XVIII Seminário Nacional de Distribuição de Energia Elétrica

SENDI 2008 - 06 a 10 de outubro

Olinda - Pernambuco - Brasil

Aplicando Mineração De Dados Para Identificar Perfis De Consumidores Fraudadores: Um Estudo De Caso Aplicado Ao Setor Elétrico

Mirela Ferreira César	Shirlei Debastiani Cortez	Sílvio César Cazella
Companhia Estadual de Distribuição de Energia Elétrica	Companhia Estadual de Distribuição de Energia Elétrica	Centro de Ciências Exatas e Tecnológicas, Universidade do Vale do Rio dos Sinos (Unisinos)
mirelac@ceee.com.br	shirleide@ceee.com.br	cazella@unisinos.br

Palavras-chave
Fiscalização
Irregularidade
Mineração

Resumo
A redução de perdas comerciais, no setor elétrico, é um desafio enfrentado pelas Companhias de Distribuição. A empresa do estudo de caso possui um enorme volume de dados referentes aos seus consumidores propiciando, desta forma, a aplicação de mineração de dados visando à descoberta de conhecimento. Este artigo apresenta a aplicação de algoritmo de classificação de dados buscando identificar o perfil do consumidor que comete fraudes no consumo de energia elétrica. A base de dados utilizada para a realização dos experimentos constitui-se em uma base de dados real disponibilizada pela empresa. No final do artigo é descrito o conhecimento descoberto e que poderá contribuir com o processo atual de fiscalização.

1. Introdução

Este artigo apresenta uma aplicação prática de mineração de dados na identificação do perfil do consumidor fraudador no setor de energia elétrica. O termo técnico perda comercial, utilizado no setor elétrico, qualifica a perda de energia elétrica proveniente de furto e da falta de manutenção dos equipamentos de medição. O furto e a fraude são crimes praticados contra a sociedade, na medida em que diminuem a arrecadação de tributos que seriam utilizados para atender às necessidades básicas (educação, saúde, segurança). Para combater estas práticas criminosas, as concessionárias investem largamente em programas de redução de perdas, quando poderiam estar conduzindo seus investimentos para melhorar os níveis de regularidade, continuidade e eficiência do fornecimento. A fiscalização é o processo utilizado para recuperar as perdas, onde equipes se deslocam até o consumidor que será fiscalizado.

O método de pesquisa utilizado neste trabalho foi o estudo de caso e a empresa selecionada foi a Companhia Estadual de Distribuição de Energia Elétrica[1]. A Companhia de Distribuição de Energia

[1] http://www.ceee.com.br/

Elétrica possui equipes que executam fiscalização com a finalidade de localizar irregularidade ou avaria no consumo de energia elétrica. Estas equipes não têm capacidade de fiscalizar todos os consumidores que compreendem sua área de atuação. A fiscalização prioriza as demandas geradas a partir de denúncias e análise do faturamento.
Para gerar o modelo referente aos perfis de fraudadores foi utilizada a classificação de dados e o algoritmo J48 implementado na ferramenta WEKA[2].

1.1 Motivação e Contribuição

A Companhia Estadual de Distribuição de Energia Elétrica possui a maioria dos processos informatizados, sendo enorme o volume de dados disponíveis de seus consumidores propiciando a aplicação da tecnologia de mineração de dados. Ao mesmo tempo, não existe muita informação sobre seus consumidores para contribuir com as equipes de fiscalização e estas dependem de denúncias ou análise do faturamento. Foi este conjunto de fatores que motivou este trabalho, pois as técnicas de mineração fornecem meios para descobrir relações interessantes e estes dados avaliados podem representar uma fonte de informação valiosa para a empresa. Através desta tecnologia, pode-se descobrir uma nova maneira para classificar estes consumidores contribuindo no processo de fiscalização da empresa.

1.2 Estrutura do artigo

O artigo encontra-se organizado em 5 seções. A seção 2 apresenta o conceito de Descoberta do Conhecimento em Base de Dados, mineração de dados e classificação de dados. A seção 3 apresenta a empresa e o estudo de caso realizado. Na seção 4 o software utilizado é descrito, e os resultados obtidos nas experimentações são apresentados. Finalmente, na seção 5 é apresentada a conclusão deste estudo e propostas de trabalhos futuros.

2. Descoberta de Conhecimento em Base de Dados

Conforme estudos de Han J. e Kamber M. (2001, p. 7), a Descoberta de Conhecimento em Base de Dados (DCDB) (do inglês *Knowledge Discovery in Databases* (KDD)) é tratado por muitas pessoas como um sinônimo para mineração de dados. Isto porque os autores entendem que mineração de dados se refere à extração ou mineração de conhecimento a partir de grandes quantidades de dados. Segundo Scheffer, T. (2001, p. 424-435), a função do processo de Descoberta de Conhecimento em Base de Dados (DCBD) é descobrir padrões desconhecidos em grandes volumes de dados, onde a análise dos dados de forma manual seria quase impossível. De acordo com Han J. e Kamber M. (2001, p. 6), a descoberta de conhecimento consiste numa seqüência iterativa e interativa dos seguintes passos (a figura 1 apresenta os passos descritos a seguir):

1. Integração de dados (onde múltiplas fontes de dados podem ser combinadas);
2. Seleção de dados (onde dados relevantes para tarefas de análise são selecionados da base de dados);
3. Transformação de dados (onde dados são transformados ou consolidados para uma forma apropriada para a mineração através de operações de sumarização ou agregação, por exemplo);
4. Mineração de dados (aplicação de algoritmos de mineração de dados conforme a tarefa de mineração selecionada; exemplo o algoritmo J48 para a tarefa de classificação de dados);
5. Avaliação de modelos (para identificar os modelos verdadeiramente interessantes, representando conhecimento baseado em alguma(s) métrica(s) interessante(s));
6. Apresentação do conhecimento (visualização e técnicas de apresentação são utilizadas para demonstrar o conhecimento minerado para o usuário).

[2] http://www.cs.waikato.ac.nz/ml/weka/

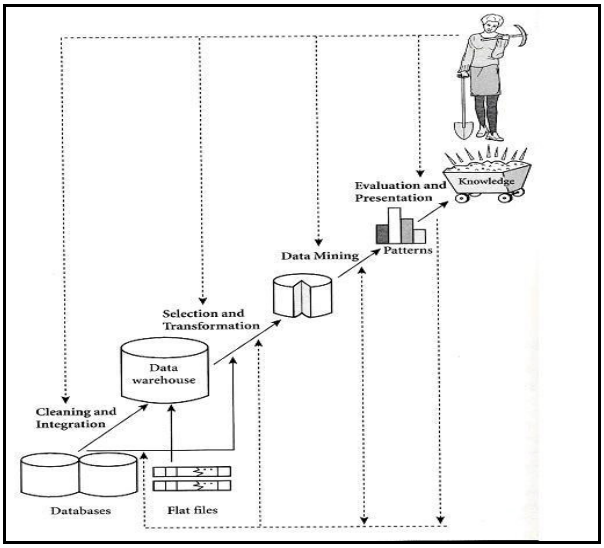

Figura 1: Descoberta do Conhecimento em Base de Dados
Fonte: Han, J. and Kamber, M. 2001 p. 6.

2.1 Mineração de Dados

De acordo com Ian H. Witten e Eibe Frank (2000, p. 3), mineração de dados é definida como um processo de descobrir padrões escondidos em dados. O processo deve ser automático ou semi-automático. Os modelos descobertos devem ser significativos e conduzir alguma vantagem, geralmente uma vantagem econômica. Os dados estão invariavelmente presentes em quantidades substanciais.

Segundo o autor Carvalho (2001, p.1-244), mineração de dados trata do uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertos a olho nu pelo ser humano. Quando aplicado na empresa, melhora a interação entre empresa e cliente, aumenta as vendas e dirige as estratégias de marketing.

Para Han J. e Kamber M. (2001, p. 7), o passo da mineração de dados pode interagir com o usuário ou uma base de conhecimento. Os padrões interessantes são apresentados para o usuário, e podem ser armazenados como novo conhecimento.

De acordo com esta visão, mineração de dados é somente um passo no processo de DCBD inteiro, apesar de essencial já que descobre padrões escondidos para avaliação.

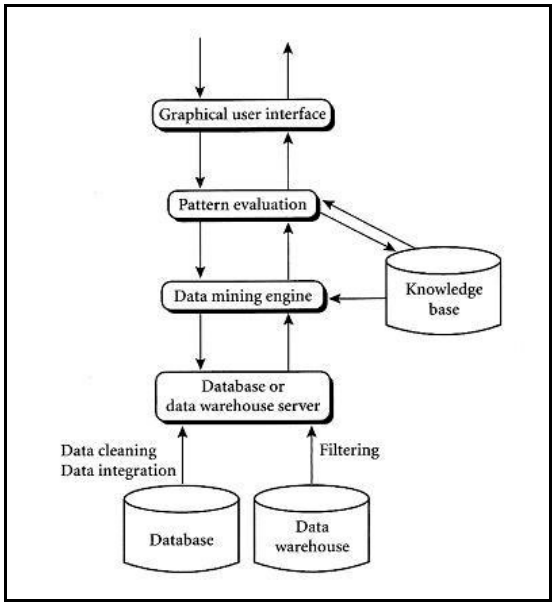

Figura 2: Arquitetura típica de um sistema de mineração de dados.

Fonte: Han, J. and Kamber, M. 2001 p. 8.

2.2 Tarefa de Mineração – Classificação de Dados

Han J. e Kamber M. (2001, p. 24), definem classificação como o processo de encontrar um conjunto de modelos que descrevem e distinguem classes de dados ou conceitos, com o objetivo de ser apto para usar o modelo e prever classe de objetos cuja classe é desconhecida. O modelo derivado pode ser representado em várias formas como: regras de classificação, árvores de decisão ou redes neurais.

A classificação pode necessitar ser precedida de análise de relevância, como tentativa de identificar atributos que não contribuem para a classificação e podem ser excluídos.

3. Estudo de Caso

Segundo Yin (2005, p. 1-101), o estudo de caso é uma das maneiras de fazer pesquisa em ciências sociais. Cada estratégia apresenta vantagens e desvantagens próprias dependendo basicamente de três condições: o tipo de questão da pesquisa, o controle que o pesquisador possui sobre os eventos comportamentais efetivos e o foco em fenômenos históricos, em oposição a fenômenos contemporâneos.

Não há fórmulas de rotina para a realização de um estudo de caso. É necessário projetar e coletar, apresentar e analisar os dados de forma imparcial. O estudo de caso inicia com a definição dos problemas ou temas a serem estudados.

O autor Yin, R.K alerta para um problema que pode surgir quando as fontes de coleta de dados são pessoas individuais ao passo que a unidade de análise do estudo de caso é organizacional. Portanto, embora as coletas de dados tenham base inteiramente em entrevistas individuais como fonte de informações, as conclusões não podem se fundamentar exclusivamente nas entrevistas.

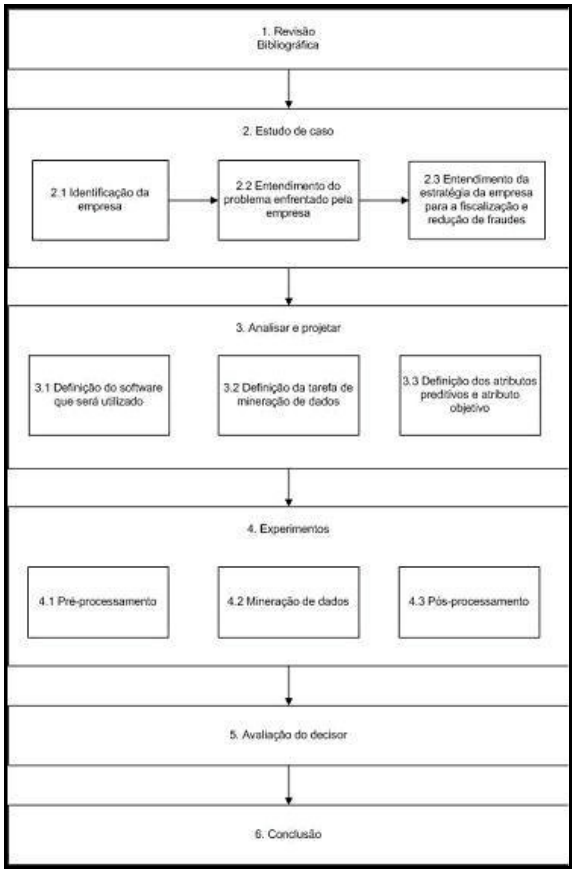

Figura 3: Desenho do modelo de pesquisa

Como fonte de coleta de dados, foi utilizado o site da empresa e entrevistas individuais com funcionários da empresa conhecedores do negócio e envolvidos no processo para o melhor entendimento do problema e o processo atual de fiscalização.

Com objetivo de explicar os passos do desenvolvimento do trabalho, foi criado um modelo de pesquisa que está representado na figura 2.

A figura 2 representa a seqüência de atividades que foram desenvolvidas no trabalho. A primeira refere-se ao levantamento bibliográfico realizado no decorrer do trabalho a fim de conhecer a técnica de mineração de dados. A segunda trata do estudo de caso realizado na empresa considerando a identificação da empresa, o entendimento do problema de perda comercial enfrentado pelas empresas do setor elétrico e a estratégia desta distribuidora de energia para a fiscalização. Na terceira atividade definiu-se a ferramenta utilizada nos experimentos, a tarefa de mineração de dados e junto ao especialista da empresa identificar os atributos relevantes para o experimento. A quarta atividade fez uso do software a partir das definições da atividade anterior. Com o resultado do pós-processamento desta atividade partiu-se para a quinta atividade onde o especialista da empresa ou decisor, avaliou o resultado. A sexta atividade descreve as conclusões deste trabalho bem como sugestões para trabalhos futuros.

3.1 A CEEE Distribuição

A Companhia Estadual de Distribuição de Energia Elétrica – CEEE-D é uma empresa de economia mista pertencente ao Grupo CEEE, concessionária dos serviços de distribuição de energia elétrica na região sul-sudeste do Estado do Rio Grande do Sul.

Com área de concessão que compreende a região Metropolitana de Porto Alegre, Litoral e Campanha gaúcha, a CEEE Distribuição atende a 72 municípios, abrangendo 73.627 km², o que corresponde

aproximadamente a 32% do mercado consumidor do Rio Grande do Sul, através de seus 47.000 km de redes urbanas e rurais, comprimento maior que o perímetro da Terra.
A CEEE Distribuição atendeu, em 2006, um total de 1.355 mil unidades consumidoras, o que equivale a cerca de 4 milhões de pessoas ou um terço da população gaúcha, distribuindo diretamente 6.287 GWh. No mesmo período a empresa investiu R$ 100 milhões em seu sistema de distribuição.
Dentre os acionistas da CEEE Distribuição, destacam-se as posições da CEEE Participações (65,92%) como *holding* controladora, e da Eletrobrás (32,59%).

3.2 O Problema

O índice de perdas comerciais na distribuidora em janeiro do ano de 2007, média dos últimos doze meses, foi de 13,92%. Em 2006, 46% das fraudes foram realizadas no medidor e 54% fora do medidor.
Irregularidades mais comuns no medidor: Manipulação na ponte de potencial externa; medidor com lacres violados ou sem lacres com manipulação interna; disco amassado ou sujeira no disco; corte do potencial interno; registrador desacoplado; eixo fora do mancal e mancal rebaixado. As irregularidades mais comuns fora do medidor: nos circuitos do sistema de medição: Fase ligada direto, inversão de fases; ligado direto no ramal de ligação; ligado direto na rede CEEE; desvio no ramal de entrada.

3.3 O Processo de Fiscalização

A CEEE-D amparada pela Resolução n.º 456/2000 da ANEEL, órgão que regulariza o setor elétrico, realiza rotineiramente inspeções nos equipamentos de medição instalados nas unidades consumidoras de sua área de concessão, com a finalidade de assegurar a qualidade e continuidade do fornecimento de energia elétrica, bem como verificar a adequação técnica e de segurança das instalações. A CEEE-D possui 33 (trinta e três) equipes de fiscalização e distribui a sua atuação em três grandes áreas: Divisão Metropolitana, Divisão Regional Litoral Norte e Divisão Litoral Regional Sul. Cada área é responsável pela fiscalização da sua região e a capacidade máxima de fiscalizações por equipe é de 15 (quinze) instalações diárias. O processo de fiscalização inicia com a Ordem de Fiscalização que é um documento emitido no sistema corporativo da empresa com algumas informações sobre a instalação que será fiscalizada. Cada equipe recebe no início do turno de trabalho 15 (quinze) ordens de fiscalizações. Caso a equipe de fiscalização identifique alguma irregularidade na instalação que está sendo fiscalizada deverá efetuar todos os procedimentos necessários para caracterização da irregularidade e regularização do fornecimento, conforme consta no artigo 72 da Resolução 456/2000 da ANEEL.

4. Experimentos

Para a realização dos experimentos, foi escolhida a tarefa de classificação de dados, pois é possível identificar o conjunto de dados de clientes que possuem irregularidade com a finalidade de descobrir alguma relação entre os atributos e o objetivo.

4.1 WEKA

O software WEKA[3] - *The Waikato Environment for Knowledge Analysis* foi escolhido para a realização dos experimentos. O Weka constitui-se em uma coleção de algoritmos de aprendizado para tarefas de mineração de dados implementados em linguagem Java. O Weka possui algoritmos para: Pré-processamento de dados; Classificação e regressão; Agrupamento; Regras de Associação; Visualização.

[3] http://www.cs.waikato.ac.nz/ml/weka/

4.2 Pré-processamento

Durante a etapa de pré-processamento dos dados não foi necessário integrar os dados porque foi utilizada uma única base de dados. Para a seleção de dados, foram considerados relevantes clientes que possuíam irregularidade. Foram analisados um total de 855 registros de clientes em situação irregular (fraudadores) da região metropolitana de Porto Alegre, referente ao período de fiscalizações do mês de abril de 2007. A tabela 1 apresenta os atributos selecionados para a etapa de mineração, sua descrição e valores distintos de cada atributo.

Tabela 1: Atributos da base de dados submetidos a mineração

Atributo	Descrição	Valores Distintos
CLASSE	Identificação da classe do consumidor.	10
TARIFA	Representa a fase que o consumidor é alimentado.	32
GERÊNCIA	Representam as gerências da empresa, que abrangem 72 municípios.	7
ATIVIDADE	Ramo de negócio do cliente.	77
ESTADO_CLIENTE	Identifica o estado do cliente com a empresa.	2
PERÍODO_IRREGULAR	Número de dias que o cliente se encontra irregular com a empresa.	Numérico
MÉDIA_CONSUMO_MENSAL	Valor médio em Khw que o cliente consome no mês.	Numérico
VALOR_FATURADO	Quantidade de energia em Khw que o cliente efetivamente pagou para a empresa durante o período irregular.	Numérico
IRREGULARIDADE	Irregularidade cometida pelo cliente.	5

O atributo irregularidade corresponde ao atributo objetivo enquanto que os demais foram utilizados como atributos preditivos.

4.3 Mineração de dados

Para realizar a etapa de mineração de dados foi utilizada a tarefa de classificação de dados, aplicando o algoritmo de classificação J48 implementado no software WEKA. A figura 3 apresenta a interface de saída da mineração de dados gerada pelo software Weka. Pode-se observar que para gerar a árvore de decisão foi utilizado o método *cross-validation* como opção de teste, aplicando dez *folds*. A figura 4 apresenta a profundidade e amplitude da árvore de decisão gerada para esta amostra de dados.

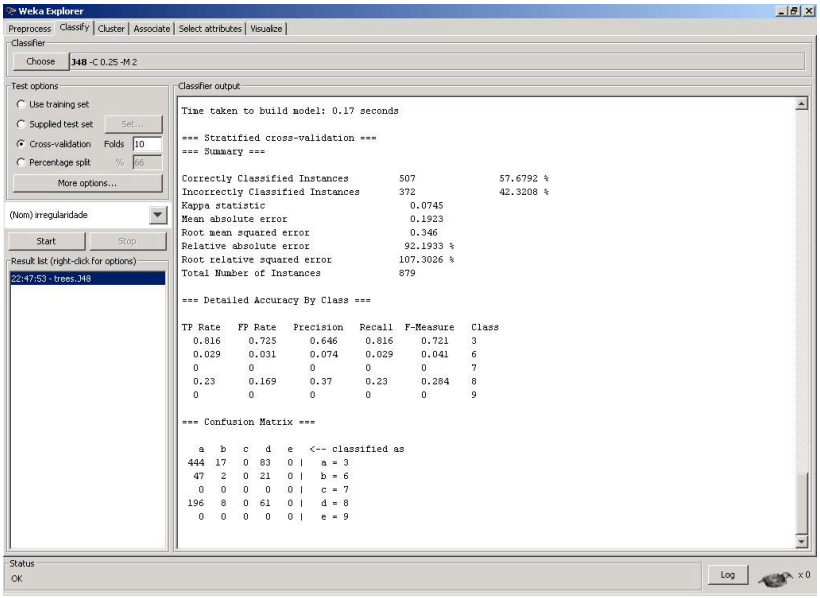

Figura 4: Resultado gerado pelo software WEKA

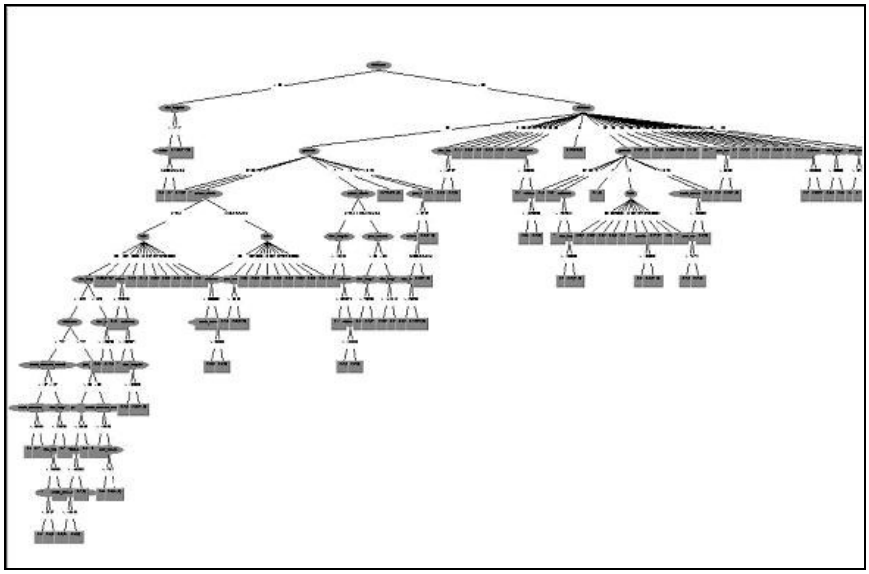

Figura 5: Árvore de decisão

4.4 Pós-processamento

Na etapa de pós-processamento as regras provenientes da interpretação da árvore de decisão gerada na etapa de mineração de dados, foi alvo de análise pelos decisores da empresa do estudo de caso. A seguir serão apresentadas as regras geradas e a avaliação realizada pelos decisores especialista no negócio da empresa.

4.4.1 Regras Geradas

Na fase de pós-processamento todas as regras geradas foram submetidas à apreciação de um especialista no negócio da empresa provedora da base de dados.

Tabela 2: Regras geradas e análise

Identificador da Regra	Corpo da Regra	Análise
R1 R2	Se dias_irregular <= 208 Então irregularidade = 6 Se dias_irregular > 208 Então irregularidade = 3	Os consumidores com fraude no medidor levam menos tempo para serem descobertos que os consumidores com fraude fora do medidor.
R3	Se atividade = 52 e Gerencia = 18 e Tarifa = 022 e Quantidade_kwh <= 1939 Então irregularidade = 8 Senão irregularidade = 3	Quando o consumidor faz o desvio, a perda de energia elétrica é maior do que a perda quando a irregularidade é feita no medidor.
R4	Se valor_faturado > 50 e Atividade = 00 e gerencia = 14 e estado_cliente = ATIVA e tarifa = 002 Então irregularidade = 8	Incidência de fraude no medidor em instalações bifásicas para o benefício da cobrança da taxa mínima de 50kwh.
R5 R6 R7 R8	Se atividade = 20 Então irregularidade = 3 Se atividade = 28 Então irregularidade = 3 Se atividade = 57 Então irregularidade = 3 Se atividade = 59 Então irregularidade = 3	As atividades do comércio varejista e consumidores de grande porte da indústria e fabricação de produtos alimentícios, metal, madeira e máquinas efetuam irregularidade fora do medidor.

3. Conclusões

Os dados referentes às fiscalizações nunca tinham sido analisados utilizando a tecnologia de mineração de dados para identificação de regras e padrões. Todas as regras obtidas com a mineração de dados apresentaram significado para a especialista do Departamento de Gestão de Perdas.

Segundo o conhecimento da especialista, nas atividades do ramo comercial e industrial estariam as maiores fraudes. Com a mineração de dados, foi possível verificar que nestas atividades a irregularidade que predomina é o desvio de energia. O desvio de energia é a irregularidade mais difícil de ser identificada pela empresa, pois geralmente está escondida em pisos, muretas, pilastras e etc. através de condutores derivativos onde é ligada a carga tipificada irregular.

A característica de consumidores predominante na área de concessão da CEEE-D é a residencial com 85%. A classe comercial possui 8% de consumidores, porém este grupo é responsável por 25% do faturamento. A classe industrial abrange somente 1% de consumidores, mas representa 23% do faturamento, já que nesta classe estão as industrias de grande porte, inclusive o Grupo A (fornecimento em Alta de Tensão). Esta informação vem reforçar a conclusão de que os maiores fraudadores pertencem às classes industrial e comercial.

Foi possível identificar um padrão para os tipos de irregularidades demonstrando que o consumidor que comete irregularidade fora do medidor permanece por um período maior irregular que o consumidor que frauda o medidor. Outra regra identificada é que a quantidade de energia fraudada é maior quando a irregularidade é fora do medidor.

O resultado obtido com a mineração de dados será utilizado na busca das unidades consumidoras no banco de dados possibilitando uma campanha direcionada ao combate às perdas comerciais. Desta forma será possível efetuar uma comparação dos resultados obtidos com a mineração de dados e os resultados encontrados em campo.

Como trabalho futuros, poderia ser seguida a mineração de dados para validar os padrões identificados neste trabalho. O atributo objetivo poderia ser tratado como irregularidade no medidor e irregularidade fora do medidor. A atividade poderia ser detalhada, pois para este trabalho foi utilizada uma visão macro desta atividade. Existe a possibilidade de criar novas relações, utilizando outros atributos preditivos e assim encontrando novos padrões no banco de dados.

4. Referências bibliográficas

Carvalho, Luís Alfredo Vidal de. Data Mining: a mineração de dados no marketing, medicina, economia, engenharia e administração. São Paulo - SP. Editora Érica, 2001, p. 1-244.

Han, J., Kamber, M. Data Mining: concepts and techniques. Morgan Kaufmann Publishers, San Francisco, USA., 2001, p. 6-24.

Scheffer, T. Finding association rules that trade support optimally against confidence. In: PKDD 2001: principles of data mining and knowledge discovery, European conference on principles of data mining and knowledge discovery N. 5, 20011973, v. 2168, 2001, p. 424-435.

Witten, Ian H.,Frank Eibe. Data Mining: practical machine learning tools and techniques with java implementations. Academic Press, 2000, p. 3.

Yin, R. K. Estudo de caso – Planejamento e Métodos. 3ª Edição. Porto Alegre, Bookman, 2005, p. 1-101.