XVIII Seminário Nacional de Distribuição de Energia Elétrica

SENDI 2008 – 06 a 10 de outubro

Olinda – Pernambuco – Brasil

Extensão de Ambiente de Detecção de Perdas Comerciais Através de Análise de Características Temporais das Curvas de Consumo de Consumidores

Flávio M. Varejão
UFES – Universidade Federal do Espírito Santo
fvarejao@inf.ufes.br

Letícia R. Margoto
UFES – Universidade Federal do Espírito Santo
leticiarm@gmail.com

Guilherme M. Nogueira
UFES – Universidade Federal do Espírito Santo
g.maionogueira@gmail.com

Evandro S. Cometti
ESCELSA – Espírito Santo Centrais Elétricas S.A.
evandros@enbr.com.br

Rodrigo Ferro
ESCELSA – Espírito Santo Centrais Elétricas S.A.
rodrigoferro@enbr.com.br

Idilio Drago
UFES – Universidade Federal do Espírito Santo
idrago@inf.ufes.br

Palavras-chave
Inspeções em Campo
Mineração de Dados
Perdas Comerciais
Séries Temporais

Resumo
Um dos maiores problemas enfrentados pelas empresas de distribuição de energia elétrica no Brasil são as perdas provenientes de ligações irregulares. O presente trabalho investiga o uso de técnicas de mineração de dados temporais para auxiliar na detecção de possíveis ocorrências de procedimentos irregulares. O objetivo é aumentar as chances de sucesso nas inspeções em campo realizadas pelas empresas. Foram utilizadas técnicas de extração de informações estáticas a partir de dados temporais e técnicas de classificação especiais. Além de descrever como essas técnicas foram empregadas, esse artigo apresenta alguns resultados experimentais alcançados.

1. Introdução

Um dos grandes problemas enfrentados pelas empresas distribuidoras de energia elétrica são as perdas comerciais, que constituem o montante de energia comprado pela concessionária e não faturado de seus consumidores, descontadas as perdas técnicas. A causa mais comum desse tipo de perda são os procedimentos irregulares, entretanto erros de medição ou defeito de equipamentos também são possíveis. Os procedimentos irregulares, além de provocarem uma grande diminuição de receita para as companhias brasileiras de distribuição de energia, representam um grande risco para a segurança pública, uma vez que sobrecarregam as redes de distribuição de energia elétrica e podem causar sérios acidentes e incêndios, por vezes, fatais.

No contexto da detecção de procedimentos irregulares em energia elétrica, alguns trabalhos de melhoria no procedimento de seleção de consumidores para inspeção, utilizando alguma forma de análise computacional, já foram realizados. Eller (2003) propõe uma arquitetura de informação para o gerenciamento de perdas comerciais de energia elétrica e emprega redes neurais na tentativa de descobrir comportamentos suspeitos no perfil dos consumidores. Cabral et al. (2004) utilizam o conceito de *Rough Sets* como técnica de redução do número de atributos usados na indução de um sistema de regras de

decisão para detecção de procedimentos irregulares em consumidores de energia elétrica. Outra abordagem de classificação automática é o sistema computacional MIP (acrônimo de Mineração de Dados para Identificação de Perdas), um sistema de mineração de dados que já vem sendo utilizado pela ESCELSA - ver Cometti & Varejão (2005); Queiroga & Varejão (2005); Varejão et al. (2007). Também para auxiliar a detecção de procedimentos irregulares em clientes da ESCELSA, Perim et al. (2007) desenvolveram um sistema baseado em conhecimento denominado SAUIPE e Margoto et al. (2007), um sistema baseado em conhecimento específico para clientes com tarifação horo-sazonal de energia.

Esse trabalho descreve a aplicação de novas técnicas de inferência e classificação, mais especializadas em lidar com dados temporais, para o processo de seleção de consumidores de energia elétrica para a inspeção. O estudo envolve a investigação de técnicas de extração de informações estáticas a partir de dados temporais e técnicas de classificação especiais que podem trazer alguma melhoria em relação aos resultados atuais do sistema computacional MIP.

Esse artigo é organizado da seguinte forma: a Seção 2 apresenta o sistema MIP. A Seção 3 mostra um breve resumo sobre mineração de dados com características temporais e descreve os novos extratores criados no sistema MIP. A Seção 4 apresenta os novos algoritmos de classificação que foram incorporados. A Seção 5 contém a descrição das bases de dados usadas na validação experimental do sistema. A Seção 6 apresenta alguns experimentos realizados e os resultados obtidos. A Seção 7 discute alguns aspectos importantes relacionados aos resultados experimentais e, por fim, a Seção 8 contém as principais conclusões deste trabalho.

2. O Sistema MIP de indicação de consumidores para inspeção

Para seleção automática de consumidores para inspeção, a ESCELSA tem utilizado uma ferramenta, apresentada por Cometti & Varejão (2005), denominada MIP. O MIP é um sistema de aprendizado automático (mais precisamente de classificação) que requer como entrada dados de consumidores nos quais se sabe se há ou não procedimento irregular, e que constrói, a partir destes dados, modelos capazes de decidir se novos consumidores devem ou não ser inspecionados.

O aprendizado do sistema MIP é dito automático porque utiliza algoritmos que não necessitam da interação de usuários ou especialistas. O aprendizado é realizado exclusivamente a partir dos dados fornecidos ao sistema, em uma etapa denominada treinamento. Durante a etapa de treinamento, o sistema produzirá modelos classificadores, que são utilizados para indicação de consumidores para inspeção, em uma etapa chamada de consulta. O sistema MIP possui dois módulos principais, denominados exatamente de Módulo de Treinamento e Módulo de Consultas.

O Módulo de Treinamento permite a escolha e a configuração dos algoritmos de classificação. Dentre as técnicas disponíveis estão: redes neurais do tipo perceptron multicamadas, apresentadas por Haykin (2001); sistemas probabilísticos, como o naive bayes e as redes bayesianas, apresentadas por Mitchell (1997); indutores de regras e árvores de decisão, como o Id3, proposto por Quinlan (1986), e C4.5, proposto por Quinlan (1993); e k-vizinhos-mais-próximos, apresentado em Mitchell (1997).

A maioria destas técnicas possui parâmetros configuráveis, que interferem decisivamente no desempenho dos classificadores gerados. O sistema MIP possui algoritmos que automatizam a busca pela melhor configuração de parâmetros dos algoritmos. Além disso, o Modulo de Treinamento permite que uma estratégia de validação seja selecionada para estimar o desempenho esperado do classificador treinado, quando este for empregado na seleção de consumidores para inspeção. Validação Cruzada e Divisão Percentual são as duas principais formas de validação disponíveis no sistema.

O Módulo de Consultas permite que novos clientes sejam avaliados e que alguns destes sejam selecionados para inspeção. Além de utilizar os modelos isoladamente, também é possível selecionar consumidores através da combinação de vários modelos, por exemplo, por voto, união ou intersecção - em cada variação, um consumidor é indicado para inspeção somente se um número mínimo de modelos selecioná-lo.

O desempenho de um sistema de aprendizado automático é absolutamente dependente da qualidade dos dados apresentados ao treinamento, o que faz do tratamento dos dados disponíveis uma etapa muito importante. Nesta etapa, as características que serão utilizadas no treinamento são selecionadas, novas

características podem ser criadas a partir das inicialmente disponíveis e inconsistências nos exemplos podem ser corrigidas. O sistema MIP possui um conjunto de ferramentas de pré-processamento para preparação dos dados, que podem ser configuradas para correção de alguns problemas nos exemplos e extração de novas características a partir das inicialmente fornecidas ao sistema. A adaptação do sistema MIP para dados temporais consistiu não somente na criação de novos algoritmos de classificação, mas também na criação de novos procedimentos para extração de características não temporais, a partir das informações temporais existentes nos dados dos clientes. A próxima seção mostra um breve resumo sobre mineração de dados com características temporais e descreve os novos extratores criados no sistema MIP.

3. Mineração de dados temporais

De acordo com Box & Jenkins (1994), uma série temporal é um conjunto de dados ordenados no tempo. Por exemplo, o consumo mensal de energia elétrica de um dado cliente registrado durante um ano é um exemplo de série temporal. Segundo Roddick & Spiliopoulou (2002), a análise de séries temporais é o processo de identificação das características e propriedades importantes das séries, utilizadas para descrever, em termos gerais, o seu fenômeno gerador. Dentre os objetivos da análise de séries temporais estão descrever o comportamento, encontrar periodicidades, controlar trajetória e prever o comportamento futuro das séries.

Todos os algoritmos existentes até então no sistema MIP não estavam preparados para manipulação de características temporais. Desta forma, a principal informação disponível sobre os consumidores da ESCELSA, que é a série temporal de consumo mensal de energia, era subutilizada pelo sistema. Além disto, os especialistas da ESCELSA conhecem a forma esperada da curva de consumo de um conjunto de consumidores (normalmente chamada de curva típica), que também é representada por uma série temporal.

Até então, os classificadores produzidos com o sistema MIP haviam sido treinados com bases compostas principalmente por três tipos de informações: dados cadastrais dos clientes, o consumo mensal de energia e dados que representam um resumo do perfil de consumo da região geográfica do cliente, por exemplo, a média de consumo da rota de leitura do cliente. Exceto a curva de consumo, que é uma série temporal, todas as demais são informações estáticas, ou seja, possuem valor fixo no tempo. Em alguns casos, como a média de consumo da rota, mesmo que os valores possam mudar no decorrer do tempo, as alterações são muito lentas e as características também foram consideradas constantes no decorrer do tempo.

No caso dos dados temporais, o sistema considerava cada valor da série como uma característica independente do cliente. Embora algumas informações estatísticas, como a média e o desvio padrão do consumo, também fossem extraídas e adicionadas à base, tal estratégia despreza uma parte importante da informação contida nos dados de consumo, que é a seqüência de acontecimento dos valores. Por exemplo, o baixo consumo de energia em um mês específico pode não dizer muito sobre um determinado cliente. O mesmo não é verdade se os valores anteriores do consumo do cliente estivessem em um patamar muito mais elevado, caracterizando uma redução abrupta e inesperada do consumo.

Uma solução possível para melhor aproveitar a temporalidade dos dados é desenvolver extratores de características mais adequados a este tipo de informação e então utilizar os mesmo algoritmos já existentes no sistema MIP com as novas características extraídas. Ao todo, o MIP contém agora 22 extratores. Novas estatísticas descritivas, como a assimetria, a curtose e a mediatriz (descritas por DeGroot & Schervish (2001)) estão disponíveis no sistema. A versão anterior já continha extratores que retornavam a maior queda absoluta e percentual entre dois valores consecutivos na série de consumo. Agora, o sistema possui também um extrator que retorna a maior queda angular, medida que indica simultaneamente a magnitude e a inclinação da redução no consumo.

Além disto, novos extratores foram criados com o objetivo de comparar a série temporal de consumo à série temporal que representa a curva típica do consumidor. Estes novos extratores recebem duas séries temporais como entrada e retornam um valor que indica o quanto elas são semelhantes. A expectativa é que clientes com consumo distante do típico tenham maior probabilidade de serem unidades com procedimento irregular. Savary (2002) e Antunes & Oliveira (2001) listam diversas medidas de comparação de séries temporais que foram adicionadas ao sistema MIP. Algumas são triviais, como a Distância Euclidiana, que determina a similaridade entre duas séries acumulando a distância ponto a ponto entre elas.

Outras, como a Distância Temporal Dinâmica (DTW) ou a Distância de Edição, calculam a similaridade entre séries após remover desalinhamentos (defasagens) locais. As métricas que foram incorporadas ao MIP para comparação de séries temporais são: a Distância de Edição, a Distância de Hamming, a Maior Subseqüência Comum e a Distância Temporal Dinâmica.

Algumas outras informações que podem descrever séries temporais também estão disponíveis no sistema. Por exemplo, as Amplitudes da Transformada de Fourier e os coeficientes da Transformada Wavelet de Haar, ambas descritas em Agrawal et al. (1993), podem ser extraídas de uma série temporal qualquer. No caso das Amplitudes da Transformada de Fourier, deslocamentos integrais da série no eixo temporal não alteram os valores das amplitudes. Neste caso, as características extraídas por esta transformação são invariantes a deslocamentos e, portanto, mais adequadas aos classificadores convencionais do que os valores no domínio original.

4. Novos algoritmos de classificação adicionados ao MIP

Dois novos algoritmos foram adicionados ao sistema MIP. O primeiro, conhecido como Rough Sets, também é um algoritmo de propósito geral, como os demais já disponíveis no sistema. A inclusão deste algoritmo, porém, é relevante porque ele já tem sido usado no mesmo contexto em outra companhia elétrica, conforme descrito por Cabral et al. (2004).

Rough Sets, ou Conjuntos Incertos, é uma ferramenta matemática que visa contornar o problema de rotulação de elementos de conjuntos que possuem diferentes classes, mas que são indiscerníveis em seus atributos. No contexto deste trabalho, essa ferramenta pode ser útil para diferenciar clientes normais e irregulares que possuem os mesmos valores em suas características, situação muito comum nos dados da ESCELSA. Nesta teoria, os dados dos clientes no treinamento sofrem uma operação de fusão, na qual todos os clientes que possuem os mesmos valores para as suas características são unidos em um único perfil. Para classificar um novo exemplo (ou um novo perfil), primeiramente é verificado se este já existe no histórico de perfis gerado no treinamento. Se não existir, o perfil inédito é automaticamente classificado como normal. Se existir e todos os clientes com este perfil no treinamento forem normais, ele também é classificado como normal. Da mesma forma, se todos forem irregulares no treinamento, o novo perfil é classificado como irregular.

Para o caso mais comum, em que há tanto exemplos de irregulares quanto normais, o cliente será classificado de acordo com um parâmetro chamado corte. Se a razão entre a quantidade de clientes normais e a quantidade de clientes irregulares para um determinado perfil, for inferior ao corte, este perfil é considerado irregular. Por outro lado, se for maior ou igual ao corte, o perfil é considerado normal.

Um detalhe importante sobre este algoritmo é que ele está apto somente para manipular informações discretas. Todos os dados numéricos, mesmo os extraídos das séries temporais, devem ser discretizados. Este algoritmo tem ainda uma debilidade relacionada ao número total de características usadas no treinamento: o total de perfis possíveis cresce exponencialmente de acordo com o número de características. Assim, se o número de características for razoável (em relação ao total de exemplo do treinamento), é provável que não existam perfis com muitos elementos e que no momento da consulta apareçam apenas perfis inéditos, que serão classificados como normais. A forma usual de contornar este problema é realizar uma seleção de características antes da criação da base de perfis do treinamento. A versão disponível no MIP permite que o usuário decida se deseja ou não que uma seleção de características seja realizada.

O segundo algoritmo incluído foi chamado de Comparador de Curvas. Neste algoritmo, na fase de treinamento, são calculadas a curva média dos clientes normais e a curva média dos clientes irregulares. Normalmente a curva faturada dos clientes irregulares tem magnitude inferior que a curva dos clientes normais, dado que há desvio de energia nestes consumidores. Na etapa de treinamento é calculado também um limiar, que representa a distância máxima que o exemplo deve estar da curva média dos consumidores normais para ser considerado um caso normal. Todos os casos com perfil de consumo abaixo da curva média dos normais, além deste limiar, são selecionados para inspeção. A etapa de classificação é realizada através da comparação da distância euclidiana entre a curva a classificar e as curvas médias de normais e irregulares e o limiar definido no treinamento.

A idéia de definição de perfil médio e de estabelecimento de limiar de corte pode também ser aplicada

à base de dados com características não temporais. Imaginando que as características sejam eixos em um espaço multidimensional, o ponto médio de cada perfil será um ponto neste espaço. Supondo que os perfis sejam separáveis, estes pontos médios estarão deslocados no espaço e os exemplos de cada classe estarão mais próximos do seu ponto representativo. O limiar pode ser encarado como o limite que define a região de decisão de cada classe (no caso, hiper-esferas).

A versão do algoritmo incluído no sistema MIP aceita atributos de qualquer tipo. Além disso, considerando que alguns atributos podem possuir maior relevância que outros para a classificação, o algoritmo distribui pesos entre as características durante o treinamento, de maneira a maximizar uma métrica de classificação. Esta ponderação atua como uma seleção de características, eliminando os atributos redundantes e inúteis da base de exemplos.

5. Bases de Dados para Validação Experimental

Dois conjuntos de dados foram selecionados pela ESCELSA para validação dos novos algoritmos desenvolvidos. A primeira, formada por todos os clientes da Grande Vitória inspecionados entre junho de 2006 e junho de 2007, possui 154.152 exemplos rotulados como Normal ou Irregular de acordo com o parecer do inspetor que visitou a unidade consumidora. Nesta base, a probabilidade a priori da classe mais relevante (Irregular) é de 9,9%. Os dados rotulados desta maneira, apesar de abundantes, não representam adequadamente a situação da ligação das unidades consumidoras. De acordo com os especialistas da ESCELSA, uma parte considerável das irregularidades não é autuada nas inspeções de campo, principalmente porque os consumidores com irregularidades podem retirar as ligações irregulares, evitando o flagrante, e porque os inspetores podem não encontrar as irregularidades, por exemplo, pela sofisticação da ligação.

A segunda base de dados foi rotulada através de uma nova tecnologia de exteriorização da medição de energia, em implantação na ESCELSA. Em alguns bairros, onde a perda comercial é elevada, foi instalado um medidor externo, que registra a energia consumida pela instalação antes que os fios da ligação cheguem à medição da unidade consumidora. Nestes bairros, o medidor convencional foi mantido nas unidades consumidoras e o novo medidor externo, inicialmente, tem atuado como uma medição comparativa. Mensalmente duas medidas estão disponíveis: a medida do medidor instalado na propriedade do cliente e a medida externa, com baixa possibilidade de violação. Desta maneira, torna-se possível comparar o gasto de energia registrado pelo medidor convencional e o obtido através do medidor externo.

De acordo com a ESCELSA, há grandes chances de um cliente estar com procedimento irregular no sistema de medição se a diferença entre os valores registrados nos dois medidores for maior que 50%. Diferenças entre 30% e 50% podem ser causadas ou por irregularidades ou por problemas causados pela defasagem entre a leitura do medidor convencional e o medidor externo. Diferenças menores que 30% são desprezíveis, causadas quase que exclusivamente pela defasagem entre a leitura do medidor convencional e o medidor externo, isto é, por diferenças entre o procedimento de leitura nas duas abordagens (enquanto a leitura exterior é automática, a leitura na casa dos clientes ainda é manual).

Uma base de validação foi formada com os clientes com medição exteriorizada. Os exemplos com diferença maior que 50% foram rotulados como Irregulares e os com diferença menor que 30% como Normais (os demais clientes, de situação desconhecida, não foram aproveitados). Nesta base de validação, 39,7% dos 2.191 exemplos foram rotulados como Irregulares. Comparando as duas bases de dados de validação é evidente que a obtida pela exteriorização de medição, pelo fato de não ser afetada por manipulação humana, possui dados mais confiáveis, sendo mais adequada para o aprendizado das técnicas e para validação do desempenho teórico dos algoritmos.

6. Experimentações e resultados

Esta seção apresenta alguns experimentos realizados com o objetivo de aferir o desempenho dos novos classificadores incluídos no sistema MIP. Além disto, os novos algoritmos, e alguns dos já disponíveis no sistema, foram treinados com as novas características extraídas das séries temporais.

Para avaliar o desempenho dos classificadores foram utilizadas métricas derivadas da matriz de confusão (apresentada por Monard & Baranauskas (2002)). Cada linha desta matriz contém a distribuição dos

exemplos de uma classe de acordo com a classificação dada pelo sistema. No caso atual, há apenas duas classes: Normal (N) e Irregular (I). A Tabela 1 ilustra a matriz de confusão de nosso problema.

Tabela 1: Matriz de confusão do problema de detecção de fraudes.

		Classe Predita	
		N	I
Classe Real	N	q_{nn}	q_{ni}
	I	q_{in}	q_{ii}

A partir dos valores nas células da matriz de confusão, as seguintes medidas podem ser calculadas para representar o desempenho de um classificador:

- Taxa de acerto (a): o percentual de classificações corretas. Determinada por:

$$a = (q_{nn} + q_{ii})/(q_{nn} + q_{ii} + q_{ni} + q_{in}) \quad (1)$$

- Especificidade (e): percentual de irregulares corretamente classificados. Determinada por:

$$e = q_{ff}/(q_{ii} + q_{in}) \quad (2)$$

- Confiabilidade negativa (c): percentual de acerto dentre os classificados como Irregulares. Determinada por:

$$c = q_{ff}/(q_{ii} + q_{ni}) \quad (3)$$

- Mérito (m): média harmônica ponderada entre a especificidade e a confiabilidade negativa, com maior peso para a confiabilidade negativa. Determinada por:

$$m = e.c/(0.7e + 0.3c) \quad (4)$$

6.1. Treinamento e avaliação por validação cruzada

A partir dos dados disponíveis, duas tarefas de treinamento foram realizadas. Nos dois casos, o conjunto de características usado no treinamento foi o mesmo, selecionado através de heurísticas de busca por características, como a Seleção Seqüencial Adiante (SFS), e medidas de qualidade, como a distância de Mahalanobis (ambas são descritas por Mitchell (1997)). Do conjunto completo de características possíveis, foram selecionadas: uma característica relacionada ao perfil dos consumidores (código da tarifa), algumas relacionadas aos bairros (quantidade de ligações clandestinas, total de irregulares e total de clientes), informações extraídas da curva de consumo (estatísticas básicas, coeficientes de Fourier, máximas diferenças entre valores consecutivos etc.) e uma informação de comparação entre a curva de consumo e a curva típica (DTW).

Na primeira tarefa de treinamento, todos os registros de inspeções da Grande Vitória foram usados em validação cruzada para estimar o desempenho dos novos classificadores criados no MIP (com as características selecionadas). Além disso, para efeito de comparação, a técnica Naive Bayes também foi treinada com as mesmas características.

Na segunda tarefa de treinamento, uma parte dos registros de clientes com medição exteriorizada foi usada para treinamento e estimativa de desempenho, por validação cruzada, dos mesmos algoritmos. Nesse caso, 400 registros de um mesmo bairro (Barramares) foram reservados para o treinamento. Os demais exemplos desta base foram reservados para uma validação independente. Todos os resultados mostrados nesta seção foram obtidos por validação cruzada com 10 conjuntos. A Tabela 2 mostra a matriz de confusão e as métricas de qualidade obtidas pelo classificador Comparador de Curva na base de dados de inspeções da Grande Vitória.

A Tabela 3 mostra os resultados do algoritmo Rough Sets. A seguir, a Tabela 4 lista os resultados do algoritmo Naive Bayes.

Tabela 2: Validação do classificador Comparador de Curvas na base de clientes da Grande Vitória.

Matriz de Confusão			Métricas	
	N	I	a	0,90
N	84.665	3.281	e	0,13
I	6.504	977	c	0,23
			m	0,19

Tabela 3: Validação do classificador Rough Sets na base de clientes da Grande Vitória.

Matriz de Confusão			Métricas	
	N	I	a	0,92
N	87.946	0	e	0,00
I	7.481	0	c	0,00
			m	0,00

Tabela 4: Validação do classificador Naive Bayes na base de clientes da Grande Vitória.

Matriz de Confusão			Métricas	
	N	I	a	0,77
N	70.914	17.032	e	0,40
I	4.470	3.011	c	0,15
			m	0,19

A partir destas três tabelas é possível, inicialmente, notar que o algoritmo Rough Sets não conseguiu identificar diferenças entre os perfis, nestes dados. Isso pode ter ocorrido devido à forma de rotulação dos exemplos, que marcou muitos casos da classe Irregular como Normal, ou porque os perfis dos clientes são mesmo muito semelhantes e não há como discernir as classes somente com estas características. No caso dos outros classificadores, os resultados são opostos: o algoritmo Naive Bayes apresenta boa especificidade e o algoritmo Comparador de Curvas boa confiabilidade. Em ambos, porém, houve ganho durante o treinamento, considerando que a probabilidade a priori da classe Irregular é menor que 10%.

A seguir, as tabelas 5, 6 e 7 apresentam os resultados com a base de clientes do bairro Barramares.

Tabela 5: Validação do classificador Comparador de Curvas na base de clientes de Barramares.

Matriz de Confusão			Métricas	
	N	I	a	0,61
N	157	88	e	0,55
I	66	82	c	0,48
			m	0,50

Tabela 6: Validação do classificador Rough Sets na base de clientes de Barramares.

Matriz de Confusão			Métricas	
	N	I	a	0,54
N	108	137	e	0,72
I	42	106	c	0,44
			m	0,49

Nestas tabelas, o resultado é melhor do que o obtido com as bases da Grande Vitória, situação absolutamente esperada, dado que a probabilidade a priori da classe Irregular é de 38% (nesta amostra) e que os dados são, em geral, mais confiáveis. Além disso, o classificador Rough Sets consegue separar as classes, obtendo inclusive a melhor especificidade. O melhor classificador, no geral, foi o Naive Bayes, seguido pelo Comparador de Curvas.

Tabela 7: Validação do classificador Naive Bayes na base de clientes de Barramares.

Matriz de Confusão			Métricas	
	N	I	a	0,67
N	176	69	e	0,59
I	61	87	c	0,56
			m	0,57

6.2. Validação com base de dados independente

Uma parte dos registros com rótulo de classe determinado pelo medidor exteriorizado foi reservada para validação independente dos modelos classificadores apresentados na seção anterior. Esta base de validação contém 1.791 exemplos, que não apareceram em nenhuma das amostras do treinamento, com 39% de exemplos da classe Irregular. As tabelas 8, 9 e 10 mostram os resultados desta validação para os algoritmos Comparador de Curvas, Rough Sets e Naive Bayes treinados com os dados da Grande Vitória.

Tabela 8: Validação independente do classificador Comparador de Curvas treinado com dados da Grande Vitória.

Matriz de Confusão			Métricas	
	N	I	a	0,63
N	1.036	24	e	0,07
I	622	48	c	0,67
			m	0,19

Tabela 9: Validação independente do classificador Rough Sets treinado com dados da Grande Vitória.

Matriz de Confusão			Métricas	
	N	I	a	0,61
N	1.060	0	e	0,00
I	670	0	c	0,00
			m	0,00

Tabela 10: Validação independente do classificador Naive Bayes treinado com dados da Grande Vitória.

Matriz de Confusão			Métricas	
	N	I	a	0,64
N	924	136	e	0,27
I	490	180	c	0,57
			m	0,43

As tabelas 11, 12 e 13 mostram a validação independente dos algoritmos treinados com a amostra de registros do bairro de Barramares.

Tabela 11: Validação independente do classificador Comparador de Curvas treinado com dados do bairro Barramares.

Matriz de Confusão			Métricas	
	N	I	a	0,67
N	929	131	e	0,34
I	443	227	c	0,63
			m	0,50

Os resultados da validação com o conjunto de dados independente mostram que o aprendizado com a base da Grande Vitória é bastante comprometido pela forma de preenchimento das classes. Mesmo o algoritmo Naive Bayes, que teve um resultado aceitável treinado desta forma, apresenta desempenho bem superior quando treinado com a base de Barramares.

Tabela 12: Validação independente do classificador Rough Sets treinado com dados do bairro Barramares.

Matriz de Confusão			Métricas	
	N	I	a	0,53
N	419	641	e	0,74
I	170	500	c	0,43
			m	0,50

Tabela 13: Validação independente do classificador Naive Bayes treinado com dados do bairro Barramares.

Matriz de Confusão			Métricas	
	N	I	a	0,61
N	556	504	e	0,73
I	178	492	c	0,49
			m	0,55

Uma avaliação preliminar dos resultados da validação com as técnicas treinadas com os dados de Barramares poderia concluir que todas as técnicas tiveram resultado relativamente bom, sempre com a função de mérito acima de 50%. É importante, porém, destacar que estes resultados dificilmente seriam reproduzidos em inspeções de campo, dado que as incertezas relacionadas ao procedimento de inspeção não existem nesta base. Considerando que nesta base a classe dos exemplos é praticamente certa, o mérito de 50% mostra que o conjunto de características ainda é pobre.

6.3. Resultado do sistema nas inspeções em campo

Ao longo desse trabalho foram submetidos 4 lotes de clientes para inspeção em campo, sendo que no último lote os clientes foram selecionados pelos classificadores treinados com a diferenciação entre Normal e Irregular feita a partir dos dados da exteriorização de medição. Foram geradas 5.128 inspeções pelos classificadores treinados com a base cuja classificação dos clientes era baseada na última inspeção, sendo encontrados 4.675 resultados normais e 453 resultados com irregularidade, obtendo uma taxa de sucesso de 8,83%. Para o lote gerado pelos classificadores treinados com os dados de medição exteriorizada, foram enviadas 2.647 inspeções e encontrados 111 irregulares, uma taxa de sucesso de 4,29%. O resultado satisfatório obtido nos experimentos de laboratório não se reproduziu em campo - as incertezas que envolvem o processo de inspeção em campo podem ser apontadas como uma das razões para esse desempenho aquém do previsto, principalmente para os classificadores treinados com dados de exteriorização de medição.

7. Análise complementar dos resultados alcançados

Em relação aos resultados da validação com dados independentes, nos quais a informação da classe foi determinada pelo medidor exteriorizado, um resultado obtido pela ESCELSA ilustra o quanto o valor das métricas na validação, apesar de relativamente bom, ainda é insuficiente para reprodução em inspeções de campo.

Após certo tempo com a medição replicada, os especialistas da ESCELSA enviaram equipes de campo para inspecionar uma amostra de clientes que apresentavam diferença maior que 30% entre os dois medidores instalados. Nestas inspeções, a equipe conseguiu autuar aproximadamente 38% dos clientes inspecionados, sabendo-se previamente que há irregularidades em 100% dos casos, o que evidencia a dificuldade da irregularidade ser efetivamente registrada no momento da inspeção, mesmo em situações onde o desvio é praticamente certo.

Esse resultado reforça a conjectura de que, mesmo o cliente sendo irregular, muitas vezes, no momento da inspeção, o inspetor não consegue encontrar evidências que comprovem a irregularidade. O principal motivo, segundo avaliação da ESCELSA, é a facilidade para o cliente descaracterizar a irregularidade nas áreas selecionadas. Estas áreas são de grande complexidade social, isto é, possuem elevados índices de criminalidade, crescimento desordenado e população de baixa renda. Os principais fatores que impedem a caracterização das irregularidades e a conseqüente autuação nestas regiões são: o simples rompimento

dos lacres sem alteração da medição, redes de distribuição e ramais com isolamento violado, abertura das caixas dos medidores possibilitando acesso sem rompimento dos lacres, entre outros. Com estas possibilidades, torna-se possível o acesso à energia não medida fora do período em que a concessionária está fiscalizando sem, contudo, deixar evidências que possibilitem a autuação no momento da fiscalização.

Supondo que nas inspeções selecionadas pelo sistema MIP este tipo de perda também ocorra, o desempenho teórico dos classificadores, mostrado nos testes das seções anteriores, está superestimado. Possivelmente o resultado real em campo será menor que 40% dos valores estimados nestes testes (supondo a mesma perda). Neste caso, o classificador Naive Bayes, por exemplo, teria confiabilidade menor que 19%, valor já não satisfatório.

Além desta questão, alguns outros testes foram realizados para investigar o poder de discriminação das características. Em uma base de dados com classes confiáveis, como as usadas nos experimentos da seção anterior, espera-se que o classificador tenha desempenho muito melhor do que o obtido. Uma possível causa para o desempenho abaixo do esperado pode ser o fato das características utilizadas no treinamento não serem representativas. Para analisar essa hipótese, recorremos a duas avaliações: análise de perfis gerados por agrupamento das séries temporais de consumo e estudo da vizinhança dos exemplos da classe Irregular.

O agrupamento de dados foi utilizado com a finalidade de formar grupos de exemplos, de forma não supervisionada, nos quais há um perfil de consumo semelhante entre os componentes de cada grupo. Para esse experimento foi utilizada uma amostra da base de dados obtida pela exteriorização de medição. Nesta amostra havia 1.133 exemplos, sendo que a probabilidade a priori da classe mais relevante (Irregular) é de aproximadamente 28,2%. Apenas as séries de consumo dos clientes foram agrupadas.

Dois dos principais algoritmos de agrupamento foram utilizados: o algoritmo k-means (proposto por McQueen (1967)) e algoritmo hierárquico (proposto por King (1967)). O k-means tem como princípio formar grupos inicialmente aleatórios e realocar os elementos iterativamente, caso seja mais vantajoso. Já o algoritmo hierárquico constrói uma árvore, sendo que, a cada passo do algoritmo, os dois objetos mais próximos são sucessivamente unidos, de acordo com alguma métrica de similaridade.

Com a finalidade de comparar os grupos formados pelas técnicas de agrupamento a grupos formados aleatoriamente, um agrupamento aleatório foi realizado, considerando a mesma base e assumindo o número de consumidores por grupo igual à média que a técnica em questão (k-means ou hierárquico) obteve para os mesmo dados.

Analisando o resultado, o k-means apresentou 12% dos grupos possuindo pelo menos 70% de unidades com procedimento irregular. No experimento de agrupamento aleatório, fixando a média de exemplos por grupo em 8 (a mesma que o k-means obteve), 7,6% dos grupos continham no mínimo 70% de unidades com procedimento irregular. Nota-se, portanto, que o k-means obteve resultado levemente superior ao agrupamento aleatório. Este ganho mostra que há realmente alguma relação entre a curva de consumo e a classe Irregular, porém a diferença do algoritmo para a hipótese aleatória é muito pequena, o que evidencia que a relação entre a curva de consumo e a classe Irregular não é tão forte assim, nesta base de exemplos. O mesmo experimento foi realizado com o algoritmo hierárquico. Com essa técnica, 6,1% dos grupos continham pelo menos 70% de unidades com procedimento irregular. Novamente, foi realizado o agrupamento aleatório, agora considerando três consumidores em cada grupo (média de elementos por grupo, obtida pelo algoritmo hierárquico). Neste caso, 7,9% dos grupos são formados por pelo menos 70% de unidades com procedimento irregular. O agrupamento hierárquico obteve resultados piores do que o agrupamento aleatório o que reforça a hipótese de pouca relação entre a curva de consumo e a classe do cliente.

O segundo experimento realizado para analisar as informações das bases de dados foi um estudo da classe dos vizinhos mais próximos dos clientes identificados como Irregulares. Novamente, apenas as curvas de consumo foram utilizadas. A hipótese a ser verificada é que os exemplos da classe Irregular possuem curvas de consumo semelhantes e, por isso, os vizinhos mais próximos de um exemplo desta classe também é da classe Irregular. Neste teste foram analisadas as classes dos primeiros 9 vizinhos de todos os exemplos de unidades com procedimentos irregulares.

O experimento foi efetuado em uma amostra da base de clientes com exteriorização, contendo 1.635 clientes e com 26% de unidades com procedimento irregular. Para cada cliente com procedimento irregular, foi contabilizado o total de Irregulares dentre os 9 vizinhos mais próximos. A Tabela 14 apresenta a contagem de casos Irregulares que tinham cada valor possível do número de vizinhos (de 0 a 9).

Tabela 14: Número de vizinhos Irregulares versus total de casos

Vizinhos Irregulares	Total de casos
0	32
1	60
2	75
3	76
4	81
5	47
6	40
7	12
8	2
9	2

Esta tabela mostra que, em geral, clientes Irregulares possuem como vizinhos clientes normais. Nesta tabela, o resultado seria melhor se os casos ficassem concentrados na sua parte inferior, ou seja, se a maior parte dos primeiros 9 vizinhos de um caso Irregular também fossem irregulares. Por exemplo, se o mesmo teste fosse realizado com uma base formada por registros com classe aleatória, a maior parte dos casos teria 2 ou 3 vizinhos irregulares, o que significa que, em uma base aleatória, dentre os 9 primeiros vizinhos de um exemplo de um caso Irregular, entre 2 ou 3 também são irregulares (porque a probabilidade a priori da classe Irregular nesta base é de 26%). Estes resultados fortalecem a hipótese de que a curva de consumo, principal característica do consumidor de energia, não possui informação suficiente para determinar se o cliente é ou não irregular.

8. Conclusões

O objetivo desse trabalho foi tentar melhorar o processo de seleção de unidades consumidoras para inspeção, através da adaptação realizada no sistema computacional MIP para permitir a manipulação de informações temporais relativas aos consumidores de energia elétrica, além da criação de novos classificadores no sistema.

Um dos novos classificadores incorporados ao sistema, o classificador Comparador de Curvas, apresentou resultados levemente superiores aos demais já existentes. Porém, verificamos que as características temporais acrescentadas não contribuíram para melhorar o desempenho do sistema nas inspeções em campo.

Os resultados obtidos nas inspeções em campo realizadas a partir da seleção indicada pelo sistema MIP foram aquém do esperado. Um dos fatores que influenciaram nesse resultado é a comprovada dificuldade do procedimento irregular ser efetivamente registrado no momento da inspeção. Assim, mesmo o cliente sendo irregular, muitas vezes, no momento da inspeção, o inspetor não consegue encontrar evidências que comprovem a irregularidade. Essa questão limita o acerto dos nossos classificadores. Além desse fato, investigamos, através de alguns experimentos, a qualidade das informações extraídas das bases de dados utilizadas para treinar os classificadores e verificamos que elas possuem um baixo poder de discriminação, não sendo capazes de discernir de maneira adequada clientes com procedimento irregular de clientes normais.

Referências

Agrawal, R., Faloutsos, C., & Swami, A. (1993). Efficient Similarity Search in Sequence Databases. *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*, pages 69–84.

Antunes, C. M. & Oliveira, A. L. (2001). Temporal Data Mining: An Overview. In *Proceedings of the Workshop on Temporal Data Mining*, San Francisco, EUA. Knowledge Discovery and Data Mining (KDD 01).

Box, G. E. P. & Jenkins, G. (1994). *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 3 edition.

Cabral, J. E., Pinto, J. O. P., Gontijo, E. M., & Filho, J. R. (2004). Fraud detection in electrical energy consumers using rough sets. *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, 4:3625–3629.

Cometti, E. S. & Varejão, F. M. (2005). Melhoramento da identificação de perdas comerciais através da análise computacional inteligente do perfil de consumo e dos dados cadastrais de consumidores. Technical report, UFES - Universidade Federal do Espírito Santo, Vitória, Brasil. Relatório final de projeto de pesquisa ESCELSA/Aneel, ciclos 2003/2004.

DeGroot, M. H. & Schervish, M. J. (2001). *Probability and statistics*. Addison Wesley, 3nd edition.

Eller, N. A. (2003). *Arquitetura de Informação para o Gerenciamento de Perdas Comerciais de Energia Elétrica*. Tese de Doutorado, Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina.

Haykin, S. (2001). *Redes Neurais: Princípios e Práticas*. Bookman, 2 edition.

King, B. (1967). Step-wise clustering procedures. *Journal of the American Statistical Association*, 62(317):86–101.

Margoto, L. R., Varejão, F. M., Côgo, P. P., & Cometti, E. S. (2007). MAMFReD - Um Sistema de Auxílio à Detecção de Fraudes em Consumidores com Tarifas Horo-Sazonais de Energia. In *CBEE 2007 - II Congresso Brasileiro de Eficiência Energética*, Vitória, Brasil.

McQueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. In *5th Berkeley Symphosium on Mathematical Statistics and Probability*, pages 281–297.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

Monard, M. C. & Baranauskas, J. A. (2002). Conceitos sobre Aprendizado de Máquina. In *Sistemas Inteligentes*, chapter 04, pages 35–53. Manole, Barueri, SP, 1 edition.

Perim, G. T., Dias, H. B. P., Varejão, F. M., & Cometti, E. S. (2007). Uma Abordagem Baseada em Conhecimento para Identificação de Perdas Elétricas. In *CBEE 2007 - II Congresso Brasileiro de Eficiência Energética*, Vitória, Brasil.

Queiroga, R. M. & Varejão, F. M. (2005). AI and GIS Working Together on Energy Fraud Detection. In *Proceedings of the North American Transmission, Distribution Conference and Expo*, Toronto, Canada.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1(1):81–106.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Roddick, J. F. & Spiliopoulou, M. (2002). A Survey of Temporal Knowledge Discovery Paradigms and Methods. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):750–767.

Savary, L. (2002). Notion of Similarity in (Spatio-)Temporal Data Mining. In *ECAI'02 Workshop on Knowledge Discovery from (Spatio-)Temporal Data*, pages 63–71.

Varejão, F. M., Loureiro, S. M., Drago, I., & Cometti, E. S. (2007). Melhoramento da Identificação de Perdas Comerciais Através da Análise Computacional Inteligente dos Dados de Consumidores. In *CBEE 2007 - II Congresso Brasileiro de Eficiência Energética*, Vitória, Brasil.