cadastro da concessionária	
Nome	Inspeções Anteriores
Documento	Resultado
Endereço	Data
Razão	Nº
Localidade	Total recuperado kWh
Livro	Total recuperado R$
Classe	Regularização
Subclasse	
Tarifa	Alteração Cadastral
Telefone	Documento
Tipo medição	Endereço
Constante	Atividade
Nº Medidor	Titular
Posto Transformador	
Poste	Histórico de consumo
Carga	Média anual de consumo
Atividade	Consumo realiz.
Ultima Insp.	Consumo Faturado
Data Ligação	Irregularidades
Disjuntor	Ocorrências
Atrasos de pagamento	Data de Leitura
Nº Cortes	
Data ultimo Corte	Débitos
Débito Autmático (Banco,Agência,Conta)	Pendentes
Ultimo serviço Nº	Atrasos de Pagamento
Ultimo serviço Data	Data Pagamento
Titulares Anteriores	Data Vencimento
	Valor Total da Fatura
	Parcelamentos

Na Seção 4.3 é apresentado um descritivo das tabelas que compõem o banco de dados utilizado, enunciando cada atributo disponível. Posteriormente, na Seção 4.4, são apresentadas as etapas de pré-tratamento utilizadas na consolidação dos dados para mineração.

Os dados disponibilizados foram em sua maioria, oriundos da base de dados já consolidada na empresa. Estas informações históricas estão organizadas em um DWH que foi

iniciado na empresa em 1998. Este tipo de técnica simplesmente sintetiza os dados de sistemas transacionais que atendem a empresa. Como estes sistemas estão sujeitos a erros, muitas vezes são enviados dados errados para compor a síntese armazenada como histórico no DWH. Um fato importante, é que as concessionárias distribuidoras de energia devem cumprir a legislação do setor elétrico. A resolução 456 da ANEEL estabelece, em seu artigo 48, um valor mínimo de fatura para UC´s do grupo B (baixa tensão), equivalentes a 30 kWh para monofásicos, 50 kWh para bifásicos e 100 kWh para trifásicos. Este dispositivo tem objetivo de garantir a remuneração financeira à distribuidora em função do ativo instalado.

Uma unidade consumidora, ao longo do ano, pode variar seu padrão de consumo, de acordo com o clima, atividades, período de férias ou outros motivos. Nestes casos, há uma alteração também no histórico das informações sobre seu consumo. As empresas de energia têm, por exemplo, autorização para realizar um faturamento mínimo em UCs conforme já citado. Estes valores de consumo mínimo mascaram o comportamento do perfil dos consumidores já que não traduzem o valor de consumo real.

Pode-se ainda citar outra situação, onde o código de uma atividade pode ser 010 e descrever frigoríficos. Num dado momento no tempo, a empresa determina que este código deveria ser em separado para cada atividade, ou seja, 010 para frigoríficos de bovinos e 011 para frigoríficos de Suínos. As informações históricas não podem ser alteradas ou perdem seu sentido. Para eliminar esta situação, na modelagem do DWH são utilizadas outras formas de identificação única. Neste trabalho, o artifício utilizado é a criação de uma segunda chave, pertencente ao modelo no DWH. Esta segunda chave existe na tabela que descreve o fato, isto é, em dado momento, uma unidade consumidora foi medida, e neste momento sua classificação era frigoríficos.

Também na tabela de descrições, é criada uma chave específica do DWH para descrever o código 010 e os novos códigos 010 e 011, o problema que isso acarreta é que se é obrigado

a lidar com uma quantidade muito maior de informações e com uma complexidade muito maior nos relacionamentos dos quais serão realizadas extrações de informação.

O DWH é uma das principais ferramentas utilizadas pelos técnicos da empresa. Normalmente eles utilizam coleta de dados dos consumidores, informações sobre faturamento, informações sobre inspeções anteriores, e sobre cortes e religações de consumidores, além das informações sobre o consumo.

4.3 Banco de dados

Pode-se definir Banco de Dados como uma coleção de informações relacionadas entre si. Estas provêm de fatos conhecidos e que apresentam significado para quem os queira armazenar. Normalmente um Banco de Dados tem alguma origem da qual os dados são derivados, algum grau de interação com eventos do mundo real, e alguns usuários que estão ativamente interessados no seu conteúdo.

Esses Bancos de Dados podem ser complementados com um Sistema Gerenciador de Banco de Dados (SGBD), que consiste em uma coleção de programas que auxiliam o usuário a criar e manter um Banco de Dados, sendo um software com propósito geral de facilitar o processo no âmbito de definir, construir e manipular Bancos de Dados de várias aplicações.

Sua definição envolve especificar os tipos dos dados, estruturas e as restrições, para os dados que serão armazenados, depois se inicia o seu processo de construção que armazena os dados em alguma mídia que é controlada por um SGBD.

Após está montagem os dados se encontram disponibilizados para manipulação que inclui diversas funções que buscam por dados específicos, atualizações no banco de dados que refletem mudanças no seu conteúdo em particular e também na geração de relatórios.

Neste trabalho, acessou-se parte de um Banco de Dados, do período de novembro de 2002 à outubro de 2003. Esta parte do banco de dados está na forma de um arquivo do Microsoft Access que contém três tabelas, cujos atributos serão enunciados abaixo:

a) Tabela *Consumo*

- *Cons_Id*: identificação única para cada unidade consumidora (ou cliente). É um atributo do tipo cadeia de caracteres (por exemplo "98.989.89.989898");
- *Cons_Mes*: ano e mês das informações contidas no registro. Consiste de um tipo numérico com seis algarismos, identificado nos quatro primeiros o ano e nos dois últimos o mês (por exemplo 200211 e 200307);
- *Cons_Munic*: identificador numérico que representa o município onde a unidade consumidora está localizada (por exemplo 31);
- *Cons_Ativ*: código numérico que enquadra a unidade consumidora em alguma atividade, tendo maior distinção entre clientes comerciais e industriais (por exemplo 1101);
- *Cons_Tarifa*: informação da classe (residencial, comercial, industrial, etc.) e do tipo de ligação (monofásica, bifásica, trifásica ou primária) concatenadas em um único identificador do tipo cadeia de caracteres (por exemplo "01.10.01");
- *Cons_Trafo*: identificação numérica do transformador (ou poste) ao qual a unidade consumidora está conectada (por exemplo 123456789011);
- *Cons_Cons*: quantidade de energia elétrica consumida em kWh, no mês e ano de referência do registro (por exemplo 125).

b) Tabela *Inspeção*

- *Insp_Id*: utilizado para relacionar um registro de Inspeção a uma unidade consumidora de *Consumo*. Desta forma, armazena a mesma informação do atributo *Cons_Id*;
- *Insp_Data*: atributo que armazena o dia, mês e ano em que ocorreu uma inspeção, no formato data (por exemplo 08/19/2003);

- *Insp_Result*: cadeia de caracteres enunciando o resultado da inspeção (por exemplo “FRAUDE”).

c) Tabela *Trafos*

- *Traf_Trafo*: identificação única para cada transformador, permitindo um relacionamento com o atributo *Cons_Trafo* da tabela *Consumo*. Também está armazenado como um atributo numérico (por exemplo 987654321098);
- *Traf_Mes*: ano e mês das informações contidas no registro, sendo semelhante ao atributo *Cons_Mes*;
- *Traf_Cons*: soma das quantidades de energia elétrica consumida em kWh pelas unidades consumidoras conectadas no transformador, no mês e ano de referência do registro (por exemplo 11001).

4.4 Descoberta de conhecimento em banco de dados

Como descrito no capítulo 3, o processo de descoberta de conhecimento, através de Banco de Dados, é um conjunto de ações diversas que formam um mecanismo para que se obtenha um resultado a partir de informações contidas dentro do universo dos diversos dados existentes.

4.4.1 Montagem do banco de dados - Seleção e coleta de dados

O presente trabalho teve como etapa inicial a execução de algumas entrevistas com especialistas no assunto de duas concessionárias distribuidoras de energia.

Através dessas entrevistas obteve-se alguns parâmetros para determinar as informações de maior relevância na busca de unidades consumidoras com possíveis irregularidades em seu

sistema de medição de energia. E ainda auxiliou no entendimento do modus operandi dos técnicos e inspetores em campo.

Após a análise das informações disponíveis no sistema de informações de clientes da concessionária e também no DWH, foram selecionados os atributos classificados nas entrevistas, agregados aos atributos definidos pela equipe do projeto como importantes.

Os tipos de dados utilizados no trabalho foram: dados de cadastro, dados de Consumo e dados de Inspeção em unidades consumidoras.

A partir da existência do banco de dados completo da concessionária foi feita a seleção e a coleta dos dados, procedimento necessário para obter amostra de dados para o desenvolvimento do sistema.

4.4.1.1.Preparação dos dados: pré-processamento e consolidação dos dados

A preparação dos dados coletados é uma etapa importante em um projeto exploratório é necessária a escolha de um ambiente de trabalho, ou seja, um programa para desenvolver diversas tarefas relativas aos arquivos, obtidos através da etapa anterior (seleção e coleta de dados), tais como: criar e modificar arquivos, editar sua estrutura, consultar, filtrar e eliminar registros e ainda, executar comandos de linguagem SQL (linguagem compreensível para Bando de Dados).

A partir da preparação dos dados, é definida a ferramenta mais adequada para a manipulação deles. Inicia-se então, a análise de todos os arquivos obtidos, observando sempre a qualidade dos dados e também o seu grau de importância.

Todos os arquivos devem ser analisados e verificados, obtendo-se a definição dos quais serão finalmente utilizados. Assim inicia-se então a limpeza dos dados, eliminando-se os registros considerados indesejáveis ao processo.

Na base de dados trabalhados foram identificados uma série de registros com problemas, os quais estão listados a seguir com as respectivas ações implementadas para a solução:

1) Dados de consumo refletem o valor faturado e não valor efetivamente medido;

Inicialmente trabalhou-se como o consumo medido, o qual possuía muitos problemas, repetição de registros, consumos negativos, etc. Posteriormente foi utilizado o consumo faturado das unidades consumidoras em função dos dados parecerem mais confiáveis.

2) Existência de grande número de valores nulos nos dados de consumo;

A condição inicial proposta teve como premissa a eliminação dos registros que contemplassem consumo nulo, em uma segunda análise foi condicionado, para os casos onde a unidade consumidora ficasse com um número de meses muito reduzidos efetuava-se a eliminação do próprio consumidor.

3) Número de meses com consumos registrados variam de consumidor para consumidor;

Este foi um dos graves problemas enfrentados. Como extraíamos médias e variâncias deste consumo, clientes com um número muito pequeno de meses (4 ou 5) eram eliminados. Somente aqueles com 6 a 12 (ou mais) meses foram considerados.

4) Existência de registros com mais de uma inspeção em dias diferentes;

Um cliente pode ser inspecionado em qualquer dia da semana de 2ª feira à 6ª, no entanto em algumas situações foram verificadas mais de uma inspeção na unidade consumidora, sendo assim optou-se para os casos onde um cliente tivesse uma inspeção com resultado de fraude, ele já era considerado fraudador.

5) Existência de valores negativos registrados no consumo;

As unidades consumidoras que foram identificadas com consumos negativos foram eliminadas da base de dados.

6) Registros com valores repetidos;

Para clientes com registros repetidos, foi usada a condição "distinct" na consulta destes clientes, tomando um exemplar dos registros repetidos.

Alguns desses problemas que foram citados são resultantes da própria natureza dos dados, ou ainda, decorrentes de um processo falho de geração das bases de dados.

Apuradas e resolvidas as questões decorrentes da base de dados, iniciou-se a transformação desses dados em formato para a sua utilização.

4.4.1.2 Pré-processamento dos dados

A tabela *Consumo* trás como principal informação o consumo de energia elétrica de cada cliente, mês a mês, no período de novembro de 2002 à outubro de 2003. Esperava-se, portanto, que cada cliente tivesse 12 registros, um para cada mês do período amostrado. Porém, a tabela *Consumo* possui 7.266.819 registros e 642.720 clientes distintos, uma média de 11,3 registros por cliente. Notou-se então que em *Consumo* existem clientes com menos de 12 registros e outros com mais de 12 registros.

Já a tabela *Inspeção* possui 81.942 registros, cada um representando uma inspeção realizada em um conjunto de 64.326 clientes distintos. Do total de clientes inspecionados, 49.514 sofreram uma única inspeção e 14.812 sofreram pelo menos duas, entre novembro de 2002 à outubro de 2003.

A tabela *Trafos* possui 326.748 registros, cada um representando o consumo de energia elétrica em um dado trafo, no mês em questão. De um total de 42.040 trafos distintos, 29.286 possuem menos ou mais de 12 registros.

As informações quantitativas para *Consumo*, *Inspeção* e *Trafos* estão simplificadas na Tabela 4.2.

Tabela 4.2: Informações quantitativas do banco de dados.

Tabela	Número de registros	Elementos distintos
Consumo	7.266.819	642.720
Inspeção	81.942	64.326
Trafos	326.748	42.040

Após esta avaliação inicial do banco de dados, inicia-se um conjunto etapas de consolidação e mineração dos dados, nas quais foram realizadas operações específicas sobre as tabelas *Consumo*, *Inspeção* e *Trafo*.

4.4.1.3 Consolidação dos dados

Esta subseção apresenta o descritivo de cada etapa da consolidação e mineração dos dados, mostrando quais decisões foram tomadas na permanência e na eliminação de dados.

4.4.1.3.1 Relacionamento de consumo com inspeção

A primeira tarefa efetuada foi o relacionamento entre registros das tabelas *Consumo* e *Inspeção*, ou seja, verificar se há inspeção para um determinado cliente, em algum mês do período de amostragem. Quando uma inspeção foi relacionada à um cliente, o resultado da mesma foi adicionado à tabela *Consumo*, derivando uma nova tabela chamada *CI*. Portanto, a tabela *CI* contém todos os registros (mês a mês) das unidades consumidoras que receberam pelo menos uma inspeção, com o acréscimo do resultado desta inspeção no registro em que o mês de consumo coincide com a data da inspeção. Um resultado de inspeção nulo foi inserido nos registros com meses em que o cliente não recebeu inspeção. A tabela *CI*, além de receber os resultados de inspeção da tabela *Inspeção*, manteve todos os demais atributos contidos em

Consumo. Por este motivo, tanto *Consumo* quanto *Inspeção* deixaram de ser necessárias nas etapas seguintes, sendo substituídas apenas por *CI*.

Outra tarefa executada nesta etapa foi a decodificação do atributo *CI_Tarifa* em *CI_Cls* e *CI_TLig*. O novo atributo *CI_Cls* corresponde aos dois primeiros algarismos de *CI_Tarifa* e informa a que classe de serviço o cliente pertence, dentre elas: residencial (1), comercial (2), industrial (3), poder público (4), etc. Já o atributo *CI_TLig* corresponde ao dois algarismos finais de *CI_Tarifa* e informa qual o tipo de ligação do cliente, ou seja: monofásica, bifásica e trifásica. Extraído estes atributos de *CI_Tarifa*, o mesmo também deixou de ser necessário nas etapas seguintes.

O atributo *CI_DCons* foi criado à partir de *CI_Cons*, representando a variação de energia elétrica consumida pelo cliente, ou seja, o consumo no mês do registro menos o consumo no mês anterior. Obviamente, valores negativos de *CI_DCons* indicam que o cliente diminuiu o consumo em relação ao mês anterior. Um valor nulo foi inserido no primeiro registro, pois o mesmo não possui registro anterior para a subtração.

Objetivando um melhor entendimento das operações realizadas nesta etapa, a Tabela 4.3 ilustra alguns atributos da tabela *CI*, tomando como exemplo os registros de uma unidade consumidora anônima.

Ao final desta etapa, a tabela *CI* possuía 659.462 registros, distribuídos por 59.489 unidades consumidoras distintas. Sendo assim, dos 64.326 clientes distintos da tabela *Inspeção*, 4.837 (7,5%) não se relacionaram com clientes da tabela *Consumo*. Esta diferença ocorreu por dois motivos:

1. Um cliente da tabela *Inspeção* não está registrado como cliente da tabela *Consumo*;
2. Há registros para um dado cliente em ambas as tabelas, porém o mês da inspeção não coincide com o mês registrado em *Consumo*. Sendo assim o cliente não terá nenhuma inspeção e nenhum de seus registros na tabela *CI*.

Tabela 4.3: Registros de uma unidade consumidora anônima da tabela *CI*.

CI_Id	***CI_Mes***	***CI_TLig***	***CI_Cls***	***CI_Cons***	***CI_DCons***	***CI_Result***
0.000.00.00000	200211	23	2	570		NORMAL
0.000.00.00000	200212	23	2	700	130	
0.000.00.00000	200301	23	2	590	-110	
0.000.00.00000	200302	23	2	640	50	
0.000.00.00000	200303	23	2	550	-90	
0.000.00.00000	200304	23	2	630	80	
0.000.00.00000	200305	23	2	510	-120	
0.000.00.00000	200306	23	2	480	-30	
0.000.00.00000	200307	23	2	460	-20	
0.000.00.00000	200308	23	2	660	200	
0.000.00.00000	200309	23	2	470	-190	NORMAL
0.000.00.00000	200310	23	2	540	70	

4.4.1.3.2 Relacionamento de consumo e inspeção com trafos

Para relacionar as tabelas *CI* e *Trafos*, adicionando a cada registro de cliente o consumo do trafo em que o mesmo está conectado, utilizam-se as informações dos códigos dos trafos e do mês de referência. Ao avaliar o atributo *CI_Trafo*, foram encontrados 33.771 registros com valor "NAO SE APLICA", os quais foram descartados pela impossibilidade de relacionamento com *Trafos*. A tabela *CI* passou a ter 625.691 registros e 57.334 unidades consumidoras distintas.

O relacionamento entre *CI* e *Trafos*, chamado *CIT*, possui 473.152 registros e 47.987 unidades consumidoras. Esta redução considerável do número de registros de *CIT* em relação à *CI* (24%) ocorreu por fatores semelhantes aos da Subseção 4.4.1.3.1:

1. Um código de trafo (*CI_Trafo*) ou mês de referência (*CI_Mes*) em *CI* não possui intersecção em *Trafos*;
2. Um registro de *CI* com resultado de inspeção não-nulo não possui associação com nenhum registro de *Trafos*, levando a eliminação dos demais registros da unidade consumidora com resultado de inspeção nulo.

4.4.1.3.3 Concentração de registros

A tabela *CIT* compreende atributos originais e derivados de *Consumo*, *Inspeção* e *Trafos*, além de um conjunto de registros para cada unidade consumidora. Com o intuito de manter apenas um registro para cada unidade consumidora, primeiramente os clientes foram agrupados pela quantidade de meses (ou registros) que possuem em *CIT*. O resultado deste agrupamento pode ser visto na Tabela 4.4 e 4.5. A maioria das unidades consumidoras (67%) possui 10 registros, que é praticamente a média de registros por clientes distintos em *CIT* (9,85). Em contrapartida, há apenas 3 unidades consumidoras acima de 16 registros.

Após o agrupamento, foram descartados os clientes com número de registros menor que 4 e maior que 16, eliminando de *CIT* 1.268 registros de 633 clientes distintos. Também foram removidos 7.263 registros de 807 clientes, os quais possuíam pelo menos um mês com valores negativos para o atributo (*CIT_Cons*), sendo que o consumo mínimo esperado é zero. A tabela *CIT*, após as eliminações acima, passou a ter 464.621 registros de 46.547 clientes distintos.

Tabela 4.4: Unidades consumidoras da tabela *CIT* agrupadas pelo número de registros 1 à 10

Número de registros ou meses	**Número de unidades consumidoras**
1	280
2	119
3	231
4	391
5	616
6	753
7	989
8	1.152
9	1.568
10	32.329

Tabela 4.5: Unidades consumidoras da tabela *CIT* agrupadas pelo número de registros 11 à 20

Número de registros ou meses	**Número de unidades consumidoras**
11	5.813
12	3.521
13	157
14	40
15	15
16	10
18	1
19	1
20	1

A concentração das informações de clientes em um único registro é feita tomando qualquer um dos valores dos atributos estáticos e realizando alguma operação sobre os atributos dinâmicos, pois os mesmos variam seus valores, mês a mês, para cada cliente. Os atributos dinâmicos de *CIT* são:

1. Consumo de energia elétrica do cliente no mês (*CIT_Cons*);
2. Variação de consumo de energia elétrica em relação ao mês anterior (*CIT_DCons*);
3. Consumo de energia elétrica do trafo no mês (*CIT_TCons*).

Os três atributos dinâmicos deram origem aos cinco novos atributos abaixo, os quais possuem um único valor para cada unidade consumidora:

1. *CIT_Cmedia*: média entre os valores de (*CIT_Cons*), representando a média de consumo do cliente;
2. *CIT_Cdp*: desvio-padrão entre os valores de (*CIT_Cons*), representando o desvio-padrão do consumo do cliente;
3. *CIT_Tmedia*: média entre os valores de (*CIT_TCons*), representando a média de consumo do trafo em que o cliente está conectado;
4. *CIT_Tdp*: desvio-padrão entre os valores de (*CIT_TCons*), representando o desvio-padrão do consumo do trafo;

5. *CIT_Delta_Cmax*: mínimo entre os valores de (*CIT_DCons*), representando a diminuição máxima do consumo de energia.

4.4.1.3.4 Seleção de clientes normais e fraudadores

A tabela *CIT* passou a concentrar em 46.547 registros, um para cada cliente, todas as informações desejadas sobre as unidades consumidoras. A Tabela 4.6 ilustra a quantidade de clientes para cada possível resultado de inspeção. Como o objetivo deste trabalho é detectar os clientes fraudulentos, somente aqueles que possuem resultado “NORMAL” ou “FRAUDE” foram selecionados. Desta forma, a tabela *CIT* foi renomeada para *CIT_NF* e passou a ter 41.290 registros, sendo 95,4% de clientes normais e 4,6% de fraudadores.

A tabela *CIT_NF* faz parte de um banco de dados do Microsoft Access, juntamente com as demais tabelas intermediárias ao pré-tratamento de dados. Porém, o processo de descoberta de conhecimento foi realizado, usando-se o programa MATLAB, o qual apresenta várias ferramentas para a manipulação de matrizes (que podem ser vistas como tabelas). Sendo assim, os atributos da tabela *CIT_NF* foram exportados para o MATLAB, onde cada atributo é um vetor numérico ou de caracteres com 41.290 elementos.

Uma última eliminação de clientes foi realizada sobre os registros (ou linhas no MATLAB) que apresentaram valor zero para média de consumo do cliente ou do trafo. Esta remoção não foi realizada na tabela *CIT_NF* do Microsoft Access, pois acreditava-se que registros com médias nulas seriam importantes no processo de mineração, o que não foi comprovado posteriormente. O tamanho final dos vetores de atributos no MATLAB é de 40.492 elementos, onde 38.621 (95,4%) possuem resultado de inspeção normal, enquanto 1.871 (4,6%) apresentam resultado fraudulento.

A tabela 4.6 apresenta a composição dos dados oriundos da tabela *CIT* conjuntamente com os resultados das inspeções efetuadas. Tais resultados são cadastrados de acordo com as opções cadastrais disponíveis no sistema de banco de dados da concessionária.

Tabela 4.6: Unidades consumidoras da tabela *CIT* agrupadas pelos resultados de inspeção

Resultado de Inspeção	**Número de unidades consumidoras**
NORMAL	39.389
FRAUDE	1.901
FALHA DE MEDIÇÃO	1.821
IRREGULARIDADE COMERCIAL	1.518
IMPEDIMENTO	1.432
AUTORELIGAMENTO	426
IRREGULARIDADE TÉCNICA	60

A tabela 4.7 apresenta o conjunto de atributos disponíveis para o processo de mineração, informando seus possíveis valores e a que tipo ou classe do MATLAB pertencem.

Tabela 4.7: Conjunto de atributos disponíveis para o processo de mineração

N°	**Atributos**	**Valores Distintos**	**Tipo**	**Distribuição**
1	*Id*	40.492	Texto	Categórico
2	*Resultado_Str*	NORMAL OU FRAUDE	Texto	Categórico
3	*Resultado_Num*	2	Numérico	Categórico
4	*Atividade*	449	Numérico	Categórico
5	*Classe*	8	Numérico	Categórico
6	*Tipo_Lig*	4	Numérico	Categórico
7	*Município*	72	Numérico	Categórico
8	*Media_Consumo*	12.834	Numérico	Contínuo
9	*Dp_Consumo*	35.171	Numérico	Contínuo
10	*Delta_Consumo*	32.250	Numérico	Contínuo
11	*Media_Trafo*	14.242	Numérico	Contínuo
12	*Dp_Trafo*	14.253	Numérico	Contínuo

4.4.2 Transformação dos dados

Após a preparação dos dados, estes foram disponibilizados em plataforma e padrão compatíveis com a ferramenta que será utilizada, permitindo fazer relacionamentos diversos com os dados disponíveis nos arquivos.

Nesta etapa, foi realizada uma nova análise dos dados disponibilizados através da integração dos mesmos. O objetivo foi delimitar aqueles que realmente poderão contribuir com o trabalho e seus objetivos finais.

Definiu-se os dados que deveriam ser desconsiderados; isto após a avaliação das especificações dos campos e seus respectivos valores, através de novas filtragens. Sendo finalmente gerada uma amostra definitiva de dados para dar seguimento no processo.

O objetivo desta fase é facilitar a etapa posterior (mineração de dados). Para isso, foi avaliada a importância de cada campo dos arquivos, bem como a compatibilidade dos dados com a técnica de mineração a ser aplicada, visando diminuir o volume de processamento, facilitar a análise e interpretação dos resultados e, caso necessário, ainda criar alguns campos e transportar dados para eles, também é objetivo da preparação e transformação dos dados a redução do número de variáveis a se considerar ou achar representações invariáveis para os dados.

4.4.3 Mineração de dados de dados utilizando Árvore de Decisão

A Mineração de dados é a etapa que se procura padrões nos dados, é uma fase essencial no contexto geral do processo de descoberta de informação através de banco de dados onde métodos inteligentes são aplicados para extrair padrões.

O processo de mineração de dados é geralmente bastante interativo, pois a seleção de dados pode ser revista sempre que a informação não atingir a expectativa esperada.

Os algoritmos podem e devem ser reajustados quando poucos fatos interessantes são descobertos, durante o passo de assimilação, tornando-se assim um processo que possui laços de realimentação.

O sistema de geração das árvores de decisão pode gerar árvores extremamente complexas que acabam perdendo o seu real poder de predição. Para se conseguir gerar uma árvore de decisão com boa precisão é necessário fazer a escolha correta dos atributos que serão usados na análise. Estes atributos devem gerar uma árvore com o menor número possível de subconjuntos, assim chegando a cada folha da árvore com um número razoável de ramos. Isto é, o ideal é escolher os atributos de modo que a árvore final seja a menor possível. Como analisar todas as possibilidades possíveis seria algo absurdo, foram desenvolvidos métodos para a escolha dos atributos e dos testes a serem utilizados, e uma vez feita à escolha as outras possibilidades não foram mais exploradas. Portanto, o objetivo inicial sempre foi tornar a árvore de decisão mais simples possível, através da simplificação e principalmente através da escolha correta dos atributos existentes no banco de dados que realmente fossem relevantes. Esta etapa foi de importância fundamental e demandou aproximadamente 40% do tempo total da pesquisa.

Após várias análises e simulações considerando os atributos relacionados anteriormente em diversas combinações, foram selecionados apenas cinco atributos abaixo identificados e que apresentaram os melhores resultados para se obter a menor árvore possível com o menor número possível de atributos:

- A classe de consumidor → [1 2 3 4 5 6 8] → sendo um atributo discreto;

- Atividade do consumidor → com 392 diferentes atividades que receberam um pré-processamento e foram reduzidas para apenas 20 diferentes atividades → sendo um atributo discreto;

- Tipo de ligação do consumidor → [1 2 3 11 12 13 21 22 23 31 32 33 43 44 53 54 63 64] → sendo um atributo discreto;

- Saldo de consumo mensal do consumidor → sendo um atributo contínuo;

- O consumo padrão da região do consumidor → sendo um atributo contínuo.

Neste trabalho foi utilizado o banco de dados de uma distribuidora de energia elétrica para buscar predizer as características dos clientes fraudadores de energia e das prováveis falhas de medição. O processo de aprendizagem envolve a apresentação de um conjunto de exemplos de dados cuja saída é conhecida, o que foi chamado de conjunto de treinamento. A classificação é uma tarefa de previsão onde um conjunto de atributos previsores é usado para prever o atributo objetivo, neste caso a fraude na medição do consumo de energia elétrica e as falhas de medição. Nesta etapa de treinamento foi utilizada a parte do banco de dados que continha os consumidores já inspecionados durante um período de um ano. Uma parte destes dados em média (50%) foi de fato utilizada para o treinamento e uma outra parte (50%) foi utilizada para teste inicial das regras de previsão obtidas na árvore de decisão.

Existem diversos algoritmos de classificação que permitem elaborar as árvores de decisão. É difícil determinar qual é o melhor algoritmo. Dependendo da situação, um pode ter melhor desempenho em relação ao outro. De forma geral, podem-se destacar os seguintes algoritmos: CART ("Classification and Regression Trees" - Breiman, 84), ID3 ("Induction Decision Tree" - Quinlan, 86) e C4.5 ("Continous" - de Quinlan, 93). CHAID ("Chi-Square Automatic Interaction Detection"). O algoritmo ID3 foi um dos primeiros algoritmos de árvore de decisão, tendo sua elaboração baseada em sistemas de inferência e em conceitos de sistemas de aprendizagem e o algoritmo C4.5 é uma extensão do ID3 que acrescenta valores contínuos. Existem diversos programas que implementam o método de classificação por árvores de decisão. Muitos deles são específicos, o que solicita pouco conhecimento do processo por parte do usuário. Por outro lado, existem também os genéricos, os quais necessita de um maior grau de conhecimento do processo por parte dos usuários. Com base no algoritmo e nos objetivos desejados e pela sua disponibilidade o programa selecionado foi o "MATLAB" da "The MathWorks" no sistema "Windows". O Matlab é uma ferramenta

matemática para computação numérica e manipulação de dados, com excelente capacidade gráfica e provê muitos recursos para a manipulação e processamento numéricos de um grande conjunto de dados e foi utilizado para implementar o método de classificação por árvore de decisão.

No “MATLAB” existe o chamado “Statistics Toolbox” que possui todos os comandos para gerar, desenhar e testar a árvore de decisão em função dos dados existentes. Existem também diversos recursos adicionais, tais como arquivos para serem executados pelo “MATLAB”, que implementam os algoritmos de árvore de decisão e que podem ser obtidos de forma adicional.

Com base nos atributos determinados e no arquivo de treinamento utilizado, foram geradas várias árvores para uma avaliação de performance do sistema de identificação de fraudes e erros de medição, a figura 4.1 mostra uma dessas árvores, onde pode-se observar a complexidade de sua topologia.

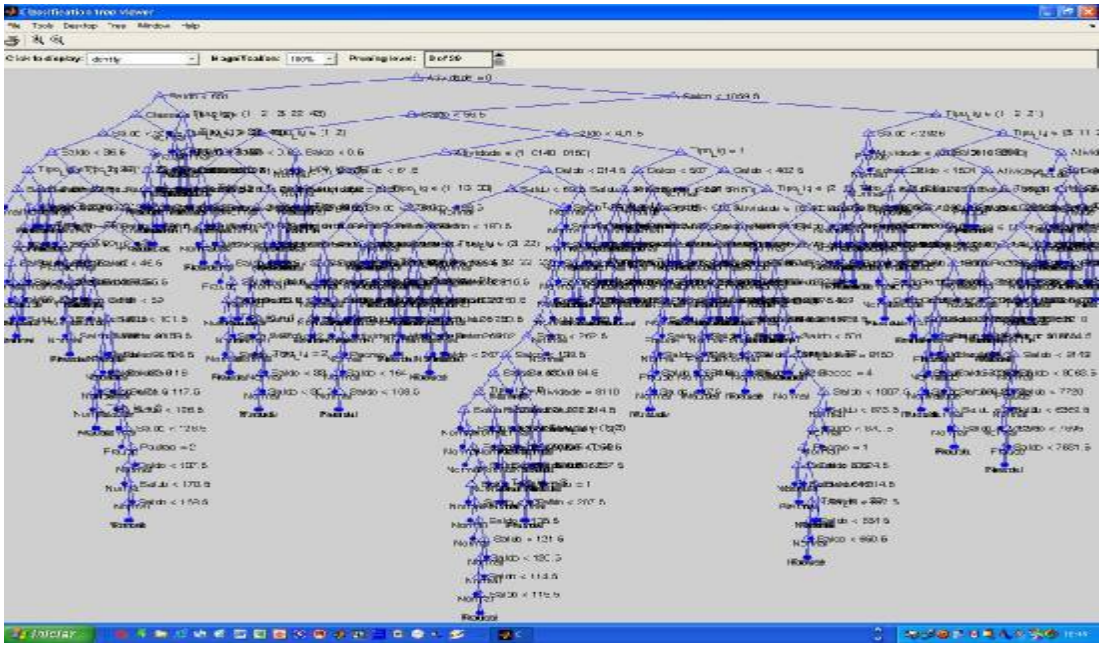

Figura 4.1 Árvore de decisão com 5 atributos

Para melhor visualização e interpretação mostramos na figura 4.2 uma ampliação (zoom) de parte da figura 4.1 do sistema de identificação de fraudes e erros de medição, onde pode-se verificar de maneira mais visível algumas ramificações da árvore com seus atributos e resultados diagnosticados.

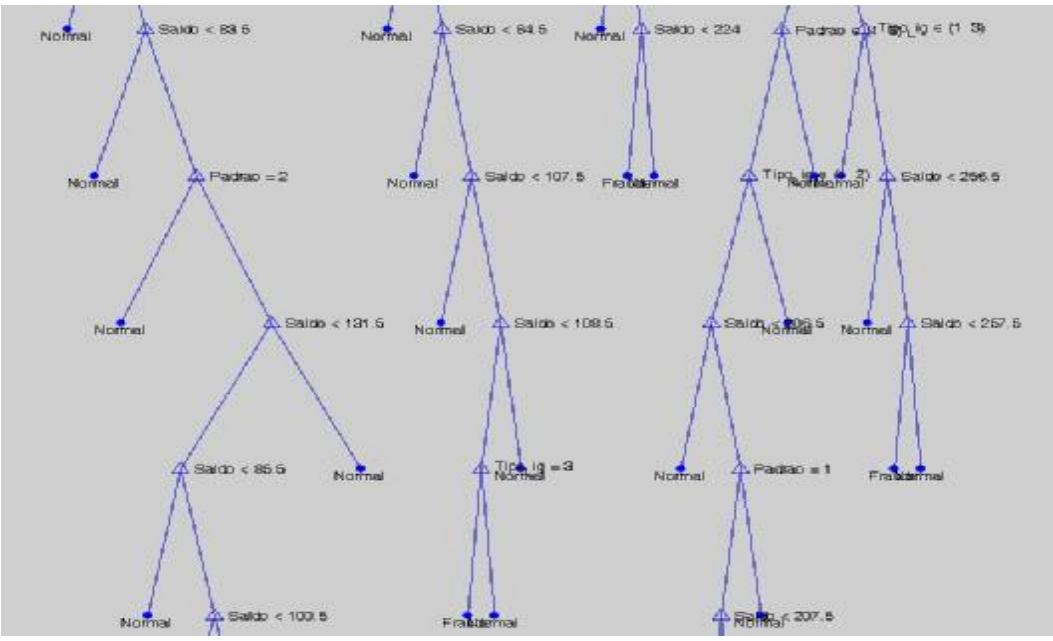

Figura 4.2 Parte da árvore de decisão com 5 atributos

Atualmente a taxa média de detecção de fraudes e falhas de medição nas inspeções é de aproximadamente 5 %. Isto é, para cada 100 consumidores inspecionados em campo apenas 5 fraudes e ou falhas de medição são detectadas.

Os atributos previamente selecionados destes consumidores foram aplicados à Árvore de Decisão previamente gerada. Foi obtida uma lista de consumidores prováveis fraudadores e ou com possíveis falhas de medição.

Utilizando-se os dados das diversas árvores de decisão que foram elaboradas no decorrer dos experimentos conforme modelo mostrada na Fig. 4.1, foram avaliados percentualmente os resultados desta pesquisa, baseados no banco de dados dos consumidores já inspecionados com resultados já conhecidos.

Os resultados obtidos alcançaram uma taxa média de detecção de fraudes e falhas de medição nas inspeções de até mais de 40%, o que representa um crescimento da ordem de 8 vezes em relação ao valor médio obtido pela concessionária.

Embora o resultado tenha apresentado um bom desempenho e aumentado consideravelmente o índice de acerto nas inspeções e desta maneira, reduzido muito o custo das inspeções para a companhia local de distribuição de energia elétrica, ele garante apenas o aumento da eficácia das inspeções, porém, ele não garante que todas as fraudes e falhas de medição estejam dentro do conjunto de consumidores a serem inspecionados. Isto significa que os custos das inspeções podem ainda ser efetivamente reduzido.

4.4.4 Estudo de casos

Mostraremos abaixo alguns dos estudos de casos que foram avaliados durante os experimentos do trabalho, foram executadas simulações com uma amostragem de dados de consumidores.

Estudou-se o comportamento do sistema, utilizando-se uma base de dados para treinamento e uma segunda amostragem de dados que foram utilizados para teste.

A partir dos resultados obtidos definiu-se uma matriz de eficiência para demonstração dos resultados.

Mostra-se a seguir as equações das referidas análises:

TA – Total da Amostra

FI – Fraudes inferidas

FIC – Fraudes inferidas e confirmadas

NFI – Não fraudes inferidas

NFIC – Não fraudes inferidas e confirmadas

$$\mathbf{TA = FI + NFI} \qquad (4.1)$$

Onde:

TA – Total da Amostra

FI – Fraudes inferidas

NFI – Não fraudes inferidas

$$\mathbf{FI = TA - NFI} \qquad (4.2)$$

Onde:

FI – Fraudes inferidas

TA – Total da Amostra

NFI – Não fraudes inferidas

$$\text{NFI} = \text{TA} - \text{FI} \qquad (4.3)$$

Onde:

NFI – Não fraudes inferidas

TA – Total da Amostra

FI – Fraudes inferidas

A partir da equação 4.4 é possível mensurar o percentual de assertividade na identificação de fraudes.

$$TAF = \frac{FIC}{FI} \Rightarrow TAF\% = \frac{FIC}{FI} * 100 \qquad (4.4)$$

Onde:

TAF – Taxa de acerto de Fraudes

FIC – Fraudes inferidas e confirmadas

FI – Fraudes inferidas

A equação 4.5 identifica percentualmente a margem de erro na identificação das fraudes no sistema.

$$TEF = 1 - \frac{FIC}{FI} = \frac{FI - FIC}{FI} \Rightarrow TEF\% = \frac{FI - FIC}{FI} * 100 \qquad (4.5)$$

Onde:

TEF – Taxa de erro de Fraude

FIC – Fraudes inferidas e confirmadas

FI – Fraudes inferidas

Ou ainda, substituindo (4.2) em (4.5)

$$TEF = \frac{(TA - NFI) - FIC}{TA - NFI} \Rightarrow TEF\% = \frac{(TA - NFI) - FIC}{TA - NFI} * 100 \qquad (4.6)$$

Onde:

TEF – Taxa de erro de Fraude

TA – Total da Amostra

NFI – Não fraudes inferidas

FIC – Fraudes inferidas e confirmadas

Para a avaliação dos casos onde é observada a performance de acerto na resposta de não fraudes utiliza-se a equação (4.7), descrita a seguir:

$$TANF = \frac{NFIC}{NFI} \Rightarrow TANF\% = \frac{NFIC}{NFI} * 100 \qquad (4.7)$$

Onde:

TANF - Taxa de acerto de não-fraude

NFIC – Não fraudes inferidas e confirmadas

NFI – Não fraudes inferidas

A equação seguinte (4.8) descreve a situação da taxa de erro dos casos de não fraudador

$$TENF = 1 - \frac{NFIC}{NFI} = \frac{NFI - NFIC}{NFI} \Rightarrow TENF\% = \frac{NFI - NFIC}{NFI} * 100 \qquad (4.8)$$

Onde:

TENF - Taxa de erro de não-fraude

NFIC – Não fraudes inferidas e confirmadas

NFI – Não fraudes inferidas

Ou ainda, substituindo (4.3) em (4.8) tem-se:

$$TENF = \frac{(TA - FI) - NFIC}{(TA - FI)} \Rightarrow TENF\% = \frac{(TA - FI) - NFIC}{(TA - FI)} * 100 \qquad (4.9)$$

A partir das equações (4.1) a (4.9) foi elaborada uma matriz conforme mostra a figura 4.1, cujo objetivo é demonstrar a eficiência do sistema desenvolvido. Os índices TAF e TANF (diagonal) devem aproximar de 1 a medida que a eficiência do sistema melhora, e ou os índices TEF e TENF se aproximar de zero.

Sistema / Inspeção	FI	NFI
FIC	TAF	TEF
NFIC	TENF	TANF

Figura 4.3 Matriz de eficiência do sistema

Mostra-se a seguir os principais estudos de casos avaliados durante o desenvolvimento deste trabalho, será apresentado para cada caso uma descrição dos atributos utilizados e suas característica, uma tabela descritiva dos valores apurados em cada simulação e ainda a matriz de eficiência do sistema demonstrando a performance através dos seus resultados.

Caso 01

O sistema é avaliado considerando os seguintes atributos:

1 - Atividade: atributo discreto com originalmente 392 diferentes atividades, as quais foram reduzidas para apenas 6 tipos diferentes;

2 - Classe: atributo discreto com 7 tipos diferentes [1 2 3 4 5 6 8];

1 – Residencial;

2 – Comercial;

3 – Industrial;

4 - Poder Publico;

5 – Rural;

6 - Serviço Publico;

7 - Iluminação Publica;

8 - Consumo Próprio.

Nota: A classe Iluminação Pública não foi considerada e as classes Residencial e Comercial corresponde a mais de 95 % dos clientes.

3 - Tipo de ligação: atributo discreto com 15 tipos diferentes: [1 2 3 11 12 13 21 22 23 31 32 33 43 44 53 54 63 64];

4 - Consumo: atributo contínuo.

Arquivo utilizado: normal_fraude_LR_junho_sem_mes.m

Arquivo de treino = 25882 consumidores

Arquivo de teste = 25882 consumidores

Tabela 4.8: Análise quantitativa das fraudes – caso 01

Análise das Fraudes	Quant.
1 - Total geral correta de Fraudes no arquivo de teste	1197
2 – Porcentagem do Total (25882)	4,62 %
3 - Total de Fraudes encontradas no arquivo de teste	380
4 - Fraudes Erradas encontradas no arquivo de teste	258
5 - Fraudes Corretas encontradas no arquivo de teste	122
6 - Relação das Corretas com o Total de Fraudes encontradas (122/380)	32,11 %
7 - Porcentagens de acertos nas Fraudes (relação ao total geral) (122/1197)	10,19 %

Sistema / Inspeção	FI	NFI
FIC	0,3211	0,6789
NFIC	0,0422	0,9578

Figura 4.4 Matriz de eficiência do sistema – caso 01

Análise do resultado: A porcentagem de acerto nas inspeções foi de 32,11 %, valor acima do valor atual de acerto nas inspeções que é de aproximadamente 5 %, contudo apenas 10,19 % dos reais fraudadores foram detectados, os demais não foram identificados pelo sistema.

Caso 02

O sistema é avaliado considerando os seguintes atributos:

1 - Atividade: atributo discreto com originalmente 392 diferentes atividades, as quais foram reduzidas para apenas 6 tipos diferentes;

2 - Classe: atributo discreto com 7 tipos diferentes [1 2 3 4 5 6 8];

3 - Tipo de ligação: atributo discreto com 15 tipos diferentes: [1 2 3 11 12 13 21 22 23 31 32 33 43 44 53 54 63 64];

4 - Consumo: atributo contínuo;

5 – Padrão de consumo: Foi divido em 3 classes: (0 - 100 kWh, 101 – 300 kWh, acima de 300 kWh);

Arquivo utilizado: normal_fraude_LR_junho_padrao2.m

Arquivo de treino = 25882 consumidores

Arquivo de teste = 25882 consumidores

Tabela 4.9: Análise quantitativa das fraudes – caso 02

Análise das Fraudes	Quant.
1 - Total geral correta de Fraudes no arquivo de teste	1197
2 – Porcentagem do Total (25882)	4,62 %
3 - Total de Fraudes encontradas no arquivo de teste	407
4 - Fraudes Erradas encontradas no arquivo de teste	278
5 - Fraudes Corretas encontradas no arquivo de teste	129
6 - Relação das Corretas com o Total de Fraudes encontradas (129/407)	31,7 %
7 – Porcentagens de acertos nas Fraudes (relação ao total geral) (129/1197)	10,78 %

Sistema / Inspeção	FI	NFI
FIC	0,3169	0,6831
NFIC	0,0419	0,9581

Figura 4.5 Matriz de eficiência do sistema – caso 02

Análise do resultado: A porcentagem de acerto nas inspeções foi de 31,7 %, valor acima do valor atual de acerto nas inspeções que é de aproximadamente 5 %, contudo apenas 10,78 % dos reais fraudadores foram detectados, os demais não foram identificados pelo sistema. Percebe-se, que o acréscimo do atributo Padrão de Consumo não alterou de forma significativa os resultados.

Caso 03

Neste caso, o sistema é avaliado considerando os seguintes atributos:

1 - Atividade: atributo discreto com originalmente 392 diferentes atividades, as quais foram reduzidas para apenas 6 tipos diferentes;

2 - Classe: atributo discreto com 7 tipos diferentes [1 2 3 4 5 6 8];

3 - Tipo de ligação: atributo discreto com 15 tipos diferentes: [1 2 3 11 12 13 21 22 23 31 32 33 43 44 53 54 63 64];

4 - Consumo: atributo contínuo;

5 – Padrão de consumo: Foi divido em 5 classes: (0 - 30 kWh, 31 – 100 kWh, 101 – 200 kWh, 200 – 300 kWh, acima de 300 kWh);

Arquivo utilizado: normal_fraude_LR_junho_padrao2.m

Arquivo de treino = 25882 consumidores

Arquivo de teste = 25882 consumidores

Tabela 4.10: Análise quantitativa das fraudes – caso 03

Análise das Fraudes	Quant.
1 - Total geral correta de Fraudes no arquivo de teste	1197
2 – Porcentagem do Total (25882)	4,62 %
3 - Total de Fraudes encontradas no arquivo de teste	457
4 - Fraudes Erradas encontradas no arquivo de teste	329
5 - Fraudes Corretas encontradas no arquivo de teste (FIC)	128
6 - Relação das Corretas com o Total de Fraudes encontradas (128/457)	28,01 %
7 - Porcentagens de acertos nas Fraudes (relação ao total geral) (128/1197)	10,69 %

Sistema / Inspeção	FI	NFI
FIC	0,2801	0,7199
NFIC	0,0420	0,9580

Figura 4.6 Matriz de eficiência do sistema – caso 03

Análise do resultado: A porcentagem de acerto nas inspeções foi de 28,01 %, valor acima do valor atual de acerto nas inspeções que é de aproximadamente 5 %, contudo apenas 10,69 % dos reais fraudadores foram detectados, os demais não foram identificados pelo sistema. Neste caso observa-se que o acréscimo do atributo Padrão de Consumo e o aumento do número de classes para o padrão de consumo não modificou significativamente o resulltado.

Caso 04

Neste caso, o sistema é avaliado considerando os seguintes atributos:

1 - Atividade: atributo discreto com originalmente 392 diferentes atividades foi considerado contínuo.

2 - Classe: atributo discreto com 7 tipos diferentes [1 2 3 4 5 6 8];

3 - Tipo de ligação: atributo discreto com 15 tipos diferentes : [1 2 3 11 12 13 21 22 23 31 32 33 43 44 53 54 63 64];

4 - Consumo: atributo contínuo;

5 – Padrão de consumo: Foi divido em 5 classes: (0 - 30 kWh, 31 – 100 kWh, 101 – 200 kWh, 200 – 300 kWh, acima de 300 kWh);

Arquivo utilizado: normal_fraude.m

Arquivo de treino = 19989 consumidores

Arquivo de teste = 19989 consumidores

Tabela 4.11: Análise quantitativa das fraudes – caso 04

Análise das Fraudes	Quant.
1 - Total geral correta de Fraudes no arquivo de teste	928
2 – Porcentagem do Total (19989)	4,64 %
3 - Total de Fraudes encontradas no arquivo de teste	230
4 - Fraudes Erradas encontradas no arquivo de teste	125
5 - Fraudes Corretas encontradas no arquivo de teste	105
6 - Relação das Corretas com o Total de Fraudes encontradas (105/230)	45,65 %
7 - Porcentagens de acertos nas Fraudes (relação ao total geral) (105/928)	11,31 %

Sistema / Inspeção	FI	NFI
FIC	0,4565	0,5435
NFIC	0,0417	0,9583

Figura 4.7 Matriz de eficiência do sistema – caso 04

Análise do resultado: Neste caso foi utilizado um novo arquivo de treino e teste com 19989 consumidores. A porcentagem de acerto nas inspeções subiu para 45,65 %, valor considerável, contudo apenas 11,31 % dos reais fraudadores foram detectados, os demais não foram identificados pelo sistema. Observa-se que para este caso onde considerou-se Atividade um atributo contínuo, o acerto teve um aumento em relação as situações anteriores contudo a detecção do total dos fraudadores continuou baixa.

Caso 05

Neste caso, o sistema é avaliado considerando os seguintes atributos:

1 - Classe: atributo discreto com 7 tipos diferentes [1 2 3 4 5 6 8];

2 - Tipo de ligação: atributo discreto com 15 tipos diferentes: [1 2 3 11 12 13 21 22 23 31 32 33 43 44 53 54 63 64];

3 - Consumo: atributo contínuo;

4 – Padrão de consumo: Foi divido em 5 classes: (0 - 30 kWh, 31 – 100 kWh, 101 – 200 kWh, 200 – 300 kWh, acima de 300 kWh);

Nota: O atributo Atividade foi retirado.

Arquivo utilizado: normal_fraude_sem_atividade.m

Arquivo de treino = 19989 consumidores

Arquivo de teste = 19989 consumidores

Tabela 4.12: Análise quantitativa das fraudes – caso 05

Análise das Fraudes	Quant.
1 - Total geral correta de Fraudes no arquivo de teste	928
2 – Porcentagem do Total (19989)	4,64 %
3 - Total de Fraudes encontradas no arquivo de teste	988
4 - Fraudes Erradas encontradas no arquivo de teste	805
5 - Fraudes Corretas encontradas no arquivo de teste	183
6 - Relação das Corretas com o Total de Fraudes encontradas (183/988)	18,52 %
7 - Porcentagens de acertos nas Fraudes (relação ao total geral) (183/928)	19,72 %

Sistema / Inspeção	FI	NFI
FIC	0,1855	0,8145
NFIC	0,0392	0,9608

Figura 4.8 Matriz de eficiência do sistema – caso 05

Análise do resultado: Neste caso foi retirado o atributo Atividade. A porcentagem de acerto nas inspeções foi baixa de 18,52 %, valor ainda acima do valor atual de acerto nas inspeções que é de aproximadamente 5 %, contudo abaixo dos resultados obtidos anteriormente. Houve melhora no número dos reais fraudadores que foram detectados (19,72%). Entende-se que houve uma piora no acerto das inspeções e uma melhora na detecção geral de fraudes.

Caso 06

Neste caso, o sistema é avaliado considerando os seguintes atributos:

1 - Atividade: atributo discreto com originalmente 392 diferentes atividades foi considerado contínuo.

2 - Classe: atributo discreto com 7 tipos diferentes [1 2 3 4 5 6 8];

3 - Tipo de ligação: Atributo discreto com 15 tipos diferentes: [1 2 3 11 12 13 21 22 23 31 32 33 43 44 53 54 63 64];

4 - Consumo: atributo contínuo;

5 – Padrão de consumo: Foi divido em 5 classes: (0 - 30 kWh, 31 – 100 kWh, 101 – 200 kWh, 200 – 300 kWh, acima de 300 kWh);

Arquivo utilizado: normal_fraude_menor.m

Arquivo de treino = 3744

Nota: O arquivo de treinamento foi diminuído para alterar a relação de consumidores normais e fraudadores para 3 / 1.

Arquivo de teste = 19989 consumidores

Tabela 4.13: Análise quantitativa das fraudes – caso 06

Análise das Fraudes	Quant.
1 - Total geral correta de Fraudes no arquivo de teste	928
2 - Porcentagem do Total (19989)	4,64 %
3 - Total de Fraudes encontradas no arquivo de teste	1363
4 - Fraudes Erradas encontradas no arquivo de teste	1087
5 - Fraudes Corretas encontradas no arquivo de teste	276
6 - Relação das Corretas com o Total de Fraudes encontradas (276/1363)	20,25 %
7 - Porcentagens de acertos nas Fraudes (relação ao total geral) (276/928)	29,74 %

Sistema / Inspeção	FI	NFI
FIC	0,2025	0,7975
NFIC	0,0350	0,9650

Figura 4.9 Matriz de eficiência do sistema – caso 06

Análise do resultado: A porcentagem de acerto nas inspeções foi baixa de 20,25 %, valor ainda acima do valor atual de acerto nas inspeções que é de aproximadamente 5 %, contudo abaixo de alguns resultados obtidos anteriormente. Houve melhora no número dos reais fraudadores detectados (29,74). Percebe-se que houve uma piora no acerto das inspeções e uma melhora na detecção geral de fraudes. A alteração nos resultados foi causada pela diminuição do arquivo de treinamento e pela alteração na sua relação de consumidores normais e fraudadores para 3 para 1.

Caso 07

Neste caso, o sistema é avaliado considerando os seguintes atributos:

1 - Atividade: atributo discreto com originalmente 392 diferentes atividades, as quais foram reduzidas para apenas 19 tipos diferentes;

2 - Classe: atributo discreto com 7 tipos diferentes [1 2 3 4 5 6 8];

3 - Tipo de ligação: atributo discreto com 13 tipos diferentes: [1 2 3 11 12 13 21 22 23 31 32 33 43] - retirado = [44 53 54 63 64];

4 - Consumo: Atributo contínuo.

Arquivo utilizado: normal_fraude_LR.m

Arquivo de treino = 19627 consumidores

Arquivo de teste = 19592 consumidores

Tabela 4.14: Análise quantitativa das fraudes – caso 07

Análise das Fraudes	Quant.
1 - Total geral correta de Fraudes no arquivo de teste	905
2 - Porcentagem do Total (19592)	4,62 %
3 - Total de Fraudes encontradas no arquivo de teste	254
4 - Fraudes Erradas encontradas no arquivo de teste	146
5 - Fraudes Corretas encontradas no arquivo de teste (FIC)	108
6 - Relação das Corretas com o Total de Fraudes encontradas (108/254)	42,52 %
7 - Porcentagens de acertos nas Fraudes (relação ao total geral) (108/905)	11,93 %

Sistema / Inspeção	FI	NFI
FIC	0,4252	0,5748
NFIC	0,0404	0,9596

Figura 4.10 Matriz de eficiência do Sistema – caso 07

Análise do resultado: A porcentagem de acerto nas inspeções foi de 42,52 %, valor acima do valor atual de acerto nas inspeções que é de aproximadamente 5 %, contudo apenas 11,93 % dos reais fraudadores foram detectados, os demais não foram identificados pelo sistema.

Caso 08

Neste caso, o sistema é avaliado considerando os seguintes atributos:

1 - Atividade: atributo discreto com originalmente 392 diferentes atividades, as quais foram reduzidas para apenas 19 tipos diferentes;

2 - Classe: atributo discreto com 7 tipos diferentes [1 2 3 4 5 6 8];

3 - Tipo de ligação: atributo discreto com 15 tipos diferentes: [1 2 3 11 12 13 21 22 23 31 32 33 43 44 53 54 63 64];

4 - Consumo: Atributo contínuo;

5 – Padrão de consumo: Foi divido em 3 classes: (0 - 100 kWh, 101 – 300 kWh, acima de 300 kWh);

Arquivo utilizado: normal_fraude_mesmo_arquivo.m

Nota: Neste caso foi utilizado o mesmo arquiva para treino e teste.

Arquivo de treino = 19989 consumidores

Arquivo de teste = 19989 consumidores

Tabela 4.15: Análise quantitativa das fraudes – caso 08

Análise das Fraudes	Quant.
1 - Total geral correta de Fraudes no arquivo de teste	936
2 - Porcentagem do Total (19989)	4,68 %
3 - Total de Fraudes encontradas no arquivo de teste	261
4 - Fraudes Erradas encontradas no arquivo de teste	60
5 - Fraudes Corretas encontradas no arquivo de teste	201
6 - Relação das Corretas com o Total de Fraudes encontradas (201/261)	77,01 %
7 - Porcentagens de acertos nas Fraudes (relação ao total geral) (201/936)	21,47 %

Sistema / Inspeção	FI	NFI
FIC	0,7701	0,2299
NFIC	0,0373	0,9627

Figura 4.11 Matriz de eficiência do sistema – caso 08

Análise do resultado: Neste caso o mesmo arquivo utilizado para treino foi utilizado no teste, o que resultou em alto índice de acerto nas Fraudes detectadas e também houve uma melhora no total de acerto nas Fraudes. Nota-se que apesar de ser utilizado o mesmo arquivo para treino e teste, o número de acertos não chegou a 100% em função da árvore de decisão

possuir alguma imprecisão, devido alguns dados dos atributos utilizados serem muito próximos.

Caso 09

Neste caso, o sistema é avaliado considerando os seguintes atributos:

1 - Atividade: atributo discreto com originalmente 392 diferentes atividades, as quais foram reduzidas para apenas 6 tipos diferentes;

2 - Classe: atributo discreto com 7 tipos diferentes [1 2 3 4 5 6 8];

3 - Tipo de ligação: [1 2 3 11 12 13 21 22 23 31 32 33 43] - retirado => [44 53 54 63 64];

4 - Consumo: atributo contínuo;

5 – Padrão de consumo: Foi divido em 3 classes: (0 - 100 kWh, 101 – 300 kWh, acima de 300 kWh);

Arquivos utilizados: normal_fraude_LR.mat e normal_fraude_LR_rs.mat

Arquivo de treino = 19627 consumidores

Arquivo de teste = 19592 consumidores

Nota: Neste caso foi realizado um pré-processamento no arquivo de treinamento. Existem alguns consumidores com perfis iguais e, alguns são classificados como fraudadores e outros como não fraudadores. Desta forma, o pré-tratamento faz uma análise no arquivo de treinamento. Quando o número de consumidores normais (não fraudadores) tiver o mesmo perfil de um consumidor fraudador e este número for, por exemplo, maior do que dez (o chamado “critério”.), estes consumidores normais passam a serem considerados fraudadores. Desta forma, foram considerados 5 casos, com o critério variando de 10, 20, 40, 60 e 100. O resultado pode ser visto na tabela 4.15.

Tabela 4.16: Análise com critério 10 a 100 – caso 09

Critério →	10	20	40	60	100
Total de Fraudes no arquivo de teste	905	905	905	905	905
Porcentagem do Total (19989)	4,62%	4,62%	4,62%	4,62%	4,62%
Total de Fraudes encontradas no arquivo de teste	755	1283	2301	3105	4418
Fraudes Erradas encontradas no arquivo de teste	588	1088	2062	2832	4089
Fraudes Corretas encontradas no arquivo de teste	167	195	239	273	329
Relação das *Corretas* em relação ao total encontrada	22,12%	15,20%	10,39%	8,79%	7,45%
Porcentagem de acertos nas Fraudes (relação a 905)	18,45%	21,55%	26,41%	30,7%	36,35%

Análise do resultado: O aumento do critério ocasionou o aumento no número total de fraudes identificadas, contudo o número de acerto nas inspeções diminuiu muito. A melhor relação se apresentou no critério (10). Quando o acerto nas inspeções foi de 22,12 % e foram identificados 18,45 % de todos os fraudadores.

Mostra-se na figura 4.12 uma visão gráfica da variação do critério e a resposta do sistema, retratando assim a relação entre os acertos nas inspeções e a identificação dos fraudadores com base na amostra total.

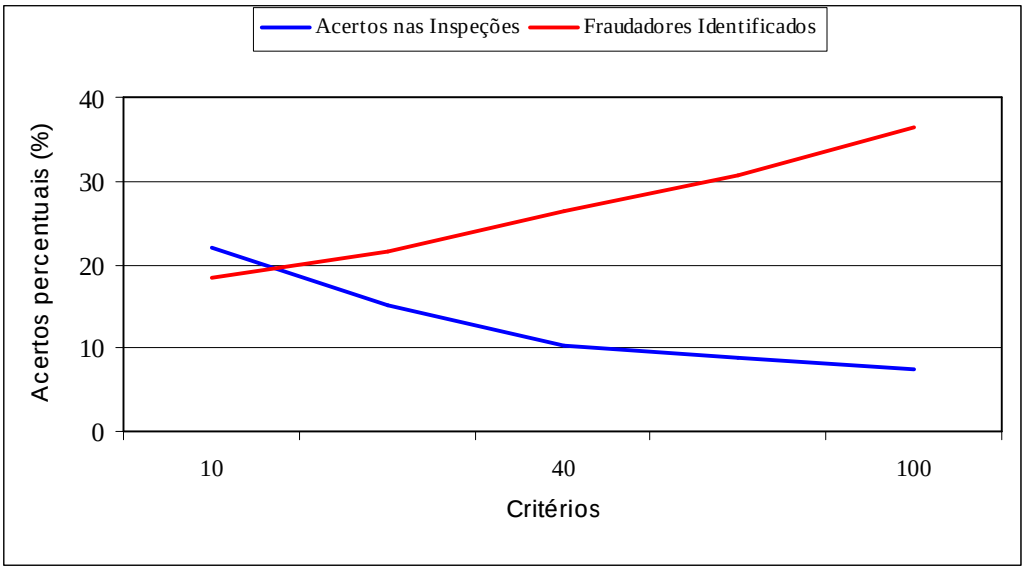

Figura 4.12 Resposta do sistema com variação dos critérios

Caso 10

Neste caso, o sistema é avaliado considerando os seguintes atributos:

1 - Atividade: atributo discreto com originalmente 392 diferentes atividades, as quais foram reduzidas para apenas 6 tipos diferentes;

2 - Classe: atributo discreto com 7 tipos diferentes [1 2 3 4 5 6 8];

3 - Tipo de ligação: [1 2 3 11 12 13 21 22 23 31 32 33 43] - retirado => [44 53 54 63 64];

4 - Consumo: atributo contínuo;

5 – Padrão de consumo: Foi divido em 3 classes: (0 - 100 kWh, 101 – 300 kWh, acima de 300 kWh);

Arquivos utilizados: normal_fraude_dados2.m

Nota: Neste caso foi utilizado o mesmo arquiva para treino e teste.

Arquivo de treino = Diversos

Arquivo de teste = 19592 consumidores

Nota: Neste caso foram realizados diversos pré-processamentos nos arquivos de treinamento, de forma que a relação Normal/Fraudador fosse variável, na forma (chamada de critério.): 1/1, 1/2, 1/3, 1/4, 1/5, 2/1, 2/2, 2/3, 2/4, 2/5, 3/1, 3/2, 3/3, 3/4, 3/5, 4/1, 4/2, 4/3, 4/4, 4/5, as tabelas 4.15 a 4.18 demonstram os resultados obtidos com a alteração do chamado critério onde buscou-se avaliar situações onde os dados tinham uma proporção conhecida entre o número de consumidores com fraudes e normais.

Tabela 4.17: Relação Normal/Fraudador – NF 1/1, 1/2, 1/3, 1/4, 1/5

Critério →	N/F 1/1	N/F 1/2	N/F 1/3	N/F 1/4	N/F 1/5
1 - Total Correta de Fraudes no teste	941	941	941	941	941
2 - Porcentagem do Total	4,71%	4,71%	4,71%	4,71%	4,71%
3 - Total de Fraudes encontradas no teste	7018	6864	7393	7444	7669
4 - Fraudes Erradas encontradas no teste	6423	6268	6758	6830	7054
5 - Fraudes Corretas encontradas no teste	595	596	635	614	615
6 - Relação entre item 5 e item 3	8,48%	8,68%	8,59%	8,25%	8,02%
7 - Relação entre item 5 e item 1	63,23%	63,34%	67,48%	65,25%	65,36%

Tabela 4.18: Relação Normal/Fraudador – NF 2/1, 2/2, 2/3, 2/4, 2/5

Critério →	N/F 2/1	N/F 2/2	N/F 2/3	N/F 2/4	N/F 2/5
1 - Total Correta de Fraudes no teste	941	941	941	941	941
2 - Porcentagem do Total	4,71%	4,71%	4,71%	4,71%	4,71%
3 - Total de Fraudes encontradas no teste	5375	4645	4592	5142	4796
4 - Fraudes Erradas encontradas no teste	4836	4123	4106	4621	4306
5 - Fraudes Corretas encontradas no teste	539	522	486	521	490
6 - Relação entre item 5 e item 3	10,03%	11,24%	10,58%	10,13%	10,22%
7 - Relação entre item 5 e item 1	57.28%	55,47%	51,65%	55,37%	52,07%

Tabela 4.19**:** Relação Normal/Fraudador – NF 3/1, 3/2, 3/3, 3/4, 3/5

Critério →	N/F 3/1	N/F 3/2	N/F 3/3	N/F 3/4	N/F 3/5
1 - Total Correta de Fraudes no teste	941	941	941	941	941
2 - Porcentagem do Total	4,71%	4,71%	4,71%	4,71%	4,71%
3 - Total de Fraudes encontradas no teste	3479	3749	3561	3900	3735
4 - Fraudes Erradas encontradas no teste	3046	3300	3123	3429	3294
5 - Fraudes Corretas encontradas no teste	433	449	438	471	441
6 - Relação entre item 5 e item 3	12,45%	11,98%	12,30%	12,08%	11,81%
7 - Relação entre item 5 e item 1	46,01%	47,72%	46,55%	50,05%	46,87%

Tabela 4.20: Relação Normal/Fraudador – NF 4/1, 4/2, 4/3, 4/4, 4/5

Critério →	N/F 4/1	N/F 4/2	N/F 4/3	N/F 4/4	N/F 4/5
1 - Total Correta de Fraudes no teste	941	941	941	941	941
2 - Porcentagem do Total	4,71%	4,71%	4,71%	4,71%	4,71%
3 - Total de Fraudes encontradas no teste	3044	2987	3250	3225	3022
4 - Fraudes Erradas encontradas no teste	2632	2582	2853	2807	2616
5 - Fraudes Corretas encontradas no teste	412	405	397	418	406
6 - Relação entre item 5 e item 3	13,53%	13,56%	12,22%	12,96%	13,43%
7 - Relação entre item 5 e item 1	43,78%	43,04%	42,19%	44,42%	41,5%

Análise do resultado: A alteração (1/1 até 4/5) do critério ocasionou o aumento no número total de fraudes identificadas, contudo o número de acerto nas inspeções diminuiu. A melhor relação se apresentou no critério (4/2). Quando o acerto nas inspeções foi de 13,56 % e foram identificados 43,04 % de todos os fraudadores.

4.4.5 Análise final dos casos simulados

Durante todos os casos simulados, ocorreu um compromisso entre o índice de acerto nas inspeções e o número total de fraudadores identificados. Para identificação do maior número possível de fraudadores, era necessário realizar o maior número possível de inspeções, o que conseqüentemente diminui o índice de acerto nas inspeções. O valor atual de acerto nas inspeções é de aproximadamente 5 %. Duplicar este valor e conseguir identificar mais de 40 % dos fraudadores é um resultado regular e foi o resultado obtido no último caso. Entretanto, possui o inconveniente de ter que fazer um pré-processamento dos dados de treinamento.

4.4.6 Padrões e Modelos – Avaliação

Nesta etapa são identificados os padrões que representam o conhecimento. Com base em medidas de interesse, pode-se dizer que as informações extraídas são expressas como padrões ou modelos. Se esses padrões são genéricos, então pode ser criado um modelo que é uma abstração do conjunto de dados original e é usado em tomadas de decisão, classificação ou predição. É desejável que tais técnicas que acham padrões apresentem-nos em formatos de fácil interpretação dos dados.

Neste passo são usadas ferramentas de visualização e técnicas de representação de conhecimento para apresentar ao usuário, o conhecimento gerado pelo minerador de forma a ser de fácil interpretação e utilização.

Os resultados obtidos através dos testes realizados em cada ciclo foram interpretados em cada etapa, na busca de melhorias.

4.5 Comentários finais

Neste capítulo foi apresentada detalhadamente uma metodologia para detecção de fraudes e ainda defeitos em medidores, pois para estas duas situações o perfil de comportamento da unidade consumidora é similar.

Foi utilizada a técnica de Árvore de Decisão, a partir do grupo de atributos selecionados, aplicou-se então a metodologia proposta.

Foram realizadas umas diversidades de estudos de casos, e simuladas várias combinações de atributos, redução de quantidade de atributos e ainda alterações de atributos do tipo contínuo para discreto e vice e versa.

A cada nova situação observou-se o desempenho do sistema através dos resultados obtidos e transcritos para a matriz de eficiência.

Atestada a eficiência da metodologia através de teste de confiabilidade, vários conjuntos de atributos foram avaliados na busca pelas informações mais relevantes para a descoberta de padrões de comportamento fraudulento ou ainda problemas de medição. Ao final, foram enunciados os atributos que beneficiaram e prejudicaram os resultados das medidas de avaliação consideradas.

No próximo capítulo são apresentadas as conclusões finais do trabalho, as contribuições alcançadas e os trabalhos futuros a serem realizados.

Capítulo V

Conclusões e Propostas de Trabalhos Futuros

5.1 Considerações finais

Este trabalho desenvolveu um sistema de auxílio à tarefa de detecção de fraudes e erros de medição em unidades consumidoras de baixa tensão. O sistema elaborado identifica consumidores potencialmente suspeitos que se enquadrem neste perfil e elabora uma lista para inspeção. O trabalho utiliza a técnica de Inteligência Artificial chamada Árvore de Decisão que é aplicada ao banco de dados da concessionária de distribuição de energia de elétrica.

O trabalho atendeu ao objetivo proposto, porém teve dificuldade de acesso direto aos dados de origem o que prejudicou a inclusão de novos atributos na pesquisa. Havia limitações de informações no banco de dados utilizado. Isto dificultou um trabalho direto na seleção de características que possuíssem maior poder discriminatório, impossibilitando a descoberta de novos conhecimentos. Desta maneira, a pesquisa se limitou a reproduzir o conhecimento dos especialistas no domínio.

Outra importante característica que influenciou nos resultados foi a qualidade dos dados. Por estarem no DWH, pressupõe-se que os dados tivessem um alto grau de limpeza e correção, contudo a realidade demonstrou uma grande diversidade de dados incompletos ou com inconsistências do ponto de vista prático da pesquisa.

Para exemplificar, podemos citar os casos em que encontramos valores de consumo negativo. A informação não está incorreta, mas o processo de inclusão deste dado no DWH e

o conceito existente por trás dele são pouco práticos para a pesquisa. Outro exemplo são os consumos com igual valor, e que correspondem ao consumo mínimo em determinada faixa. Estes valores são arbitrários e não demonstram o real consumo de energia. Tais fatos provocam distorções que dificultam a interpretação pelos algoritmos de mineração. Com relação à qualidade dos dados, também devemos citar a grande quantidade de dados nulos encontrados.

Muitas vezes as estratégias utilizadas para correção desta informação podem ter influenciado negativamente no desempenho dos classificadores. Podemos ponderar também a possibilidade da existência de valores nulos como sendo um tipo de informação importante e que foi, por algum motivo, perdida.

Consideramos para esta pesquisa apenas duas classificações possíveis: Não Fraudador (Normal) e Fraudador. Esta simplificação pode ter significado uma redução do poder discriminatório dos dados usados para treinamento e testes. Os resultados de inspeções que classificam as UC´s não são limitados a estas duas classes, mas a sete classes distintas: Normal, Fraude, Falha na Medição, Irregularidade Técnica, Irregularidade Comercial, Auto-religação e Impedimento.

Ainda com respeito à classificação de unidades consumidoras, devemos considerar também a ausência de campo de cadastro no sistema para indicação de suspeita de fraude a partir de inspeção em campo. Esta informação poderia demonstrar o comportamento real dos chamados falsos-normais.

Neste trabalho foi abordada a detecção de fraudes em unidades consumidoras de energia elétrica através da aplicação de uma metodologia baseada em conceitos de Árvore de Decisão.

O estudo aprofundado desta técnica de Inteligência Artificial permitiu compreender sua atuação em dados organizados em Sistemas de Informação ou Tabelas de Decisão. Ao aplicar o conceito de classificação nos dados de clientes consumidores de energia elétrica, foi

possível analisar o relacionamento entre os padrões de comportamento normais e fraudulentos.

A avaliação detalhada utilizando a classificação de cada atributo e ainda a variância entre padrões contínuos e discretos é o ponto principal da metodologia proposta. Esta avaliação permite gerar várias situações de classificação que levam a caracterização de uma unidade consumidora potencialmente fraudadora, cada qual focado em diferentes estimativas de taxa de acerto de inspeção e quantidade de fraudes detectadas.

Portanto, o resultado final determina um lote de inspeções a serem realizadas em campo, caracterizando unidades consumidoras com os seus perfis de comportamentos potencialmente fraudulentos.

Foram utilizados procedimentos adequados ao processo de aquisição do conhecimento e, destes experimentos, obtivemos resultados satisfatórios. Os resultados experimentais trariam uma melhora na identificação de suspeitos de fraudes, porém tais resultados necessitam de uma comprovação prática. Acreditamos, pelas razões já citadas, que os resultados tendem a melhorar, na medida em que novas possibilidades associadas a esta pesquisa forem incorporadas.

Embora este trabalho tenha abordado especificamente a detecção de fraudes em consumidores de energia elétrica e problemas em medidores de energia, a metodologia proposta pode ser entendida para a detecção de outros seguimentos de negócios.

Portanto, este trabalho representa uma importante contribuição, visto que as publicações na área de detecção de fraudes não detalham suas metodologias e resultados, prejudicando o aperfeiçoamento das técnicas e ferramentas contra fraudes.

Este trabalho enunciou em detalhes a fundamentação da teoria de Descoberta de conhecimento em Banco de Dados através da técnica de mineração de dados Árvore de Decisão, como também apresentou uma abordagem das perdas no setor elétrico com ênfase

no seguimento de distribuição. Por este motivo, o trabalho contribui como uma referência ou fonte de estudo na área de inteligência artificial para aplicabilidade em fraudes.

5.2 Trabalhos futuros

O presente trabalho suscitou a possibilidade de desenvolvimento de um Sistema Baseado em Conhecimento, cujas regras em muitos casos já foram levantadas através das várias entrevistas com especialistas do domínio. Certamente, um sistema baseado em conhecimento seria de grande valia também para a patrocinadora. Ainda dentro das modalidades desses sistemas, outra pesquisa derivada pode ser aprimorada para a Análise de Memória de Massa, cujo protótipo já foi desenvolvido. Trata-se da análise automatizada de informações de consumo provenientes de medidores especificamente instalados em clientes com medição em alta tensão para coleta de informações com intervalos de 5 minutos. O conjunto de informações geradas por este tipo de dispositivo é extremamente grande tanto do ponto de vista quantitativo quanto do qualitativo, e possibilita a análise de várias características presentes no perfil de consumo. Porém sua análise requer algum tipo de processamento automatizado, uma vez que se realizada de forma manual seria lenta e muito difícil.

Em continuação às propostas de trabalhos futuros, o uso de bases de dados específicas por tipo de atividade das unidades consumidoras poderia gerar classificadores específicos e com possibilidades de desempenho ainda superiores aos até agora encontrados. Estas bases de dados poderiam contribuir para a especialização de alguns classificadores em determinados tipos de consumidores de energia, e aumentar significativamente seu desempenho.

O uso de técnicas mais avançadas sobre séries temporais pode ser explorado e acreditamos que poderá agregar grandes avanços na pesquisa, incluindo a possibilidade de descoberta de novos conhecimentos, não apenas naqueles já formalizado pelos especialistas.

Neste sentido, a inclusão de novas características, que até o momento não foram utilizadas seria de grande valia para a tentativa de estabelecimento de relações entre estas características e a classificação do consumidor. Podemos citar, como exemplo, a utilização de características do medidor instalado no consumidor, bem como as características sócias econômicas. Desta forma, estas características usadas poderão proporcionar novos resultados e merecem uma investigação.

Acredita-se que existe a possibilidade de melhoria no sistema proposto, baseado em duas premissas:

- Melhoria no banco de dados: Foram identificados vários problemas nos dados existentes e falta de dados. O que compromete os resultados, considerando que a qualidade do banco de dados era baixa;

- Otimização do sistema proposto: Poderia ser feita uma reavaliação dos atributos utilizados.

5.3 Artigos Publicados

Artigos em Congressos Internacionais:

Artigo 1
Título: *Fraud Identification In Electricity Company Costumers Using Decision Tree*
Congresso : *IEEE - System, Man and Cybernetics Annual Conference*
Outubro/2004 - The Hague, Holanda

Artigo 2
Título: *Fraud Detection In Electrical Energy Consumers Using Rough Sets*
Congresso: *IEEE - System, Man and Cybernetics Annual Conference*
Outubro/2004 - The Hague, Holanda

Artigo 3

Título: *Rough Sets Based Detection in Eletrical Energy Consumers*

Congresso : World Engineering Academy and Socyet - WSEAS
Maio/2004 - Cancun – México

Revista: *WSEAS TRANSACTIONS*

Artigos em Congressos Nacionais:

Artigo 4 (Painel)

Título: Sistema de Identificação de Fraudes Utilizando Árvore de Decisão
SENDI 2004 – XVI Seminário Nacional de Distribuição de Energia Elétrica
Novembro/2004 – Brasília – Brasil

Artigo 5

Título: Sistema de Detecção de Fraudes em Consumidores de Energia Elétrica baseada em Rough Sets.
SENDI 2004 – XVI Seminário Nacional de Distribuição de Energia Elétrica
Novembro/2004 – Brasília – Brasil

Artigo 6

Título: Sistema de Detecção de Fraudes em Consumidores de Energia Elétrica baseada em Rough Sets.
CITENEL 2005 – III Congresso de Inovação Tecnológica em Energia Elétrica
Dezembro/2005 – Florianópolis - Brasil

Fraud Identification In Electricity Company Costumers Using Decision Tree *

José Reis Filho
Edgar M. Gontijo
ENERSUL S.A.
Campo Grande, MS, Brazil
jreis@enersul.com.br
emg@notes.escelsa.com.br

Antonio Carlos Delaiba
Faculdade de Engenharia Elétrica
Universidade Federal de Uberlândia
Uberlândia, MG, Brazil
delaiba@ufu.br

Evandro Mazina
José E. Cabral
João Onofre P. Pinto
Electrical Engineering Department
Federal University of Mato Grosso do Sul
Campo Grande, MS, Brazil
mazina@del.ufms.br
Jcabral@nin.ufms.br
jpinto@ieee.com

Abstract - *The objective of this work is to develop a system that pre-select electricity energy company costumers which will undergo in-site inspection for frauds or faulty measurement equipments identification. The pre-selection system was built based on the electricity company database. It was used attributes such as monthly energy consumption, type of consumers, previous inspection outcome, and others. A Decision Tree based classification system was used to reach such goal. The identification was designed, trained and tested using MATLAB code. The fraud/faulty equipments identification per number of in-site inspection rate was 40% of the total of pre-selected costumers, which was above the expectation.*

Keywords: Decision Tree, Fraud Detection, Data Mining, Knowledge Discovery in Database.

1 Introduction

Fraud is one of the main cause of revenue lost in many areas of business. Among those, credit card, cellular phone and insurance are the top ones. Therefore, a lot of research work have been done addressing fraud identification problem [1], [2], [3]. In order to solve this problem, hard computing and soft computing or computational intelligence have been used. The most used computational techniques are: Artificial Neural Networks, Fuzzy Logic Systems, and Decision Tree.

Decision Tree is a technique extensively explored for classifications problems [4], [5], [6], [7]. The advantages of this technique are: simple way to represent knowledge, easy to be built, training based in case, etc [8].

As in other business areas, electricity distribution companies, can also suffer fraud from their costumers. In Brazil, as in many other countries, the revenue losses of electricity companies due to frauds can go as high as 3%.

On the other hand, fraud identification via in-site inspection is costly. As an example, fraud identification per number of in-site inspection rate can be as low as 5%, which sometimes makes the inspection process unviable. The objective of this work is to built a classification system, based on Decision Tree, that can pre-select costumers that will undergo inspection. The goal is to reach a fraud identification per number of in-site inspection rate of 30%. This will reduce considerably inspection cost. This paper presents the complete knowledge based database process involved in this problem. First the pre-selector goal is established. In the sequence the attributes selection strategy is described. Next, the data pre-processing process is presented. Then, the development of the Decision Tree used by the pre-selector is addressed. Finally results are given and analyzed.

2 Basics on Decision Tree

Decision Trees are data based classification models. Basically, for this type of classification models, a set of structured examples with some non-categorical variables, the inputs, and one categorical variable or class, the

The project was supported by ENERSUL S.A. from Brazil through its research and development program.

* 0-7803-8566-7/04/$20.00 © 2004 IEEE.

output. The problem then is to find a model, decison tree, that can correctly classify to what category the non-categorical data presented with new values belongs to. The input variables may be continuos or discrete type of variables. On other hand, the output is discrete, and in general binary type, i. e., the output assumes values 1 or 0, meaning belongs or not belongs to a category. Table 1 gives a description of typical variables that are found in this type of problem.

Table 1 – Typical Variables in Decison Trees

Non-Categorical Attributes	**Variable Type**	**Values/Range**
Variable 1	Discrete	type 1, type 2, type 3
Variable 2	Discrete	0; 1
Variable 3	Continuous	[a b]
Variable 4	Continuous	[c d]
Categorical Attributes	**Variable Type**	**Values**
Output Variable	Discrete	0; 1

The examples used to build a decision tree, using the variables described in Table 1, are in general given as shown in Table 2. The examples in Table 2 would be the, or part of the, training data.

The decision tree training algorithm uses a set of training data, as given in Table 2 and creates a set of rules that express what is know about the problem. These rules can be graphically represented as given in Figure 1.

In Figure 1, the rectangles are the decision tree nodes. Each node, represents a non-categorical variable. The lines are the branchs, and each branch represents the value that the variable may assume. Finally, the circles are the leaves, and each leaf represents the expected classification of the categorical variable. Therefore, by following a path from the root, which is the top node, to a leaf, it is possible to create a rule. All the paths from the root to the leaves are the set of rules defined by the decison tree.

After the decision tree being trained, it is necessary to verify if it will correctly classify the new cases. So, in order to check the decision tree performance, a test data set has to be presented. The outcome of this test will then represent the evaluation of this classification model.

Table 2 – Example of a Simple Training Data Set

	Non-Categorical Variable/Input				Categorical Variable/Oupu t
Example	Var. 1	Var. 2	Var. 3	Var. 4	Class
1	1	0	25	350	0
2	1	1	20	400	0
3	2	0	23	280	1
4	3	0	10	460	1
5	3	0	08	300	1
6	3	1	05	200	0
7	2	1	04	150	1
8	1	0	12	450	0
9	1	0	09	200	1
10	3	0	15	300	1
11	1	1	15	200	1
12	2	1	12	400	1
13	2	0	21	250	1
14	3	1	11	300	0

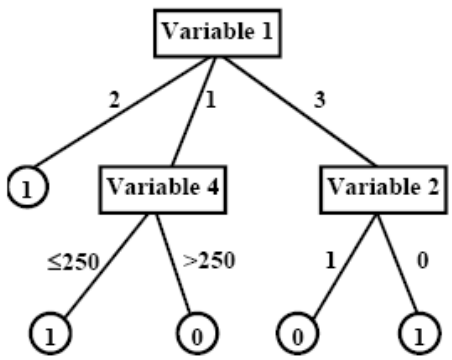

Figure 1. Graphical Representation of The Decision Tree

There are some different classification algorithms that allow to develop decision trees. Among these algorithms are: CART – Classification and Regression Trees [9], ID3

– Induction Decision Tree [10], C4.5 Continuous [11], and others. ID3 was one of the first decision tree algorithms and it was based in inference systems and learning systems. The algorithm C4.5 is an extension of the ID3, however it includes continuous variables. The performance of each algorithm depends on the addressed problem. The explanation of such algorithm is not in the scope of the work.

There are many softwares that implement decision trees, some require more knowledge of the user, some require less. The software used in this work was the Statistics Toolbox, which is part of MATLAB by Mathworks. The toolbox has functions to built and test decision trees.

3 The Pre-Selector

3.1 The Pre-Selector Goal

The objective of a pre-selector system is to built the matrix given in Table 3, in such way that it maximizes the diagonal entries, and minimizes the off-diagonal entries. In other words, the system must maximizes the number of right alarms ("frauder" or not "frauder" costumers as well as faulty or not faulty equipments). On the other hand, false alarms must be minimized, i.e., the costumer was selected as "frauder" or the equipment was identified as faulty when in fact neither the costumer was "frauder" or the equipment was faulty; or the costumer was selected as not "frauder" or the equipment was identified as not faulty when in fact the costumer was "frauder" or the equipment was faulty.

Table 3. The Pre-Selector Efficiency Matrix

System Outcome / Inspection Outcome	Fraud or Faulty Equipment (S)	Not Fraud or Not Faulty Equipment (NS)
Fraud or Faulty Equipment (F)	S/F	NS/F
Not Fraud or Not Faulty Equipment (NF)	S/NF	NS/NF

Since it is very difficult to find the entries of the last column, the efficiency of the pre-selector will be evaluated only based on the first column entries, which can be easily calculated after the inspection process.

3.2 Attributes Selection

This work is being done using the database of the local electricity company. The database is composed of a set of 52 variables or costumer attributes, however not all attributes were correlated to the desired output, i.e., selected or not selected for in-site inspection. Therefore, the first step was to select the attributes subset that was correlated to the output. Table 4 shows a subset sample of the set of 52 attributes existing in the company database. The selection of attributes database was based on company experts interviews and correlation function. The outcome of this process was the selection of five attributes. These attributes, as well as their features, are given in Table 5.

Table 4. Sample of the Existing Attributes in the Company Database

Name
Document
Address
Book
Class
Subclass
Fare
Type of measurement
Equipment number
Transformer point
Post
Load
Activity
Last inspection
Late payments
Number of disconnections
Date of the last disconnection
Last service #
Date of last service
Later inspections:
Result
Date
kWh recovered
$ recovered
regularization
Identification alterations
Average annual consumption

3.3 Data Pre-Processing

Data from a time period of one full year was used as the database. So, considering the universe of 600,000 costumer, 5 attributes, 12 months, the database was composed of a matrix of 3,000,000 x 12, which is

Table 5. Selected Attributes

Attributes	Variable Type
Consumer Class	Discrete
Consumer Activity	Discrete
Electrical Connection	Discrete
Monthly Consumption	Continuous
Local Average Consumption	Continuous

Table 6. Results for the Pre-Selector Efficiency Matrix

System Outcome / Inspection Outcome	Fraud or Faulty Equipment (S)	Not Fraud or Not Faulty Equipment (NS)
Fraud or Faulty Equipment (F)	**0.4**	**0.75**
Not Fraud or Not Faulty Equipment (NF)	**0.6**	**0.25**

somewhat large and too noisy to be used as a training data. Therefore the data has to be pre-processed. This stage did not use any special pre-processing technique, only trivial operations was done.

First, costumers that undergone inspection in the last 12 months were selected. Through this strategy, the number of costumers was reduced from around 600,000 to around 100,000. In the sequence, only costumers with inspection outcomes normal, fraud or faulty equipments were selected. This put the total of costumers down to around 50,000. Finally, all inconsistent data like negative monthly consumption was disregarded. The result was a database with 40,000 costumers, with 12 months historical data containing 5 attributes, 3 discrete and 2 continuous.

In order to build the decision tree, the reduced database was then divided into two subsets: training and testing subsets. A MATLAB code was then developed to, using the training dataset, train the decision tree for the pre-selector system.

4 Results

After the classification system had being trained, it was tested. The results showed 40% right classification rate, i.e., from every 100 consumers inspected, 40 was found to be "frauders". The problem with this result is that there still are a high number of clients that are "frauders" and were not pre-selected. These rate is of 25%, i.e., in the testing data set, from every 100 costumers with fraud, only 25 was inside of the pre-selected set for inspection. The problem with this system, as in most of cases, is the database, which was not composed for this particular use, and also, in many cases, does not correspond to reality. Therefore, it is difficult to the system learn the pattern of behavior of the people that commit frauds. Further work in the pre-processing stage is going on, aiming to get hid of the unreliable data. Table 6 shows the efficiency matrix of the pre-selector system.

5 Conclusions

This paper proposed a pre-selection system to improve in-site inspection performance for fraud detection based on decision tree. The system used the database from the electricity company. The performance of the main based system being used by the company today is as low as 5%. The proposed pre-selection system had a performance of 40%. Although the proposed system had such a good performance in making right selections, it left a large number of "frauders" outside of the set of costumers to be inspected. Further work is being done in the pre-processing stage in order to decrease the number of "frauders" outside of the inspection set..

References

[1] *Stefano, B.; Gisella, F;*. Insurance fraud evaluation: a fuzzy expert system, The 10th IEEE International Conference on Fuzzy Systems, 2001, Volume: 3 , 2-5 Dec. 2001, Pages:1491 – 1494

[2] *Deshmukh, A.; Talluru, T.L.N.;*A rule based fuzzy reasoning system for assessing the risk of management fraud, IEEE International Conference on Systems, Man, and Cybernetics, 12-15 Oct. 1997, Pages:669 - 673 v

[3] *Syeda, M.; Yan-Qing Zhang; Yi Pan;* Parallel granular neural networks for fast credit card fraud detection, Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, Volume: 1 , 12-17 May 2002. Pages:572 - 577

[4] *Wei Fan; Haixun Wang; Yu, P.S.; Stolfo, S.J.;* A fully distributed framework for cost-sensitive data mining - Proceedings. of 22nd International Conference on Distributed Computing Systems, 2002. , 2-5 July 2002

[5] *Srivastava, A.; Eui-Hong Sam Han; Singh, V.; Kumar, V.;* Parallel formulations of decision-tree

classification algorithms, Proceedings International Conference on Parallel Processing 1998, 10-14 Aug. 1998 Pages:237 – 244

[6] *Hambaba, M.L.;* Intelligent hybrid system for data mining, Proceedings of the IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering, 24-26 March 1996

[7] *Khoshgoftaar, T.M.; Seliya, N.; Yi Liu;* Genetic programming-based decision trees for software quality classification, 15th IEEE International Conference on Tools with Artificial Intelligence Proceedings., 3-5 Nov. 2003, Pages:374 – 383

[8] *Hongyan Liu; Jeffrey Xu Yu; Hongjun Lu; Jian Chen;* Unifying decision tree induction and association based classification, IEEE International Conference on Systems, Man and Cybernetics, 2002 Volume: 7 , 6-9 Oct. 2002

[9] L. Breiman, J. Friedman, R. Olshen, C. Stone, "Classification and Regression Trees", Wadsworth, Pacific Grove, CA, 1984.

[10] J.R. QUINLAN, "Induction of decision trees", Machine Learning, 1:81-106,1986.

[11] J.R. QUINLAN, "C4.5: Programs for Machine Learning", Morgan Kaufmann, 1993.

Fraud Detection in Electrical Energy Consumers Using Rough Sets*

José E. Cabral
João Onofre P. Pinto
Electrical Engineering Department
Federal University of Mato Grosso do Sul
Campo Grande, MS, Brazil
jcabral@nin.ufms.br
jpinto@ieee.com

Edgar M. Gontijo
José Reis Filho
ENERSUL S.A.
Campo Grande, MS, Brazil
emg@notes.escelsa.com.br
jreis@enersul.com.br

Abstract – *Rough sets is an emergent technique of Soft Computing that have been used in many knowledge discovery in database applications. This work describes an application of rough sets in the fraud detection of electrical energy consumers. From an information system, rough sets concept of reduct was used to reduce the number of conditional attributes and the minimal decision algorithm (MDA) was used to reduce some values of conditional attributes. The reduced information system derives a set of rules that reaches consumers behavior, allowing the classi cation rule system to predict many fraud consumers pro les. Rough sets proves that it is a powerful technique with application in many systems based in data.*

Keywords: Soft computing, knowledge discovery in database, rough sets, fraud detection.

1 Introduction

The recovery of pro t losses caused by frauds is important to keep the nancial balance of the energy companies. Often, inspections in electrical energy consumers are used to detect frauds. Due the very high number consumers, these inspections are made without prior behavior analysis, causing low rightness rate.

The energy companies have many information about electrical energy consumers stored in databases. These information can be used to identify pro les of fraud behavior, guiding which electrical energy consumers must be inspected. However, due the very large mass of data stored, it is necessary that the pro le identi cation process must be automatic.

Soft Computing techniques try to implement some human beings abilities in machines, such that it can realize some tasks as: decision problems, learning and recognition. Rough sets is one of these techniques, which supplies resources to manipulate uncertainty and imprecision in data, mainly in the knowledge discovery in databases (KDD). Many works applied Soft Computing techniques in fraud detection, mainly in credit cards [1], [2], [3], [4]. However, it was not found works using rough sets in the detection of any type of fraud. Similar work (based on the reduction of rules provided from database) apply the knowledge discovery from diagnostic cases of slope-failure danger [5].

The employed solution uses rough sets in KDD process to design an automatic classi cation system that stores many pro les of fraud behavior optimizing energy companies inspections.

Firstly, the most important concepts of rough sets theory are reviewed. Then is presented the application of rough sets at fraud detection in electrical energy consumers, showing step by step the KDD process. Finally, the application conclusions and some improvement proposals are presented.

2 Rough Sets Theory

The rough sets theory was proposed in 1982 by Zdzislaw Pawlak [6]. In 1991, Pawlak published the book "*Rough Sets: Theoretical Aspects of Reasoning about Data*" to consolidate his contribution [7]. Considering that real life is not accurate or precise (crisp), the datas that represent this universe can be indiscernible or uncertain (rough). Rough sets tries to pro le these uncertainties in data, aiming at the difculty to transform data to knowledge. It uses the indiscernibility relation between examples in databases, where this relation is associated to values of attributes that compose the database.

Rough sets have often been compared to fuzzy sets [8], sometimes considering that are competing models of imperfect knowledge. However this comparison is misfounded because indiscernibility and vagueness are distinct facets of imperfect knowledge [7].

2.1 Data Representation

Information about the real world can be organized in database, which was simpli ed in a table called information

*0-7803-8566-7/04/$20.00 © 2004 IEEE.

system [9]. Table 1 shows an example of information system.

Table 1: Information system of fraud in electrical energy consumers.

Consumer	Connection	Class	Average	Fraud
e1	1	1	Normal	No
e2	1	1	High	Yes
e3	1	1	Low	Yes
e4	2	1	Normal	No
e5	2	2	High	No
e6	2	1	Low	Yes

The rows in Table 1 represent the examples, objects, registries or cases (electrical energy consumers). The columns are the conditional attributes (Connection, Class, Average) and decision attributes (Fraud) to each example. Any system that is based in cases can be represented in an information system, where the rows content the examples and the columns content the attributes. Formally, the information system of Table 1 is composed for the following sets:

- U: the set of all examples, $U = \{e1, e2, e3, e4, e5, e6\}$.
- A: the set of all conditional attributes, $A = \{Connection, Class, Average\}$.
- D: the set of all decision attributes, $D = \{Fraud\}$. The conditional attribute can be represented by d.

Among the de nitions contained in rough sets theory, there are two that can directly be applied to information systems. They are: reducts and concepts.

2.2 Reducts

Considering the set A of the Table 1, all the elements belonged to U are distinct. Either, considering the attributes $Connection$, $Class$ and $Average$, the set U is partitioned in the elementary subsets $\{e1\}$, $\{e2\}$, $\{e3\}$, $\{e4\}$, $\{e5\}$ and $\{e6\}$. Now, considering the subset $\{Connection, Class\}$ of A, the set U is partitioned in the not-elementary subsets $\{e1, e2, e3\}$, $\{e4, e6\}$ and $\{e5\}$. Due to this fact, only the attributes $Connection$ and $Class$ cannot discern all examples of the Table 1. However, the subset $\{Connection, Average\}$ can divide the set U in elementary subsets. Only the attributes $Connection$ and $Average$ can distinguish all examples of Table 1. Then, it is concluded that the attribute $Class$ is *redundant*. The set $P = \{Connection, Average\}$ does not contain redundant attributes and it is called *reduct* of set A.

Formally, a set of attributes P is reduct of A if $P \subseteq A$ keeps the indiscernibility relations of A. In other words, if P has cardinality less or equal than A and can represent all examples of an information system, then P is a reduct of A. Considering the reduct $P = \{Connection, Average\}$ of $A = \{Connection, Class, Average\}$, a new information system is shown in Table 2.

Table 2: Information system reduced by the reduct.

Consumer	Connection	Average	Fraud
e1	1	Normal	No
e2	1	High	Yes
e3	1	Low	Yes
e4	2	Normal	No
e5	2	High	No
e6	2	Low	Yes

Although Table 2 shows a reduct to the information system of Table 1, it is not necessarily unique. It may exist more than one reduct to a set of attributes [7]. Due to the minimization of attributes through reducts, there is reduction of data without lost of knowledge. This reduction is more relevant where the information system has many conditional attributes. To nd the reducts is one of the bottlenecks of the rough sets methodology [10]. However, heuristics based on genetic algorithms compute many reducts in often acceptable time [10].

2.3 Concepts

Beyond the analyzed conditional attributes for the search of reducts in information systems, decision attributes are also important to rough sets theory. Considering the set of decision attributes D of Table 2, $\{Fraud\}$, it divides the set U in two subsets: $\{e1, e4, e5\}$ and $\{e2, e3, e6\}$. Each subset is called *concept*. The rst concept represents the not fraudulent electrical energy consumers, when the second contents the examples with fraud. The concepts determine the classes to which examples belong. It is possible to determine the concept of an example from its conditional attributes. Considering Table 2, a set of *classi cation rules* can be derived:

1. $Average = Normal \rightarrow Fraud = No$
2. $Connection = 2 \wedge Average = High \rightarrow Fraud = No$
3. $Connection = 1 \wedge Average = High \rightarrow Fraud = Yes$
4. $Average = Low \rightarrow Fraud = Yes$

Each row of Table 2 derives one distinct rule. The rules derived from examples $e1$ and $e4$ were simpli ed and reduced to rule 1. The same happens with $e3$ and $e6$ to reach rule 4. The derived rules can classify all the examples, however, this directly classi cation method cannot be applied in all information system. To show a problematic situation, consider the Table 3, that is a copy of Table 2 plus two new examples ($e7$ and $e8$).

The concepts of Table 3 are $\{e1, e4, e5, e8\}$ and $\{e2, e3, e6, e7\}$. However, the examples $e5$ and $e7$ are in

different classes, although they content the same values of conditional attributes. The same happens with examples $e6$ and $e8$. These inconsistence makes impossible the generation of two rules:

1. $Connection = 2 \wedge Average = High \rightarrow Fraud = ?$

2. $Connection = 2 \wedge Average = Low \rightarrow Fraud = ?$

To pro le these problems, rough sets theory de nes three subsets of U.

2.4 Lower Approximation, Upper Approximation and Boundary Region

Let X be a concept of an information system. It can be found a subset of X with examples that *certainly* are members of X. This subset is called *lower approximation* of X, or simply $\underline{X}$. Considering Table 3, if $X = \{e1, e4, e5, e8\}$, then $\underline{X} = \{e1, e4\}$. Similarly, if $X = \{e2, e3, e6, e7\}$, then $\underline{X} = \{e2, e3\}$. Note that always $\underline{X} \subseteq X$.

The *upper approximation* of X, or simply $\overline{X}$, corresponds to a subset of U with examples that *can* be contained in a concept X. Considering Table 3, if $X = \{e1, e4, e5, e8\}$, then $\overline{X} = \{e1, e4, e5, e6, e7, e8\}$. Similarly, if $X = \{e2, e3, e6, e7\}$, then $\overline{X} = \{e2, e3, e5, e6, e7, e8\}$. It notes that always $X \subseteq \overline{X}$.

The *boundary region* of X, or simply BX, corresponds to a subset of U with examples that belong to $\overline{X}$, but do not belong to $\underline{X}$, i.e., $BX = \overline{X} - \underline{X}$. If $BX = \emptyset$, then $\overline{X}$ and $\underline{X}$ are the same sets. In other words, the information system does not contain inconsistent examples. Consequently, the bigger the cardinality of BX, the greater is the indiscernibility between the concepts.

If the examples of an information system belong to the sets $\underline{X}$ and $\overline{X}$, their organization are in accordance with its relevancies to the concept. If it is desired to nd the fraudulent examples, it is enough to determine $\underline{X}$. When the certainty is not necessary and it is desired to determine the possible fraudulent examples, it determines $\overline{X}$. The Figure 1 represents the approximation sets of Table 3.

Table 3: Inconsistent information system.

Consumer	Connection	Average	Fraud
e1	1	Normal	No
e2	1	High	Yes
e3	1	Low	Yes
e4	2	Normal	No
e5	2	High	No
e6	2	Low	Yes
e7	2	High	Yes
e8	2	Low	No

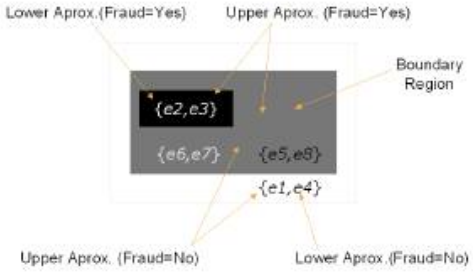

Figure 1: Lower approximation, upper approximation and boundary region.

3 Application

This section shows the KDD process applied to an electrical energy company database. The Figure 2 illustrates each step of the KDD process [11].

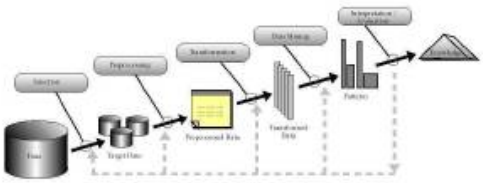

Figure 2: The KDD process.

3.1 Problem De nition

It was desired to discover patterns (or pro les) of fraudulent consumers in the direction of better guiding the inspections. The electrical energy company realizes about 1,000,000 inspections per year with a rightness rate of 0,1 (10%). It was desired to reach a rightness rate of 0.3 (30%).

3.2 Selection of Relevant Attributes

First, all the conditional attributes were listed. Then, interviews and quarrels with specialists of the electrical energy company was done with the intention to understand the information of the attribute contains. The result of these meetings was the election of 10 attributes (statics and dynamics).

3.3 Data Preprocessing

The biggest part of the process time was used in this step. The full database had about 600,000 consumers until 2003 and about 6,000,000 registers. Only considering almost the 100,000 consumers inspected in 2003, it was selected around 1,000,000 registers (each month sampled to January until December). Each consumer were classi ed by the attribute $Outcome$ that corresponded to the inspection result. The

possible values to this attribute was: normal, fraud, measurement fault, auto-reconnection, impediment, comercial irregularity, technical irregularity and null. Consumers with values $Normal$ and $Fraud$ was accepted. After this consideration, the number of consumers fell to around 40,600, with about 3,900 (9.6%) classi ed as defrauding and 36,700 (91.4%) as normal.

Since rough sets theory is only applied in information systems with values of attributes that are categorical, the continuous attributes were converted to discrete. Attributes with categorical values, however with unappropriate types (as string), were converted to integer. Then, all the values of attributes were composed of integer numbers.

3.4 Data Transformation

Some attributes were created from others as the $Average$ that is based on the monthly consume of each electrical energy consumer.

3.5 Data Mining

In this step, the database was ready for data mining using rough sets algorithms. Of the 40,600 registers, 20,300 were random selected to create the training set and the 20,300 remain to form the testing set. As only the training set was submitted to data mining, it was considered an information system. The rst employed algorithm was the one that computed the reduct $P \subseteq A$. The reached reduct kept 8 of 10 conditional attributes selected *a priori*. With the columns reduction of the information system, it became less complex. However, this "horizontal reduction" generated indiscernibility and similar examples.

The main objective of this work was to detect the pro les of fraudulent electrical energy consumers. Consequently, the concept X started to represent the examples with decision attribute $Fraud = Yes$. Considering that $\underline{X}$ determines the elements that with *certainty* are in the concept X, examples of the set $\underline{X}$ were selected and the examples of the set $U - \underline{X}$ was discarded. The set $\underline{X}$ had 630 examples.

After nding $\underline{X}$, each example of this set would generate a classi cation rule. All the rules would indicate the pro les of fraudulent units consumers. However, before generating these rules, the MDA (*Minimal Decision Algorithm*) [5] was applied. This algorithm compares each rule, one by one, to detect values of conditional attributes that can be rejected without the rule losing classi cation ef cienc y. This job reduced the number of conditional rule attributes and, consequently, the amount of comparisons when they are going to be tested. The set of nal simpli ed rules totalized 425 rules, therefore 205 of the 630 initial rules had become similar after the simpli cation.

3.6 Test

The set of rules was decently sorted by the number of rule attributes. Each example of the testing set was compared with the set of rules. When the values of the rule attributes were equal to the values of the example attributes, the example was marked as correctly classi ed and a rightness rate of the rule was updated. At the end of the test stage, each example of the set of rules had a ag symbolizing if it was classi ed correctly or not by the rst rule with similar attribute values. In the same way, each rule had a rightness rate. The total rightness rate was the number of examples with a ag symbolizing rightness divided by the total number of examples. It was reached a total rightness rate of 20%, considered good, however below the nal objective. Rules with low rightness rate represented noises of the information system, and was discarded. Rules with high rate was selected as good generalizations of fraudulent pro les.

4 Conclusions

Rough sets proves that it is a powerful technique with application in fraud detection. The concepts of rough sets (mainly lower approximation, reduct and MDA) was used to reach a reduced information system, which is a classi cation rule system. This system could predict fraud consumers pro les with rightness rate of 20%.

The main dif culty to detect fraudulent electrical energy pro les is the reason between normal and fraudulent examples (1/10). To aggravate this disadvantage, many fraudulent consumers behavior seems like normal behavior. This fact makes that many fraudulent examples belong to the lower approximation. Further work is being done in the data preprocessing stage in order to increase the number of fraudulent examples.

References

[1] M. Syeda, Yan-Qing Zhang and Yi Pan, "Parallel granular neural networks for fast credit card fraud detection", Proc. of IEEE Int. Conf. on Fuzzy Systems 2002, Vol.1, pp.572–577, May 2002.

[2] S. Ghosh and D.L. Reilly, "Credit Card Fraud Detection with a Neural-Network", Proc. of 27th Annual Hawaii Int. Conf. on System Science, IEEE Comp. Soc. Press, Vol.3, pp.621–630, 1994.

[3] P. K. Chan, W. Fan, A. L. Prodromidis and S. J. Stolfo, "Distributed data mining in credit card fraud detection", Proc. of IEEE Intelligent Systems on Data Mining, Vol.14, pp.67–74, December 1999.

[4] E. Aleskerov, B. Freisleben and B. Rao, "CARDWATCH: A Neural Network based Database Mining System for Credit Card Fraud Detection", Proc. of IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, pp.220–226, 1997.

[5] L. Polkowski, J. Kacprzyk and A. Skowron, *Rough Sets in Knowledge Discovery 2: Applications, Case Studies, and Software Systems*, Physica-Verlag, 1998.

[6] Z. Pawlak, "Rough Sets", *International Journal of Computer and Information Sciences*, Vol.11, pp.341–356, 1982.

[7] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Dordrecht, 1991.

[8] L. A. Zadeh, "Fuzzy Sets", *Information and Control*, Vol.8, pp.338–353, 1965.

[9] Z. Pawlak, J. Grzymala-Busse, R. Slowinski and W. Ziarko, "Rough Sets", *Communications of the ACM*, Vol.38, No.11, pp.89–95, November 1995.

[10] S. K. Pal and A. Skowron, *Rough-fuzzy hybridization: a new trend in decision-marking*, Springer-Verlag, Singapore, 1999.

[11] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery in databases", *AI Magazine*, Vol.17, pp.37–54, 1996.

Rough Sets Based Fraud Detection in Electrical Energy Consumers

José E. Cabral Junior, João Onofre P. Pinto, Edgar M. Gontijo, José Reis Filho

Electrical Engineering Department
Federal University of Mato Grosso do Sul
Cidade Universitária, s/n – CP 549
Campo Grande, MS - 79070-900
BRAZIL

ENERSUL S.A.
Av. Gury Marques s/n
Bairro Santa Felicidade
Campo Grande, MS - 79072-900
BRAZIL

jcabraljr@hotmail.com, jpinto@del.ufms.br, emg@notes.escelsa.com.br, jreis@enersul.com.br

Abstract: - This article describes the theory and application of Rough Sets of fraud detection in electrical energy consumers from databases. The Rough Sets concept of reduct was used to remove conditional attributes and the minimal decision algorithm (MDA) was used to remove insignificant classes of each conditional attribute. The minimized database approach the consumers behavior, allowing a classification rule system to predict fraud consumers profiles. The achieved results are good enough to demonstrate that Rough Sets is a very powerful technique for this type of problem.

Key-Words: - Rough sets, Fraud Detection, Electrical Energy Consumers, Reduct, Minimal Decision Algorithm (MDA), Classification Rules.

1 Introduction

Intelligent Fraud Detection Systems have been intensively addressed in recent past. So far, mostly of the work has been done for credit card and cell phone fraud detection. The most popular soft computing techniques used for the purpose are artificial neural networks [1],[2], and fuzzy logic [3],[4].

However, fraud detection of electrical energy consumers has barely been reported in literature. In general, this problem is solved by in-site inspection. Most of the time, as reported by some electricity companies, the fraud identification rate of this strategy is 5% or below. This because the decision of who has to be inspected is done by a worker, who although is a specialist, cannot efficiently look into all the data available from all company consumers and make a decision. The result is a very expensive process that sometimes does not results in cost reduction for the company. Rough Sets is a soft computing technique that, recently, is being widely applied to Knowledge Data Discovery (KDD) problems. For instance, Rough Sets was shown to be effective in classification rules determination [5]. However, for the best knowledge of the authors, it has never been applied for any type of fraud detection.

Initially, a brief description of Rough Sets theory is made, approaching the main concepts. In the sequence, the solution in the detection of frauds from databases is presented and finally the results of the gotten system are given.

2 Rough Sets Theory

Rough Sets theory was developed by Zdzislaw Pawlak in early 1980's[6]. It deals with the classificatory analysis of data tables (or databases). The main goal of Rough Sets analysis is to synthesize approximation of concepts from the acquired data. Often this concepts are "rough" or "fuzzy", and consequently, some methods or algorithms are necessary to reach them. This justify the applicability of Rough Sets in knowledge discovery in databases. Some concepts of Rough Sets will be presented on next subsections.

2.1 Information and Decision Systems

A data set is represented as a table. The rows represent the objects (examples, cases), and each column an attribute (variable, property). This table is called an information system [6]. Formally, it is a pair $\mathcal{A}$=(U,A), where U is a non-empty finite set of objects, and A is a non-empty finite set of attributes. Often, an information system has one (or more) special attribute representing a decision or an

The project was supported by ENERSUL S.A. from Brazil through its research and development program.

outcome. It is called decision attribute. The information system plus the decision attribute defines a decision system. Formally, $A=(U, A \cup \{d\})$, where $d \notin A$ is a decision attribute. The attributes in the set A are called conditional attributes. An information system and its decision system are showed in Table 1.

Cand.	Diploma	Experience	Decision
x1	MSc	Medium	Accept
x2	MBA	Medium	Accept
x3	MCE	Low	Reject
x4	MBA	Medium	Reject
x5	MBA	High	Accept
x6	MSc	Medium	Reject
x7	MSc	High	Accept
x8	MCE	Low	Reject

Table 1- Information system (gray) and decision system (all the table).

2.2 Indiscernibility in Objects

Some objects in the Table 1 are indiscernible. For example, considering the attributes Diploma and Experience, the objects in each subset {x1,x6}, {x2,x4} and {x3,x8} are indiscernible. Considering only the attribute Experience, {x3,x8}, {x1,x2,x4,x6} and {x5,x7} are indiscernible. The indiscernibility relation is an equivalence relation. For more details, see [6].

2.3 Set Approximation

Analyzing the decision attributes in a decision system, the class set is found. It is just the set of decision values. For the decision system of Table 1, the class set is {Accept, Reject}.

As it can be observed in Table 1, some objects can represent conflicting information. For example, the objects x1 and x6 possess the same values of conditional attributes, however different values in the decision attribute. To deal with this kind of problem, Rough Sets theory defines yours sets approximation. Either $X \subseteq U$ the set of objects with one determined class, is defined:

- Lower approximation ($\underline{X}$): set of all objects of class X that are not indiscernible with none another object. For the decision system of Table 1, the lower approximation for the class "Accept" is {x5, x7}, and for the class "Reject" is {x3, x8}. Maybe $\underline{X}$ has less elements than X, due to elimination of the indiscernible elements in X to reach $\underline{X}$;
- Upper approximation ($\overline{X}$): set of all objects of class X plus the objects of others classes that are indiscernible with some object of class X. For the decision system of Table 1, the upper approximation for the class "Accept" is {x1, x2, x4, x5, x6, x7} and for the class "Reject" is {x1, x2, x3, x4, x6, x8}. Maybe $\overline{X}$ has more elements than X, due to addiction of some elements to reach $\overline{X}$;
- Boundary region (BnX): set of all objects of class X that are indiscernible. For the decision system of Table 1, the boundary region between the classes "Accept" and "Reject" is {x1, x2, x4, x6}.

The lower and upper approximations, together with boundary region, define the regions for the classes. These regions inform how much an object can be said to be inside a class or not. Fig. 1 illustrates the distribution of the objects inside the regions. The dark blue region delimits all the candidates (objects) that for certain are classified as Accept (a crisp set). The white region delimits all the candidates that positively are classified as Reject (a crisp set). Already the blue region (between dark blue and white) defines the candidates that can be classified as Accept or Reject (a rough set). In other words, the darker the blue, greater is the acceptance certainty.

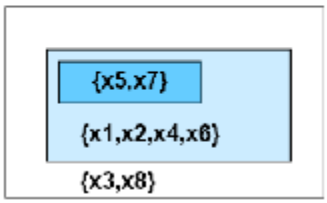

Fig. 1 – Regions for the classes.

It can be interesting to know how much a class is represented or not in a decision system. For such, the accuracy of approximation is defined as:

$$\alpha(X) = |\underline{X}| / |\overline{X}| \qquad (1)$$

If $\alpha(X)=1$, X is crisp (X is precise), and otherwise, if $\alpha(X)<1$, X is "rough" (X is vague).

The project was supported by ENERSUL S.A. from Brazil through its research and development program.

Candidate	Diploma	Experience	French	Reference	Decision
x1	MBA	Medium	Yes	Excellent	Accept
x2	MBA	Low	Yes	Neutral	Reject
x3	MCE	Low	Yes	Good	Reject
x4	MSc	High	Yes	Neutral	Accept
x5	MSc	Medium	Yes	Neutral	Reject
x6	MSc	High	Yes	Excellent	Accept
x7	MBA	High	No	Good	Accept
x8	MCE	Low	No	Excellent	Reject

Table 2 – Decision System

2.4 Reduct

Given the set of attributes of the decision system defined by Table 2, the reduct of this system can be found. This task consists of eliminating the linear dependent attributes. Or either, to eliminate conditional attributes that do not add any real information to the object. Finding a minimal reduct, a discernibility matrix must be created [6]. This matrix compares each object, identifying in each comparison which attributes possess different values. Later, a discernibility function is applied to the matrix and the linear dependent attributes (and the reduct) are found. To find this minimal reduct from the discernibility matrix is NP-hard. Fortunately, there exist some heuristics that find reducts with a viable computational cost. Although they do not guarantee that the reduct is minimal, the heuristics are more used. The software Rosetta [7] implements some of these heuristics.

2.5 Minimal Decision Algorithm

The minimal decision algorithm (MDA) [5] is used for reductions in decision systems or rule bases. MDA compares the attribute values of an object with the others objects. If it finds attribute values that can be eliminated without the objects becoming indiscernible, the MDA removes this attribute value from the object. Considering object x1 of Table 2. If its first attribute value is eliminated, x1 continues different of all objects. The second attribute value can be eliminated in the same way. Already, the third attribute value of x1 cannot be eliminated because, in case it was eliminated, the object x1 become indiscernible with the object x8.

3 Problem Solution

3.1 Aplication

Rough Sets theory addresses the analysis of tables (database) aiming to approximate concepts and information from these repositories. Often, this information is imprecise and/or has uncertainty, and it needs algorithm or special methodology to determine it.

At first, to solve the fraud detection of electrical energy consumers problem, costumers data was divided into training data and testing data. This is a standard procedure for supervised learning. In the sequence, the repeated registers were eliminated, and for the training data, only the distinct registers remained. Then, rough sets concepts were used. The lower approximation for the concepts (normal and fraud) was found and the registers that did not belong to this subset were eliminated. After this step, a valid reduct was determined, and the linear dependent attributes were eliminated. The elimination of some attributes reduced the dimension of the training data. Again, since after some attributes elimination some registers become repeated. The repeated ones were eliminated. Then, the Minimal Decision Algorithm (MDA) was applied to the reduced training data. This algorithm was able to significantly reduce the training data. As before, after the application of the MDA, some registers became repeated, and they were eliminated. Finally, for each remaining register a classification rule was derived. The whole set of rules is called classification rules system. The classification rules system can then be tested using the testing data.

3.2 Results

The database had about 100,000 registers, considering only the inspected units. After filtering inconsistent and irrelevant data, the number of registers fell down to 40,000,

The project was supported by ENERSUL S.A. from Brazil through its research and development program

which 90% was classified as normal and 10% was classified as fraud. The database was equally divided in training data and testing data. Following the steps explained in aplication subsection, the training data was reduced

from 20000 to 1980 registers. The remaining registers resulted in sparse rules, i.e., not all attributes were used to all rules. This makes the classification rules system not so computation intensive. The application of the classification rules system to the testing data resulted in 20% right classification. This is a very promising result, since the ultimate goal is to reach 30% right classification.

4 Conclusion

- Rough Sets is a powerful tool for fraud detection, mainly when does not exist any previous knowledge of the system, but the database;

- Although it is a computation intensive tool, the rough sets algorithms are easy to understand and to implement;

- An hybrid system involving Rough and Fuzzy sets seems to be a good approach for fraud detection problems;

- The obtained system reaches 20% true classification, but further work is being done in order to reach 30%;

- The main problem to reach the percentage of true classification, as in most data mining cases, was the quality of the data, which in many register did not correspond to reality;

- The upper approximation for the concepts (normal and fraud) has been studied to reach a better goal or the ultimate goal of 30% right classification.

References:

[1]R.Brause, T. Langsdorf, M.Hepp, Neural data mining for card fraud detection. *Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference*,1999, pp.103-106.

[2]Garavaglia, S, An application of a counter-propagation neural network: simulating the Standard and Poor's Corporate Bond Rating system. *Artificial Intelligence on Wall Street, Proceedings, First International Conference*, 1991, pp. 278-287.

[3] Deshmukh, A.; Talluru, A rule based fuzzy reasoning system for assessing the risk of management fraud, *Systems, Man, and Cybernetics, 1997, Computational Cybernetics and Simulation, 1997 IEEE International Conference*, Vol. 1, 1997, pp. 669-673.

[4] Syeda, M., Yan-Qing Zhang, Yi Pan, Parallel granular neural networks for fast credit card fraud detection, *Proceedings of the 2002 IEEE International Conference*, Vol. 1, 2002, pp. 572-577.

[5] H. Furuta, M. Hirokane, Y. Mikumo, Extraction Method Based on Rough Sets Theory of Rule-Type Knowledge from Diagnostic Cases of Slope-Failure Danger Levels, *Rough Sets in Knowledge Discovery 2 Application, Case Studies and Software Systems*, Part 1, Chapter 10, pp. 178-192.

[6] Pawlak, Z., Rough Sets, *International Journal of Computer and Information Sciences*, 1982, pp. 341-356.

[7] Ohrn, A. Rosetta: Technical reference manual. Technical report, Knowledge System Group, Norwegian University on Science and Technology, NO. http://rosetta.lcb.uu.se/general/

The project was supported by ENERSUL S.A. from Brazil, through its research and development program.

SISTEMA DE IDENTIFICAÇÃO DE FRAUDES UTILIZANDO ÁRVORE DE DECISÃO.

José Reis Filho
Edgar M. Contijo
ENERSUL S. A.
Campo Grande, MS, Brasil
jreis@enersul.com.br
emg@notes.escelsa.com.br

Antonio Carlos Delaiba
Faculdade de Engenharia Elétrica
Universidade Federal de Uberlândia
Uberlândia, MG, Brasil
delaiba@ufu.br

Evandro Mazina
José E. Cabral
João Onofre P. Pinto
Electrical Engineering Department
Federal University of Mato Grosso do Sul
Campo Grande, MS, Brasil
mazina@del.ufms.br
jcabral@nin.ufms.br
jpinto@ieee.com

OBJETIVO:

Desenvolver um sistema que pré-selecione os consumidores de uma concessionária de energia elétrica que deverão ser inspecionados para identificação de fraudes ou erros de medição por parte dos equipamentos.

1 - JUSTIFICATIVA

Coibir a fraude, uma das principais causas de perda de receita.

Reduzir as inspeções em campo devido ao seu alto custo operacional.

Melhorar a relação número de inspeções por fraudes ou erro de medição.

2 - A ÁRVORE DE DECISÃO

São modelos de classificação baseados em dados.

O problema: encontrar um modelo, uma árvore de decisão, que possa classificar corretamente a categoria de um conjunto valores das variáveis de entrada.

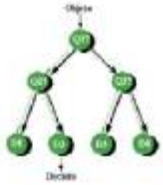

Tabela 1 - Variáveis Típicas em Árvores de Decisão.

Atributos Não Categóricos	Tipo de Variável	Valores/Intervalo
Var vel	Discreta	tipo 1, tipo 2, tipo 3
Var vel	Discreta	0; 1
Var vel	Contínua	[a b]
Var vel	Contínua	[c d]
Atributos Categóricos	**Tipo de Variável**	**Valores**
Variável de Saída	Discreta	0; 1

O significado de treinamento da árvore de decisão usa um conjunto de dados que são mostrados na tabela 2 e cria um conjunto de regras que expressam o que se sabe sobre o problema. Estas regras são mostradas graficamente na figura 1.

Tabela 2 - Conjunto de Dados para Treinamento.

	Variáveis Não Categóricas / Saída				Variável Categórica Saída da
Exemplo	Var. 1	Var. 2	Var. 3	Var. 4	Decisão
1	1	0	25	350	0
2	1	1	20	400	0
3	2	0	23	280	1
4	3	0	10	460	1
5	3	0	08	300	1
6	3	1	05	200	0
7	2	1	04	150	1
8	1	0	12	450	0
9	1	0	09	200	1
10	3	0	15	300	1
11	1	1	15	200	1
12	2	1	12	400	1
13	2	0	21	250	1
14	3	1	11	300	0

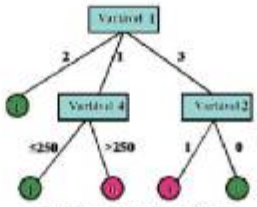

Figura 1 - Representação Gráfica da Árvore de Decisão

3 - O PRÉ-SELETOR

3.1 - Meta do Pré-Seletor

Maximizar os elementos da diagonal principal e minimizar os de elementos fora da diagonal principal.

Tabela 3 - Matriz de Efeitos do Pré-seletor

Fraude no Poste no Equipamento (FE)	[illegible]	[illegible]
Erro Medição no Sistema Medição no Equipamento (EM)	[illegible]	[illegible]

3.2 - Atributos de Seleção

O trabalho começou com uma base de dados da concessionária local composta de 52 atributos. Após a análise dos subconjuntos de atributos relacionados com a saída desejada, o número de atributos foi reduzido para cinco conforme tabela 4.

Tabela 4 - Atributos Selecionados.

Atributos	Tipo de Variável
Classe do Consumidor	Discreta
Atividade do Consumidor	Discreta
Consumo Médio	Discreta
Consumo Mensal	Contínua
Consumo Mínimo Mensal	Contínua

Uma base de dados de um ano considerando 12 meses com um universo de 600.000 clientes foi usada, composta uma matriz 5.000.000x12. Neste artigo apenas operações envolvendo faturamento foram feitas.
Após uma seleção criteriosa dos clientes, reduziu-se o total de clientes a 40.000, 12 meses de histórico e 5 atributos.
Para construção da árvore de decisão a base de dados reduzida foi dividida em dois sub-conjuntos: treinamento e teste.

4 - RESULTADOS

Os resultados mostraram uma taxa de 40 % de acerto, isto é, de 100 clientes inspecionados 40 foram classificados como fraudadores. Porém ainda permanece um número de clientes que são fraudadores não-flagrados-selecionados.
A tabela 5 mostra a acuracia de eficiência do sistema pré-seletor.

Tabela - 10 Resultados da Matriz de Efeitos do Pré-seletor.

Fraude no Poste no Equipamento (FE)	[illegible]	[illegible]
Erro Medição no Sistema Medição no Equipamento (EM)	[illegible]	[illegible]

5 - CONCLUSÃO

Foi proposto um sistema de pré-seleção baseado em árvore de decisão, visando melhorar a performance na detecção de fraudes. O desempenho do sistema usado hoje pela concessionária é abaixo de 5 %. O sistema proposto de pré-seleção teve um desempenho de 40 %.
Apesar de este sistema ter uma boa performance ele ainda deixa um grande número de clientes fraudadores fora do sub-conjunto de clientes a serem inspecionados.

SENDI 2004
XVI SEMINÁRIO NACIONAL DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

Sistema de Detecção de Fraudes em Consumidores de Energia Elétrica Baseado em Rough Sets

José E. Cabral – UFMS, João Onofre Pereira Pinto – UFMS
José Reis Filho – ENERSUL/SA, Edgar M. Gontijo – ENERSUL/SA

E-mail: jeabral@nin.ufms.br

Palavras-chave – Detecção de fraudes, descoberta de conhecimento em banco de dados (KDD), computação flexível, rough sets.

Resumo – Este artigo descreve a teoria e a aplicação de rough Sets na detecção de fraudes em unidades consumidoras de energia elétrica, a partir de banco de dados. O conceito de reduto em rough sets foi usado para remover atributos condicionais e o algoritmo da decisão mínima (MDA) foi aplicado para remover valores insignificantes de atributos condicionais. O banco de dados minimizado aprendeu o comportamento dos consumidores, permitindo ao sistema de regras de classificação predizer perfis de consumidores fraudulentos. Os resultados obtidos foram bons o suficiente para demonstrar que rough sets é uma técnica poderosa para este tipo de problema.

Abstract – This article describes the theory and application of rough sets in fraud detection in electrical energy units consumers from databases. The rough sets concept of reduct was used to remove conditional attributes and the minimal decision algorithm (MDA) was used to remove insignificant classes of each conditional attribute. The minimized database approach the consumers behavior, allowing a classification rule system to predict fraud consumers profiles. The achieved results are good enough to demonstrate that rough sets is a very powerful technique for this type of problem.

1 INTRODUÇÃO

A recuperação de perdas de receitas ocasionadas por fraudes é essencial para manter o equilíbrio financeiro do caixa das empresas distribuidoras de energia elétrica. Porém, a identificação das unidades consumidoras com comportamento fraudulento é uma tarefa complexa. Normalmente, esta tarefa envolve inspeção *in loco*. Considerando-se o elevado número de unidades consumidoras e a não-linearidade do problema, os custos envolvidos assumem valores inviáveis. Esta inviabilidade ocorre porque geralmente tais inspeções são feitas aleatoriamente, ou a partir da experiência do responsável, ou seja, não existe nenhum sistema automático que possa indicar a probabilidade de um determinado consumidor estar fraudando. Como resultado disso, o número de fraudes detectadas na inspeção é muito baixo comparado com o número total de inspeções. O percentual de acerto, de maneira geral, chega a menos de 5%.

Por outro lado, é sabido que sistemas inteligentes de classificação, baseados em Computação Flexível (Soft-Computing), são empregados nas mais diversas áreas, comerciais e acadêmicas, na construção de sistemas de suporte a tomada de decisão. Os resultados oriundos de tais sistemas têm se mostrado bastante satisfatórios.

Na literatura foram reportados muitos trabalhos utilizando técnicas de computação flexível na detecção de fraudes em cartões de crédito. Dentre as técnicas utilizadas, destacam-se Redes Neurais Artificiais [1] e Lógica Nebulosa [2].

Rough sets é uma técnica emergente de computação flexível que vem sendo usada em muitas aplicações de descoberta de conhecimento em banco de dados, como por exemplo na determinação de regras de classificação [3]. No entanto, apesar do seu potencial, tal técnica tem sido preterida para problemas de detecção de fraude.

Este trabalho aborda a teoria e a aplicação de rough sets na detecção de fraudes em unidades consumidoras de energia elétrica, a partir de banco de dados. Inicialmente é feita um breve descrição da Teoria de rough sets, abordando os principais conceitos. Na sequência, a solução na detecção de fraudes à partir de banco de dados é apresentada e finalmente são dados os resultados do sistema.

2 TEORIA DE ROUGH SETS

A teoria de rough sets foi proposta por Zdzislaw Pawlak na década de 80 [4]. Ela aborda basicamente a análise de tabelas (ou banco de dados) com o objetivo de aproximar conceitos e informações contidas nesses repositórios. Muitas vezes estas informações são imprecisas ou incertas, necessitando de métodos ou algoritmos para serem determinadas. Este motivo justifica a grande aplicabilidade da teoria de rough sets na descoberta de conhecimento em banco de dados. Alguns conceitos de rough sets devem ser apresentados para melhor consolidar sua teoria.

2.1 Sistema de Informação e Decisão

Um conjunto de dados é representado por uma tabela. As linhas representam os objetos (exemplos) e as colunas os atributos. Cada objeto caracteriza-se pelos valores de atributos que possui. Esta tabela é chamada sistema de informação [5]. Formalmente, o sistema de informação é definido por $A=(U, A)$ onde U é um conjunto finito e não vazio de objetos e A é um conjunto finito e não vazio de atributos. Os sistemas de informação vêm geralmente acompanhados de outra informação, a classificação do objeto. Esta classificação é representada por outra coluna de atributo. O sistema de informação complementado com este atributo de classificação é chamado de sistema de decisão. Ele é definido por $A=(U, A \cup \{d\})$, onde $d \neq A$ é o atributo de decisão. Os demais atributos de A são chamados de atributos condicionais. Um sistema de informação e decisão é ilustrado na Tabela 1.

2.2 Redutor

Considerando o conjunto A da Tabela 1, todos os elementos pertencentes a U são distintos. Ou seja, considerando os atributos *Tipo de Ligação, Classe e Média de Consumo*, o conjunto U é particionado nos subconjuntos elementares *{e1}, {e2}, {e3}, {e4}, {e5}, {e5}* e *{e6}*. Agora, considerando o subconjunto *{Tipo de Ligação,Classe}* de A, o conjunto U é particionado nos subconjuntos *{e1,e2,e3}, {e4,e6}* e *{e5}*, que são subconjuntos não-elementares. Sendo assim, somente os atributos *Tipo de*

Ligação e *Classe* não conseguem discernir todos exemplos da Tabela 1. Porém, o subconjunto {*Tipo de Ligação,Média de Consumo*} pode particionar o conjunto *U* em subconjuntos elementares. Somente os atributos *Tipo de Ligação* e *Média de Consumo* podem discernir todos exemplos da Tabela 1. Então, conclui-se que o atributo *Classe* é *redundante*. O conjunto *P*={*Tipo de Ligação,Média de Consumo*} não contém atributos redundantes e é chamado *reduto* do conjunto *A*.

Formalmente, o conjunto de atributos *P* é reduto de *A* se *P*⊆*A* mantém as relações de discernibilidade de *A*. Em outras palavras, se *P* tem cardinalidade menos ou igual a *A* e pode representar todos elementos de um sistema de decisão, então *P* é um reduto de *A*. Considerando o reduto *P*={*Tipo de Ligação,Média de Consumo*}, um novo sistema de decisão é mostrado na Tabela 2. Embora a Tabela 2 mostre uma redução (a partir do reduto) do sistema de decisão da Tabela 1, redutos não são necessariamente únicos. Pode existir mais de um reduto para um dado sistema de informação qualquer[5].

Tabela 1 – Sistema de informação (cinza) e decisão (toda a tabela).

Cliente	Tipo de Ligação	Classe	Média de Consumo	Fraude
e1	1	1	Normal	Não
e2	1	1	Alta	Sim
e3	1	1	Baixa	Sim
e4	2	1	Normal	Não
e5	2	2	Alta	Não
e6	2	1	Baixa	Sim

Tabela 2 – Sistema de decisão considerando o reduto.

Cliente	Tipo de Ligação	Média de Consumo	Fraude
e1	1	Normal	Não
e2	1	Alta	Sim
e3	1	Baixa	Sim
e4	2	Normal	Não
e5	2	Alta	Não
e6	2	Baixa	Sim

Esta redução em sistemas de decisão é mais relevante quando o mesmo possui muitos atributos condicionais. Encontrar os redutos é um dos gargalos da teoria de rough sets. Porém, heurísticas baseadas em algoritmos genéticos computam os redutos com menor tempo computacional [5].

2.3 Aproximações

Analisando os atributos de decisão em um sistema de decisão, encontra-se o conjunto dos conceitos. Ele nada mais é que o conjunto dos possíveis valores de classificação que um elemento pode ter. Para o sistema de decisão das Tabelas 2, o conjunto de conceitos é *{Sim,Não}*, informando se o elemento é classificado como fraudador ou não. Considerando a Tabela 2, os elementos de *U* estão bem definidos. Para ilustrar uma situação problemática, será adicionado à Tabela 2 mais dois elementos, dando origem a Tabela 3. Os conceitos da Tabela 3 são representados pelos subconjuntos *{e1,e4,e5,e8}* e *{e2,e3,e6,e7}*. Porém, os elementos *e5* e *e7* têm classificação diferente e possuem os mesmos valores de atributos condicionais. O mesmo acontece com os exemplos *e6* e *e8*. Para tentar contornar esse problema, rough sets define três subconjuntos de *U*.

Seja *X* um conceito de um sistema de decisão. Pode ser encontrado um subconjunto de *X* com exemplos que *com certeza* pertençam ao conceito *X*. Este subconjunto é chamado *aproximação inferior* de *X*, ou simplesmente $\underline{X}$. Considerando a Tabela 3, se *X*={*e1,e4,e5,e8*}, então $\underline{X}$ ={*e1,e4*}. Similarmente, se *X*={*e2,e3,e6,e7*}, então $\underline{X}$ ={*e2,e3*}. Note que sempre $\underline{X} \subseteq X$.

A *aproximação superior* de X, ou simplesmente $\overline{X}$, corresponde ao subconjunto de U com exemplos que *podem* pertencer ao conceito X. Considerando a Tabela 3, se $X=\{e1,e4,e5,e8\}$, então $\overline{X}=\{e1,e4,e5,e6,e7,e8\}$. Similarmente, se $X=\{e2,e3,e6,e7\}$, então $\overline{X}=\{e2,e3,e5,e6,e7,e8\}$. Note que sempre $X \subset \overline{X}$.

A região de fronteira de X, ou simplesmente BX, corresponde a um subconjunto de U com exemplos que pertencem a $\overline{X}$, mas não pertencem a $\underline{X}$, ou seja, $BX = \overline{X} - \underline{X}$. Se BX é vazio, então $\overline{X}$ e $\underline{X}$ possuem os mesmo elementos; em outras palavras, o sistema de decisão, neste caso, não contém elementos inconsistentes. Consequentemente, quanto maior a cardinalidade de BX, maior é a indiscernibilidade entre os conceitos. A Figura 1 ilustra a distribuição das aproximações para o sistema de informação da Tabela 3.

Tabela 3 – Sistema de decisão inconsistente.

Cliente	Tipo de Ligação	Média de Consumo	Fraude
e1	1	Normal	Não
e2	1	Alta	Sim
e3	1	Baixa	Sim
e4	2	Normal	Não
e5	2	Alta	Não
e6	2	Baixa	Sim
e7	2	Alta	Sim
e8	2	Baixa	Não

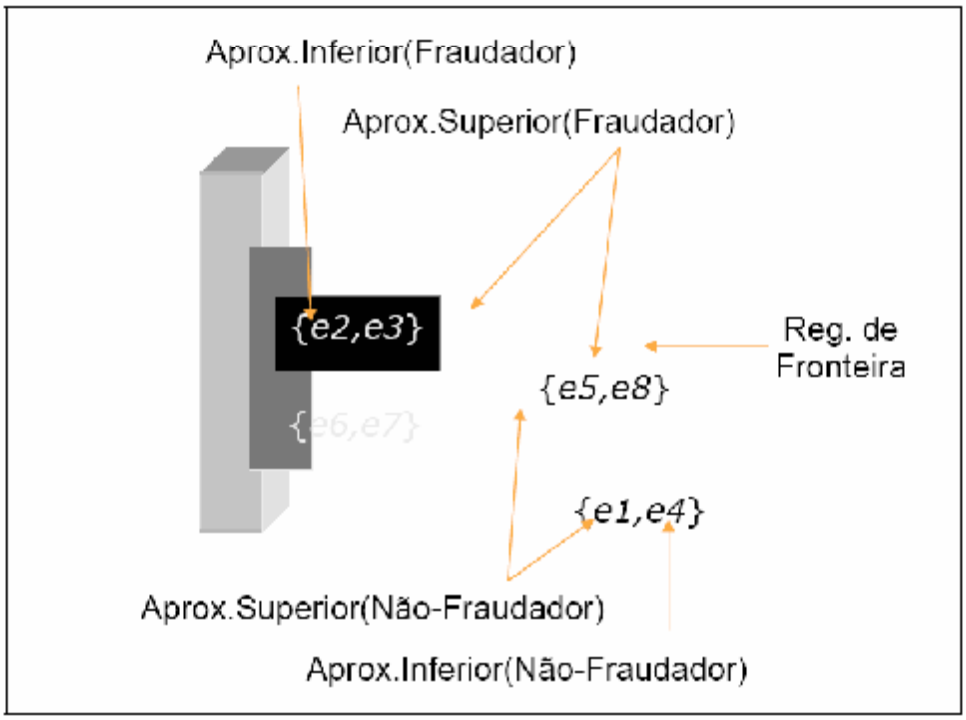

Figura 1 – Aproximação inferior, superior e região de fronteira.

2.4 Coeficiente de Incerteza

Pode ser interessante saber o quanto um conceito é bem definido ou não dentro de um sistema de decisão. Para tal, define-se o coeficiente de incerteza pela equação 1:

$$\alpha(X) = |\underline{X}| / |\overline{X}| \qquad (1)$$

O coeficiente de incerteza pode ser entendido como a qualidade da aproximação do conceito X. Ou seja, quanto mais $\alpha(X)$ se aproxima de 1, mais o conceito X é definido de forma exata (crips). E quanto mais $\alpha(X)$ se aproxima de 0, mais o conceito X é definido de forma imprecisa (rough).

2.5 Algoritmo da Decisão Mínima

O algoritmo da decisão mínima (MDA) é utilizado para reduções em sistemas de decisão ou banco de regras [6]. O MDA compara os valores dos atributos de um objeto com os demais objetos do sistema de decisão. Caso encontre valores de atributos que possam ser eliminados sem que dois objetos tornem-se contraditórios, o MDA retira este valor do objeto.

3 APLICAÇÃO

O software MATLAB [7] foi utilizado na implementação dos conceitos e algoritmos da teoria de rough sets, aproveitando sua facilidade de uso e portabilidade.

A solução aplicada consistiu em dividir o banco de dados em dados de treinamento e teste. Esta divisão é um procedimento típico do aprendizado supervisionado. O primeiro passo empregado foi eliminar registros repetidos, isto é, deixar os dados de treinamento somente com objetos distintos. No segundo passo, iniciou-se o uso dos conceitos de rough sets apresentados. Utilizou-se o software Rosetta [8] para determinar um redutor. A partir dele, os atributos linearmente dependentes foram eliminados dos dados de treinamento. Esta eliminação diminui diretamente a dimensão dos dados. Embora tenham sido eliminados atributos, alguns objetos também podem ser suprimidos. Isto porque novos objetos tornam-se idênticos com a retirada de atributos. A aproximação inferior para o conceito *Fraudador* = {*Sim*} foi encontrada e os demais registros eliminados. Conseqüentemente, não existiram mais objetos da região de fronteira. Garante-se assim que os objetos restantes nos dados de treinamento estão totalmente no conceito. A seguir, foi aplicado o MDA sobre os dados de treinamento reduzidos nos passos à cima. O algoritmo conseguiu minimizar significantemente os dados de treinamento. Novamente, após a aplicação do MDA, outros objetos idênticos surgiram e foram removidos.

Finalmente, para cada objeto dos dados de treinamento, uma regra de classificação foi derivada, a qual determina um perfil de fraudador. O conjunto de regras de classificação é chamado sistema de regras de classificação. Com o sistema de regras de classificação em mãos, bastou testar a qualidade das regras nos dados de teste.

4 RESULTADOS

O banco de dados utilizado possuía aproximadamente 40.600 registros (exemplos, objetos), dos quais 90% são classificados como Normal e 10% como Fraude. O conjunto de dados foi separado aleatoriamente em 20.300 registros para treinamento e 20.300 registros para teste. Tomando somente o conjunto de registros para treinamento, foi encontrado o reduto para este conjunto. A aproximação inferior, correspondendo aos exemplos fraudulentos, constou de 630 registros. Finalmente foi aplicado o MDA, resultando em 450 registros, sendo que estes registros resultaram em regras esparsas, i. e., nem todos os atributos foram utilizados em cada regras. Assim, de posse das regras geradas, o sistema foi submetido ao conjunto de teste com índice de acerto da ordem de 20 %, o que ficou abaixo do objetivo final que é de 30%.

5 Conclusões

- Rough sets é uma poderosa ferramenta de detecção de fraudes, principalmente quando não existem informações preliminares sobre o sistema, além do banco de dados;
- Apesar de alto custo computacional, os algoritmos de rough sets são de fácil implementação e compreensão;
- Um sistema híbrido, explorando conceitos de lógica fuzzy, pode ser a base de futuros trabalhos na detecção de fraudes usando rough sets.
- O sistema obtido resultou em acertos da ordem de 20%, o que ficou abaixo do objetivo final, que é de 30%.
- A melhoria está em desenvolvimento, porém constatou-se que o resultado obtido até o presente está abaixo do desejado principalmente devido à qualidade (fidelidade à realidade) dos dados.
- Novos estudos estão sendo feitos, utilizando-se das aproximações, no sentido de melhorar o pré-tratamento dos dados.

Referências Bibliográficas

[1] R.Brause, T. Langsdorf, M.Hepp: Neural data mining for card fraud detection. Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on, 9-11 Nov.1999, pp. 103 – 106.

[2] Deshmukh, A.; Talluri, T.L.N.: A rule based fuzzy reasoning system for assessing the risk of management fraud. Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation', 1997 IEEE International Conference on , Volume: 1 , 12-15 Oct. 1997, pp. 669 - 673 vol.1.

[3] H. Furuta, M. Hirokane, Y. Mikumo: Extraction Method Based on Rough Sets Theory of Rule-Type Knowledge from Diagnostic Cases of Slope-Failure Danger Levels. Rough Sets in Knowledge Discovery 2 - Application, Case Studies and Software Systems. Part 1. Applications. Chapter 10, Pages: 178-192.

[4] Pawlak, Z. (1982). Rough Sets. International Jornal of Computer ans Information Sciences, pages 341-356.

[5] S. Pal, A. Skowron. Rough-fuzzy hybridization: a new trend in decision-marketing. Springer-Verlag Singapore Pte. Ldt. 1999.

[6] H. Furuta, M. Hirokane, Y. Mikumo: Extraction Method Based on Rough Sets Theory of Rule-Type Knowledge from Diagnostic Cases of Slope-Failure Danger Levels. Rough Sets in Knowledge Discovery 2 - Application, Case Studies and Software Systems. Part 1. Applications, Chapter 10, Pages: 178-192.

[7] MATLAB: The Language of technical Computing. Copyright 1984-2004 The Mathworks, Inc. http://www.mathworks.com

[8] Ohrn, A. Rosetta: Technical reference manual. Technical report, Knowledge System Group, Norwergian University on Science and Technology, NO. http://rosetta.lcb.uu.se/general/

Sistema de Detecção de Fraudes em Consumidores de Energia Elétrica Baseado em Rough Sets

J. E. Cabral, UFMS, J. O. P. Pinto, UFMS, J. Reis Filho, ENERSUL S.A. e
E. M. Giontijo, ENERSUL S.A.

Resumo – **Este artigo descreve a teoria e a aplicação de rough Sets na detecção de fraudes em unidades consumidoras de energia elétrica, a partir de banco de dados. O conceito de redutor em rough sets foi usado para remover atributos condicionais e o algoritmo da decisão mínima (MDA) foi aplicado para remover valores insignificantes de atributos condicionais. O banco de dados minimizado aprendeu o comportamento dos consumidores, permitindo ao sistema de regras de classificação predizer perfis de consumidores fraudulentos. Os resultados obtidos foram bons o suficiente para demonstrar que rough sets é uma técnica poderosa para este tipo de problema.**

Palavras-chave – **Detecção de fraudes, descoberta de conhecimento em banco de dados (KDD), computação flexível, rough sets.**

I. INTRODUÇÃO

A recuperação de perdas de receitas ocasionadas por fraudes é essencial para manter o equilíbrio financeiro do caixa das empresas distribuidoras de energia elétrica. Porém, a identificação das unidades consumidoras com comportamento fraudulento é uma tarefa complexa. Normalmente, esta tarefa envolve inspeção *in loco*. Considerando-se o elevado número de unidades consumidoras e a não linearidade do problema, os custos envolvidos assumem valores inviáveis. Esta inviabilidade ocorre porque geralmente tais inspeções são feitas aleatoriamente, ou a partir da experiência do responsável, ou seja, não existe nenhum sistema automático que possa indicar a probabilidade de um determinado consumidor estar fraudando. Como resultado disso, o número de fraudes detectadas na inspeção é muito baixo comparado com o número total de inspeções. O percentual de acerto, de maneira geral, chega a menos de 5%.

Por outro lado, é sabido que sistemas inteligentes de classificação, baseados em Computação Flexível (Soft Computing), são empregados nas mais diversas áreas, comerciais e acadêmicas, na construção de sistemas de suporte à tomada de decisão. Os resultados oriundos de tais sistemas têm se mostrado bastante satisfatórios.

Na literatura foram reportados muitos trabalhos utilizando técnicas de computação flexível na detecção de fraudes em cartões de crédito. Dentre as técnicas utilizadas, destacam-se Redes Neurais Artificiais [1] e Lógica Nebulosa [2].

Rough sets é uma técnica emergente de computação flexível que vem sendo usada em muitas aplicações de descoberta de conhecimento em banco de dados, como por exemplo na determinação de regras de classificação [3]. No entanto, apesar do seu potencial, tal técnica tem sido preterida para problemas de detecção de fraude.

Este trabalho aborda a teoria e a aplicação de rough sets na detecção de fraudes em unidades consumidoras de energia elétrica, a partir de banco de dados. Inicialmente é feita um breve descrição da Teoria de rough sets, abordando os principais conceitos. Na sequência, a solução na detecção de fraudes a partir de banco de dados é apresentada e finalmente são dados os resultados do sistema.

II. TEORIA DE ROUGH SETS

A teoria de rough sets foi proposta por Zdzislaw Pawlak na década de 80 [4]. Ela aborda basicamente a análise de tabelas (ou banco de dados) com o objetivo de aproximar conceitos e informações contidas nesses repositórios. Muitas vezes estas informações são imprecisas ou incertas, necessitando de métodos ou algoritmos para serem determinadas. Este motivo justifica a grande aplicabilidade da teoria de rough sets na descoberta de conhecimento em banco de dados. Alguns conceitos de rough sets devem ser apresentados para melhor consolidar sua teoria.

A. Sistema de Informação e Decisão

Um conjunto de dados é representado por uma tabela. As linhas representam os objetos (exemplos) e as colunas os atributos. Cada objeto caracteriza-se pelos valores de atributos que possui. Esta tabela é chamada sistema de informação [5]. Formalmente, o sistema de informação é definido por $\mathcal{A}=(U, A)$ onde U é um conjunto finito e não vazio de objetos e A é um conjunto finito e não vazio de atributos. Os sistemas de informação vêm geralmente acompanhados de outra informação, a classificação do objeto. Esta classificação é representada por

---

Este projeto foi desenvolvido no programa de Pesquisa e Desenvolvimento da ENERSUL S.A.

outra coluna de atributo. O sistema de informação complementado com este atributo de classificação é chamado de sistema de decisão. Ele é definido por $A = (U, A \cup \{d\})$, onde $d \notin A$ é o atributo de decisão. Os demais atributos de A são chamados de atributos condicionais. Um sistema de informação e decisão é ilustrado na Tabela I.

TABELA I
SISTEMA DE INFORMAÇÃO (CINZA) E DECISÃO (TODA A TABELA)

Cliente	Tipo de Ligação	Classe	Média de Consumo	Fraude
e1	1	1	Normal	Não
e2	1	1	Alta	Sim
e3	1	1	Baixa	Sim
e4	2	1	Normal	Não
e5	2	2	Alta	Não
e6	2	1	Baixa	Sim

B. Reduto

Considerando o conjunto A da Tabela I, todos os elementos pertencentes a U são distintos. Ou seja, considerando os atributos *Tipo de Ligação, Classe e Média de Consumo*, o conjunto U é particionado nos subconjuntos elementares *{e1}, {e2}, {e3}, {e4}, {e5}, {e5}* e *{e6}*. Agora, considerando o subconjunto *{Tipo de Ligação,Classe}* de A, o conjunto U é particionado nos subconjuntos *{e1,e2,e3}, {e4,e6}* e *{e5}*, que são subconjuntos não-elementares. Sendo assim, somente os atributos *Tipo de Ligação* e *Classe* não conseguem discernir todos exemplos da Tabela I. Porém, o subconjunto *{Tipo de Ligação,Média de Consumo}* pode particionar o conjunto U em subconjuntos elementares. Somente os atributos *Tipo de Ligação* e *Média de Consumo* podem discernir todos exemplos da Tabela I. Então, conclui-se que o atributo *Classe é redundante*. O conjunto P=*{Tipo de Ligação,Média de Consumo}* não contém atributos redundantes e é chamado *reduto* do conjunto A.

Formalmente, o conjunto de atributos P é reduto de A se $P \subseteq A$ mantém as relações de discernibilidade de A. Em outras palavras, se P tem cardinalidade menor ou igual a A e pode representar todos elementos de um sistema de decisão, então P é um reduto de A. Considerando o reduto P=*{Tipo de Ligação,Média de Consumo}*, um novo sistema de decisão é mostrado na Tabela II. Embora a Tabela II mostre uma redução (a partir do reduto) do sistema de decisão da Tabela I, redutos não são necessariamente únicos. Pode existir mais de um reduto para um dado sistema de informação qualquer [5].

Esta redução em sistemas de decisão é mais relevante quando o mesmo possui muitos atributos condicionais. Encontrar os redutos é um dos gargalos da teoria de rough sets. Porém, heurísticas baseadas em algoritmos genéticos computam os redutos com menor tempo computacional [5].

C. Aproximações

Analisando os atributos de decisão em um sistema de decisão, encontra-se o conjunto dos conceitos. Ele nada mais é que o conjunto dos possíveis valores de classificação que um elemento pode ter. Para o sistema de decisão da Tabela II, o conjunto de conceitos é *{Sim,Não}*, informando se o elemento é classificado como fraudador ou não. Considerando a Tabela II, os elementos de U estão bem definidos. Para ilustrar uma situação problemática, será adicionado à Tabela II mais dois elementos, dando origem à Tabela III. Os conceitos da Tabela III são representados pelos subconjuntos *{e1,e4,e5,e8}* e *{e2,e3,e6,e7}*. Porém, os elementos *e5* e *e7* têm classificação diferente e possuem os mesmos valores de atributos condicionais. O mesmo acontece com os exemplos *e6* e *e8*. Para tentar contornar esse problema, rough sets define três subconjuntos de U.

TABELA II
SISTEMA DE DECISÃO CONSIDERANDO O REDUTO

Cliente	Tipo de Ligação	Média de Consumo	Fraude
e1	1	Normal	Não
e2	1	Alta	Sim
e3	1	Baixa	Sim
e4	2	Normal	Não
e5	2	Alta	Não
e6	2	Baixa	Sim

TABELA III
SISTEMA DE DECISÃO INCONSISTENTE

Cliente	Tipo de Ligação	Média de Consumo	Fraude
e1	1	Normal	Não
e2	1	Alta	Sim
e3	1	Baixa	Sim
e4	2	Normal	Não
e5	2	Alta	Não
e6	2	Baixa	Sim
e7	2	Alta	Sim
e8	2	Baixa	Não

Seja X um conceito de um sistema de decisão. Pode ser encontrado um subconjunto de X com exemplos que *com certeza* pertencam ao conceito X. Este subconjunto é chamado *aproximação inferior* de X, ou simplesmente $\underline{X}$. Considerando a Tabela III, se X=*{e1,e4,e5,e8}*, então $\underline{X}$ =*{e1,e4}*. Similarmente, se X *{e2,e3,e6,e7}*, então $\underline{X}$ =*{e2,e3}*. Note que sempre $\underline{X} \subseteq X$.

A *aproximação superior* de X, ou simplesmente $\overline{X}$, corresponde ao subconjunto de U com exemplos que *podem* pertencer ao conceito X. Considerando a Tabela III, se X *{e1,e4,e5,e8}*, então $\overline{X}$ =*{e1,e4,e5,e6,e7,e8}*. Similarmente, se

$X = \{e2,e3,e6,e7\}$, então $\overline{X} = \{e2,e3,e5,e6,e7,e8\}$. Note que sempre $X \subseteq \overline{X}$.

A região de fronteira de X, ou simplesmente BX, corresponde a um subconjunto de U com exemplos que pertencem a $\overline{X}$, mas não pertencem a $\underline{X}$, ou seja, $BX = \overline{X} - \underline{X}$. Se BX é vazio, então $\overline{X}$ e $\underline{X}$ possuem os mesmo elementos, em outras palavras, o sistema de decisão, neste caso, não contém elementos inconsistentes. Conseqüentemente, quanto maior a cardinalidade de BX, maior é a indiscernibilidade entre os conceitos. A Figura 1 ilustra a distribuição das aproximações para o sistema de informação da Tabela III.

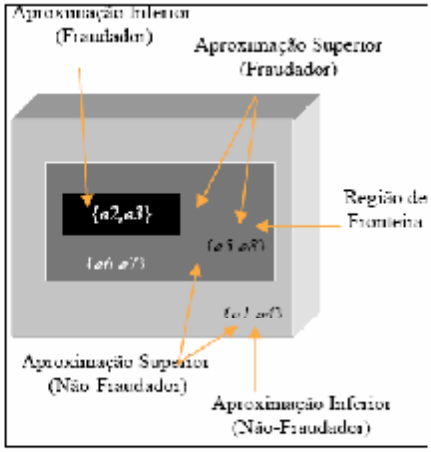

Figura 1 – Aproximação inferior, superior e região de fronteira

D. Coeficiente de Incerteza

Pode ser interessante saber o quanto um conceito é bem definido ou não dentro de um sistema de decisão. Para tal, define-se o coeficiente de incerteza por "(1)"

$$\alpha(X) = |\underline{X}| / |\overline{X}| \qquad (1)$$

O coeficiente de incerteza pode ser entendido como a qualidade da aproximação do conceito X. Ou seja, quanto mais $\alpha(X)$ se aproxima de 1, mais o conceito X é definido de forma exata (crisp). E quanto mais $\alpha(X)$ se aproxima de 0, mais o conceito X é definido de forma imprecisa (rough).

E. Algoritmo da Decisão Mínima

O algoritmo da decisão mínima (MDA) é utilizado para reduções em sistemas de decisão ou banco de regras [6]. O MDA compara os valores dos atributos de um objeto com os demais objetos do sistema de decisão. Caso encontre valores de atributos que possam ser eliminados sem que dois objetos tornem-se contraditórios, o MDA retira este valor do objeto.

III. APLICAÇÃO

O software MATLAB [7] foi utilizado na implementação dos conceitos e algoritmos da teoria de rough sets, aproveitando sua facilidade de uso e portabilidade.

A solução aplicada consistiu em dividir o banco de dados em dados de treinamento e teste. Esta divisão é um procedimento típico do aprendizado supervisionado. O primeiro passo empregado foi eliminar registros repetidos, isto é, deixar os dados de treinamento somente com objetos distintos. No segundo passo, iniciou-se o uso dos conceitos de rough sets apresentados. Utilizou-se o software Rosetta [8] para determinar um reduto. A partir dele, os atributos linearmente dependentes foram eliminados dos dados de treinamento. Esta eliminação diminui diretamente a dimensão dos dados. Embora tenham sido eliminados atributos, alguns objetos também podem ser suprimidos. Isto porque novos objetos tornam-se idênticos com a retirada de atributos. A aproximação inferior para o conceito *Fraudador* = {*Sim*} foi encontrada e os demais registros eliminados. Conseqüentemente, não existiram mais objetos da região de fronteira. Garante-se assim que os objetos restantes nos dados de treinamento estão totalmente no conceito. A seguir, foi aplicado o MDA sobre os dados de treinamento reduzidos nos passos à cima. O algoritmo conseguiu minimizar significantemente os dados de treinamento. Novamente, após a aplicação do MDA, outros objetos idênticos surgiram e foram removidos.

Finalmente, para cada objeto dos dados de treinamento, uma regra de classificação foi derivada, a qual determina um perfil de fraudador. O conjunto de regras de classificação é chamado sistema de regras de classificação. Com o sistema de regras de classificação em mãos, bastou testar a qualidade das regras nos dados de teste.

IV. RESULTADOS

O banco de dados utilizado possuía aproximadamente 40.600 registros (exemplos, objetos), dos quais 90% são classificados como Normal e 10% como Fraude. O conjunto de dados foi separado aleatoriamente em 20.300 registros para treinamento e 20.300 registros para teste. Tomando somente o conjunto de registros para treinamento, foi encontrado o reduto para este conjunto. A aproximação inferior, correspondendo aos exemplos

fraudulentos, constou de 630 registros. Finalmente foi aplicado o MDA, resultando em 450 registros, sendo que estes registros resultaram em regras esparsas, i. e., nem todos os atributos foram utilizados em cada regras. Assim, de posse das regras geradas, o sistema foi submetido ao conjunto de teste com índice de acerto da ordem de 20 %, o que ficou abaixo do objetivo final que é de 30%.

V. CONCLUSÕES

- Rough sets é uma poderosa ferramenta de detecção de fraudes, principalmente quando não existem informações preliminares sobre o sistema, além do banco de dados;
- Apesar do alto custo computacional, os algoritmos de rough sets são de fácil implementação e compreensão;
- Um sistema híbrido, englobando conceitos de lógica fuzzy, pode ser a base de futuros trabalhos na detecção de fraudes usando rough sets.
- O sistema obtido resultou em acertos da ordem de 20%, o que ficou abaixo do objetivo final, que é de 30%.
- A melhoria está em desenvolvimento, porém constatou-se que o resultado obtido até o presente está abaixo do desejado principalmente devido à qualidade (fidelidade à realidade) dos dados.
- Novos estudos estão sendo feitos, utilizando-se das aproximações, no sentido de melhorar o pré-tratamento dos dados.

VI. REFERÊNCIAS BIBLIOGRÁFICAS

[1] R. Brause, T. Langsdorf, M. Hepp: Neural data mining for credit fraud detection. Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on, 9-11 Nov 1999, pp. 103 – 106.

[2] Deshmukh, A.; Talluru, T.L.N.: A rule based fuzzy reasoning system for assessing the risk of management fraud. Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation', 1997 IEEE International Conference on , Volume: 1 , 12-15 Oct 1997, pp. 669 - 673 vol.1.

[3] H. Furuta, M. Hirokane, Y. Mikumo: Extraction Method Based on Rough Sets Theory of Rule Type Knowledge from Diagnostic Cases of Slope-Failure Danger Levels. Rough Sets in Knowledge Discovery 2 - Application, Case Studies and Software Systems, Part 1. Applications, Chapter 10, Pages. 178-192.

[4] Pawlak, Z. (1982). Rough Sets. International Jornal of Computer and Information Sciences, pages 341-356.

[5] S. Pal, A. Skowron. Rough-fuzzy hybridization: a new trend in decision making. Springer Verlag Singapore Pte Ltd. 1999.

[6] H. Furuta, M. Hirokane, Y. Mikumo: Extraction Method Based on Rough Sets Theory of Rule-Type Knowledge from Diagnostic Cases of Slope-Failure Danger Levels. Rough Sets in Knowledge Discovery 2 - Application, Case Studies and Software Systems. Part 1. Applications, Chapter 10, Pages. 178-192.

[7] MATLAB. The Language of technical Computing. Copyright 1984-2004 The Mathworks, Inc. http://www.mathworks.com

[8] Olav, A. Rosetta. Technical reference manual. Technical report, Knowledge System Group, Norwegian University on Science and Technology, NO. http://rosetta.idi.ntnu.no/general/

Bibliografia

Agência Nacional de Energia Elétrica - ANEEL, Condições Gerais de Fornecimento de Energia Elétrica, Resolução 456, 2000.

Aleskerov, E., Freisleben, B. and Rao, B., "CARDWATCH: A Neural Network Based Data Mining System for Credit Card Fraud Detection", Proceedings of the IEEE/IAFE, 1997.

Anderson, D., McNeil, G., Artificial Neural Networks Technology, 1992.

Associação Brasileira de Distribuidores de Energia – ABRADEE, CODI 08-05 Perdas Comerciais, 1998.

Associação Brasileira de Distribuidores de Energia – ABRADEE, CODI 19-34 Metodologia para determinação, Análise e Otimização de Perdas Técnicas em Sistemas de Distribuição, 1994.

Braga, A. P., Carvalho, A. C. P. L. F., Ludermir, T. B., Fundamentos de redes neurais artificiais, 1998.

Bolton, R. J., Hand, D. J., Unsupervised Profiling Methods for Fraud Detection, 2001.

Breiman, L., Friedman,R.A., Olshen, J.H. e Stone, C.J., Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.

Cabral, J.E.; Gontijo, E.M.. Fraud detection in electricalenergy consumers using rough sets. Systems, Man and Cybernetics, 2004.

Eller, N. A., Arquitetura de informação para o gerenciamento de perdas comericiais de energia elétrica, Programa de Pós Graduação, Engenharia da Produção, UFSC, 2003.

Engels, R. e Theusinger, C. Using a Data Metric for Preprocessing Advice for Data Mining Applications, European Conference on Artificial Intelligence, ECAI 1998.

Fayyad, U., Piatetsky-Shapiro, G. e Smyth, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data, ACM, 1996

Han, J., Kamber M., Data Mining Concepts and Techniques, Morgan-Kaufmann Publishers, 2001.

Haykin, S., Redes Neurais Princípios e Prática, Ed. Bookman, 2ª Edição, 2001.

KPMG Transition and Forensic Services Ltda. A fraude no Brasil Relatório de Pesquisa, 2004.

Michell, T., Machine Learning. Mcgraw Hill. 1997.

Passini, S.R.R.; Toledo, Mineração de Dados para Detecção de Fraudes em Ligações de Água. Dissertação de Mestrado. PUC-Campinas. Mar 2002.

Quinlan, J. R., Induction of Decision Trees, Centre of Advanced Computer Sciences, New South Wales Institute of Technology, Sidney, Australia, 1985.

Quinlan, J. R., Induction of decision trees, Machine Learning, 1986.

Quinlan, J. R., C4.5: Programs for Machine Learning. San Mateo, CA, 1993.

Reis, J. Filho; Gontijo, E.M.. Fraud Identification In Electricity Company Costumers Using Decision Tree Systems, Man and Cybernetics, 2004.

Russel, S. J.; Norvig, Peter. Artificial intelligence: a modem approach. Prentice Hall. 1995.

Souza, F. J. de: Modelos Neuro-Fuzzy Hierárquicos. Tese de Doutorado. Puc-Rio, 1999.

Wermter, S. e Sum, R. An Overview of Hybrid Neural Systems, 2000.

Y. Kou, C.T. Lu, S. Sirwongwattana, Y.P. Huang, .Survey of Fraud Detection Techniques,. Proceedings of the 2004 International Conference on Networking, Sensing, and Control, pp. 749-754, Taipei, Taiwan, March 21-23, 2004.

Z. Pawlak, Rough Sets - Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1991