Caracterização de Perdas Comerciais – Uma Ferramenta de Gestão de Recuperação de Receitas

Edson Amaral CPFL Paulista ; Carlos Barioni de Oliveira ENERQ-USP

Resumo **– O presente estudo mostra uma metodologia para caracterização de perdas comerciais que emprega o software estatístico SPSS e o software de mineração de dados Clementine na busca por padrões que permitam identificar unidades consumidoras com perdas comerciais. Algumas técnicas foram analisadas, entre elas, análise discriminante, regressão logística e redes neurais, que utilizaram dados cadastrais, dados de inspeções e histórico de consumo das unidades consumidoras na construção de modelos. Os padrões obtidos demonstraram excepcionais índices de acerto na identificação de unidades consumidoras com perdas comerciais.**

Palavras-chave **– Análise Discriminante, Análise de Agrupamento, Perdas Comerciais, Redes Neurais, Regressão Logística.**

I. INTRODUÇÃO

A preocupação das distribuidoras de energia com perdas comerciais é crescente. Na CPFL, os índices de perdas comerciais chegam a 3,0 % da energia comercializada. Por isso, existe grande necessidade de metodologias que aumentem a eficiência do processo de seleção das unidades consumidoras que possivelmente apresentem perdas comerciais. Atualmente, o índice de acerto nas inspeções realizadas pela CPFL está abaixo de 20%. Este trabalho procurou maximizar esse índice, de forma a obter um melhor retorno dos recursos destinados às inspeções das unidades consumidoras.

A evolução dos sistemas de informação das empresas do setor elétrico e a existência de bases de dados cada vez mais completas criam um cenário muito propício à aplicação de técnicas de mineração de dados. Essas técnicas buscam, a partir dos dados existentes na base da empresa, identificar padrões e relações que podem ser muito úteis nas tomadas de decisões dentro dos diversos setores da empresa. Em particular, a mineração dos dados cadastrais, dados de inspeções e histórico de consumo das unidades consumidoras são muito úteis na detecção de perdas comerciais. A partir desses dados e da utilização de técnicas estatísticas e de inteligência artificial, como análise discriminante, regressão logística, cluster e redes neurais, é possível criar modelos de classificação visando melhorar os índice de acertos.

II. METODOLOGIA

O diagrama esquemático da Fig. 1, apresenta as bases de dados envolvidas e os processos relacionados com a extração de dados, banco de dados de perdas comerciais, preparo dos dados, emissão de relatórios e atualização das bases de dados.

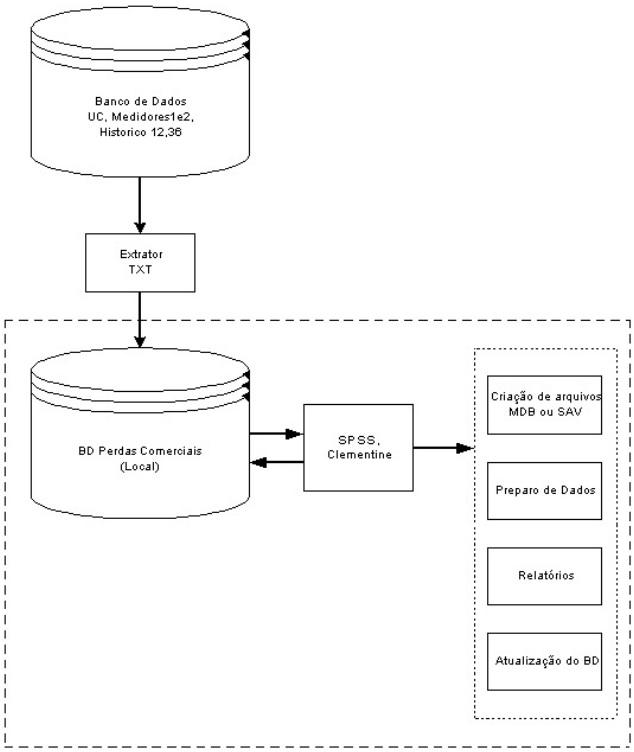

Figura 1. Diagrama Esquemático.

O banco de dados do sistema coorporativo disponibiliza uma série de informações das unidades consumidoras. Para cada unidade consumidora (UC) tem-se o seu histórico de consumos (12 ou 36 meses) como a informação mais relevante a ser analisada.

A metodologia desenvolvida utiliza técnicas de estatística e inteligência artificial para analisar o histórico de consumos das unidades consumidoras. A base de dados utilizada para se obter o modelo, corresponde à base de inspeções que possui o resultado das inspeções das UCs como sendo perdas sim (PS) ou perdas não (PN), em conjunto com a base de dados do histórico de cada UC. Após a obtenção do modelo aplica-se no banco de dados de UCs não inspecionadas.

A metodologia aplicada no Data Mining está associada com o nível de informações que se tem do problema, permitindo realizar uma descoberta supervisionada de conhecimento ou uma descoberta não supervisionada de conhecimento.

O conhecimento da área e da relação que se deseja encontrar permite realizar a Descoberta Supervisionada, para isto, um determinado campo é selecionado como alvo e o Data Mining será aplicado sobre este (Ex. Campo

PSPN, atributo que identifica a situação de perda das UC's inspecionadas).

Quando não se sabe nada sobre o comportamento do fenômeno, realiza-se a Descoberta não Supervisionada, e as técnicas Data Mining irão encontrar essas relações.

A. Extrator de Dados

Tem a finalidade de obter os dados provenientes do sistema corporativo (Mainframe). A CPFL irá disponibilizar os dados gerados nos respectivos arquivos **TXT** que são utilizados no banco de dados local.

A extração de dados gera arquivos de: Unidades Consumidoras, Histórico de Consumos Grupo A, Histórico de Consumos Grupo B, Cadastro de Medidores – Tipo 1, Cadastro de Medidores – Tipo 2, Unidades Consumidoras Inspecionadas.

B. Banco de Dados Local de Perdas Comerciais

Este banco de dados está formado por um conjunto de tabelas, com informações provenientes da importação de dados gerada pelo extrator.

Arquivos em formato .SAV são adequados para seu uso no ambiente SPSS e Clementine. Porém, arquivos em outros formatos podem ser manipulados pelas ferramentas.

C. Critérios Estatísticos

A medida de posição para tratamento dos dados considerada na metodologia é a *média*. Serve para localizar a distribuição de freqüências sobre o eixo de variação da variável estudada (consumos – kWh). De todas formas, a informação fornecida pela medida de posição necessita ser complementada pelas medidas de dispersão.

As medidas de dispersão a se considerar são: o *desvio padrão* e o *coeficiente de variação*. As medidas de dispersão servem para indicar o quanto os dados se apresentam dispersos em torno da região central, ou seja, caracterizam o grau de variação existente no conjunto de valores.

A analise de dados multivariada pode-se realizar mediante regressão logística, analise discriminante e analise agrupamentos. As duas primeiras técnicas visam obter o modelo baseadas em um banco de dados que já fornece informação em relação a uma variável de agrupamento (PS e PN). A técnica de analise agrupamentos pretende classificar as unidades consumidoras formando grupos (segundo o comportamento mostrado pelo histórico) onde se consiga diferenciá-los, seguindo um critério (maior coeficiente de variação) que permita proceder à auditagem de campo das respectivas UCs.

Regressão Logística

A regressão logística, assim como a regressão linear e múltipla, estuda a relação entre uma variável resposta (dependente) e uma ou mais variáveis independentes. A diferença entre estas técnicas de regressão se deve ao fato de que na regressão logística as variáveis dependentes estão dispostas em categorias, enquanto na regressão linear estas variáveis são dados contínuos. Outra diferença é que na regressão logística a resposta é expressa por meio de uma probabilidade de ocorrência, enquanto na regressão simples obtém-se um valor numérico.

Em regressão logística, as variáveis independentes podem ser tanto fatores quanto co-variantes; já as variáveis dependentes poderão estar dispostas em duas ou mais categorias.

Na regressão logística, a variável dependente tem duas categorias (logística binária) e o caso mais geral (mais de duas categorias de resultado) é tratado pela regressão logística multinomial. A regressão logística prediz o valor de uma variável que varia em uma escala de 0 para 1, fazendo sentido ajustar uma curva em forma de S para os dados.

Análise discriminante

A Análise Discriminante é um conjunto de técnicas estatísticas com a finalidade de alocar um elemento em uma de k populações distintas, previamente conhecidas, admitindo-se que este elemento realmente pertence a uma das k populações. O tratamento estatístico dado a esse problema de alocação reside no fato de que os dados utilizados são os valores de um conjunto de variáveis aleatórias.

Os objetivos da Análise Discriminante podem se sintetizar em dois: analisar se existem diferenças entre os grupos em quanto a seu comportamento (com respeito às variáveis consideradas e averiguar em que sentido se dão ditas diferenças) e elaborar procedimentos de classificação sistemática de indivíduos de origem desconhecido, em um dos grupos analisados.

Analise de Clusters

A análise de conglomerados ou clusters é um dos tipos de análise multivariada, baseada na proximidade dos objetos em relação a uma medida definida. Para a formação dos clusters os dados são tratados a partir das imagens geométricas, correspondentes às expressões algébricas que definem esses dados.

A análise de conglomerados pode ser feita a partir de métodos que consideram a similaridade. O método mais amplamente utilizado realiza a verificação da similaridade através das distâncias euclidianas entre os elementos.

D. Inteligência Artificial

Em relação às técnicas de inteligência artificial, utilizadas em mineração de dados encontram-se as redes neurais. Redes neurais Perceptron Multicamada e Redes de Kohonen podem ser utilizadas para este propósito.

As redes neurais são modelos computacionais não lineares, inspirados na estrutura e modo de operação do cérebro humano, com o objetivo de reproduzir

características humanas, tais como: aprendizagem, associação, generalização e abstração.

As redes neurais são muito úteis na aprendizagem de padrões a partir de dados não lineares incompletos, com ruídos compostos de exemplos contraditórios.

E. Mineração de Dados

Data Mining é o processo de descobrir informações relevantes, como padrões, associações, mudanças, anomalias e estruturas, em grandes quantidades de dados armazenados em bancos de dados ou depósitos de informação. É uma técnica de transformação de dados de baixo nível em informação de alto nível, ajudando no processo de tomada de decisões organizacionais através do uso de técnicas automáticas de exploração de dados, de forma a descobrir novos padrões e relações, que devido a esse volume não seriam descobertas a olho nu.

O Data Mining ou Mineração de Dados, na maioria das vezes, consiste basicamente na análise de dados multivariada. A mineração de dados utiliza técnicas estatísticas e da inteligência artificial para realizar a mineração em grandes massas de dados que estão representados por uma matriz de dados.

A extração de conhecimento em bases de dados é um processo complexo que envolve a formulação do problema, a preparação dos dados, a análises dos resultados e as respectivas avaliações.

As *etapas* previstas no processo de Data Mining podem ser aplicadas à maioria dos casos. Estas etapas compreendem: preparação de dados, definição do problema, descoberta das relações, analise de novas relações e avaliação dos resultados.

O resultado da aplicação do Data Mining propriamente dito, é um conjunto de novas relações descobertas mecanicamente com ajuda de programas computacionais.

As técnicas utilizadas no Data Mining são de caráter genérico e podem ser implementadas por meio de ferramentas de Inteligência Artificial e Estatística. Em um problema, são usadas as técnicas de acordo com o tipo de conhecimento que se deseja adquirir. As técnicas mais utilizadas são: Classificação, Estimação, Previsão ou Predição e Associação.

Após a escolha da técnica e da ferramenta em função do problema a analisar, é importante realizar a etapa de preparação dos dados.

F. Preparação dos dados

Devido ao volume de dados disponível num ambiente coorporativo, é importante adequar os dados para seu uso com as ferramentas de analise estatístico e de mineração de dados.

A preparação de dados tem a finalidade de limitar a base de dados e o número de variáveis envolvidas, incluir dados ausentes com um valor padrão e eliminar dados errados ou não representativos (ruídos). Estas atividades que fazem parte do processo de preparação de dados podem se resumir em: seleção, complementação e eliminação de dados.

III. SOFWARE

As ferramentas utilizadas para aplicação da metodologia são o SPSS e o Clementine. A primeira é uma ferramenta de analise estatístico que apresenta técnicas analítica eficientes. O Clementine é uma ferramenta de mineração de dados com uma série de módulos apropriados para propósitos de Data Mining.

SPSS

O SPSS e um software que possibilita o acesso e a preparação de dados, realização de analises, segmentações e desenvolvimento de diversos modelos. É composto por vários módulos opcionais: SPSS Base, SPSS Modelos de Regressão e outros. O modulo base disponibiliza as opções de analise discriminante, cluster e estatísticos e o módulo de modelos de regressão apresenta a possibilidade de uso de regressão logística.

Clementine

Clementine é um software que permite um processo interativo de data mining através da criação de modelos direcionados a cada questão especifica do problema em analise.

Possibilita a determinação de padrões e grupos através de poderosos algoritmos como: redes neurais, árvores de decisão, regras de associação, agrupamento (cluster), regressão logística, etc. Os modelos neste ambiente são construídos de maneira gráfica.

IV. EXEMPLOS

A Fig. 2 apresenta uma tela do ambiente Clementine, onde se observa o fluxo para um modelo de rede neural. As variáveis de entrada, consideradas são: *média*, *desvio padrão* e *coeficiente de variação*. Os dados correspondem a Unidades Consumidoras Residenciais Monofásicas, da região de Ribeirão Preto – SP.

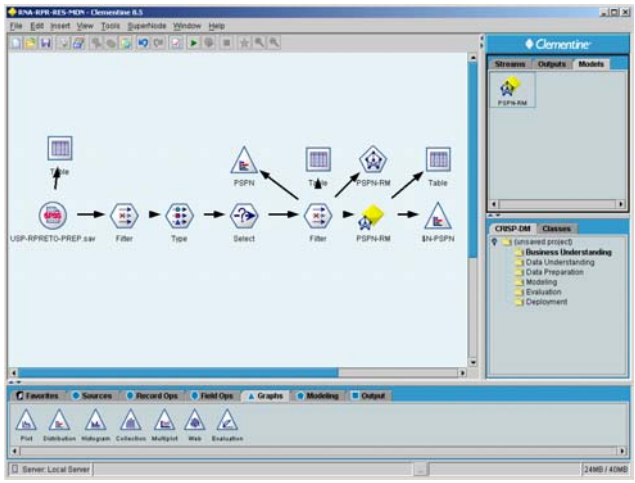

Fig. 2. Modelo utilizando Rede Neural.

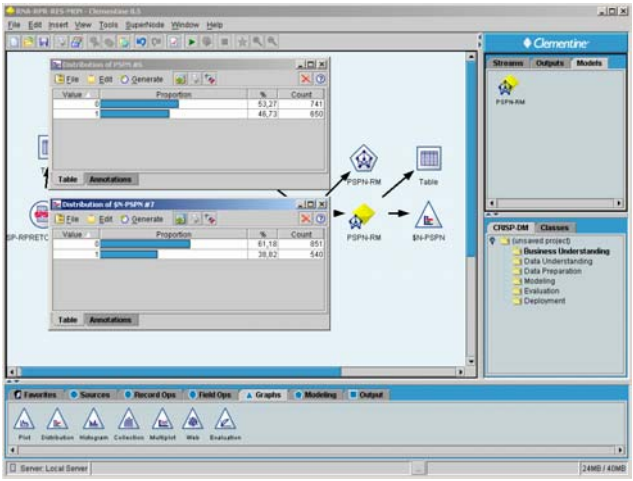

Fig. 3. Resultados da classificação.

Na Fig. 3, considerando os dados originais e o resultado do modelo tem-se:

UCs (Residencial - Monofásico)
PS (dados) 46,73 % PS (modelo) 38,82 %
PN (dados) 53,27 % PN (modelo) 61,18 %

Na Fig. 4 se observa o modelo para regressão logística, onde a diferença do modelo com relação ao modelo de rede neural está na utilização da técnica, já que as variáveis (atributos) utilizadas como variáveis independentes são as mesmas.

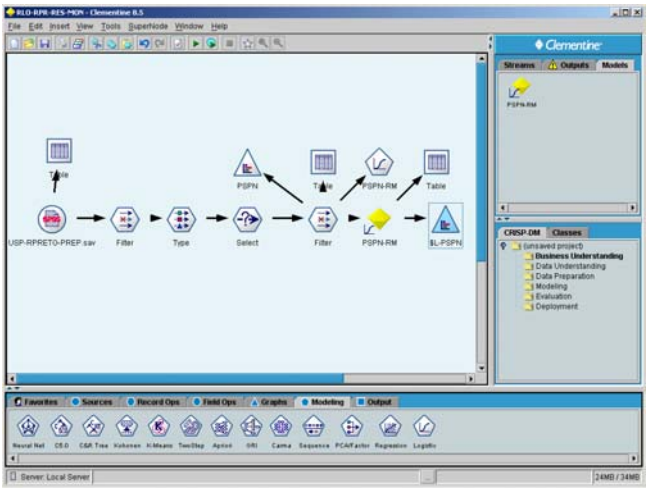

Fig. 4. Modelo utilizando Regressão Logística.

A versatilidade da ferramenta permite realizar uma série de simulações, considerando segmentos de UCs por classe de consumidor e fase ou também obter simulações para modelos integrados.

A Fig. 5 mostra os resultados da analise discriminante para uma amostra de UCs, utilizando a ferramenta SPSS. Analogamente, considerando os mesmos atributos pode se aplicar a regressão logística neste ambiente.

Classification Results[a,b]

				Predicted Group Membership		
			PSPN	0	1	Total
Cases Selected	Original	Count	0	643	98	741
			1	334	316	650
		%	0	86,8	13,2	100,0
			1	51,4	48,6	100,0
Cases Not Selected	Original	Count	0	0	0	0
			1	0	0	0
		%	0	,0	,0	100,0
			1	,0	,0	100,0

a. 68,9% of selected original grouped cases correctly classified.
b. ,0% of unselected original grouped cases correctly classified.

Fig. 5. Resultados da classificação.

V. REFERÊNCIAS BIBLIOGRAFICAS

[1]. Pedro Luiz de Oliveira Neto. *Estatística*. 2ª. Edição. Editora Edgar Blücher Ltda. 2002.

[2] Luís Alfredo Vidal de Carvalho. Datamining - *A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração* - 2ª Edição. Editora Erica.

[3] SPSS Inc. *SPSS 11.0 – Módulo Base, Conceitos e Recursos*. 2002.

[4] SPSS Inc. *Statistical Analysis Using SPSS*. 2001

[5] SPSS Inc. *Advanced Statistical Analysis Using SPSS*. 2003.

[6] SPSS Inc. *Introduction to Clementine*. 2003.

[7] SPSS Inc. *Data manipulation with Clementine*. 2003

[8] Enerq – USP. Desenvolvimento de Metodologia para Caracterização de Perdas Comerciais - Uma Ferramenta de Gestão da Recuperação de Receitas. Consolidação da Base de Informações e Efeitos Sazonais. 2004.

[9] Enerq – USP. Desenvolvimento de Metodologia para Caracterização de Perdas Comerciais - Uma Ferramenta de Gestão da Recuperação de Receitas Critérios Estatísticos. Relatório Técnico. 2004

[10] Abraham Laredo Sicsú. *Análise Discriminante*. Dissertação de Mestrado. USP. São Paulo. 1975.

[11] Rodolfo Coli da Cunha. *Proposição de metodologia para controle da qualidade de fornecimento de energia elétrica a partir da segmentação do mercado consumidor em famílias de rede elétricas*. Dissertação de Mestrado. USP. São Paulo. 2002.

[12] Renata Neves Penha. *Um estudo sobre regressão logística binária*. Universidade Federal de Itajubá. 2002.

[13] Data Mining. Em: http://www.inf.aedb.br/datamining/. Acesso: [3 – maio – 2004]