X Congresso Brasileiro de Inteligência Computacional (CBIC'2011), 8 a 11 de Novembro de 2011, Fortaleza, Ceará

UM SISTEMA INTELIGENTE PARA SELEÇÃO DE UNIDADES CONSUMIDORAS COM SUSPEITA DE IRREGULARIDADES NO CONSUMO DE ENERGIA ELÉTRICA

Leonardo Conegundes Martinez[*], **Crisnamurti E. S. Vale**[†], **Bruno L. A. Alberto**[†], **Gisele L. Pappa**[*], **Renato A. C. Ferreira**[*]

[*]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
{leocm,glpappa,renato}@dcc.ufmg.br

[†]Proativa Software
{cris.svale,bruno.araujo}@proativasoft.com.br

Resumo – Esse trabalho apresenta os resultados obtidos por um Sistema Inteligente de Detecção de Irregularidades desenvolvido junto à Cemig (Companhia Energética de Minas Gerais). O principal objetivo do sistema é auxiliar na identificação das unidades consumidoras com alta probabilidade de irregularidades de consumo, e funciona em quatro etapas. A primeira delas, a de pré-processamento, inclui o tratamento dos históricos de consumo para identificação de mudanças de padrão de comportamento dos usuários, e é a maior contribuição desse artigo. O sistema faz uso de uma rede neural para aprender a diferenciar consumidores fraudadores de não-fraudadores, e utiliza dados cadastrais, de consumo e requisições de serviço do consumidor, além de um histórico de inspeções de unidades suspeitas anteriormente fiscalizadas pela Cemig. Avaliações do sistema com métricas de aprendizado de máquina e inspeções em campo mostraram um ganho de 50% em relação à metodologia atualmente utilizada pela empresa na identificação de consumidores fraudadores, gerando uma enorme economia de recursos.

Palavras-chave – Pré-processamento de dados, Redes neurais artificias, Perdas comerciais, Irregularidades no consumo de energia elétrica

Abstract – This work presents the results obtained by an Intelligent System for detecting irregularities in energy consumption, developed in partnership with Cemig. The main goal of the developed system is to help experts identifying costumers with a high probability of being fraudsters. The system works in four main steps, being data preprocess step together with the way we deal with consumption historic series the major contribution of this paper. The system uses a neural network to learn to separate fraudsters users from non-fraudsters, using data coming from four different sources, including: consumption history, service requirements, consumer unit characteristics, and a historic of inspections performed by Cemig in the past. The system was evaluated using traditional machine learning evaluation metrics, as well as a real-world field validation. The latter showed an improvement of 50% when compared to the current methodology used by Cemig to identify fraudsters users, leading to significant resource savings.

Keywords – Data preprocessing, Artificial neural networks, Commercial loss, Electrical Energy Consumption Irregularities

1. INTRODUÇÃO

Estima-se que, no Brasil, as perdas na rede de distribuição elétrica sejam responsáveis atualmente por cerca de 15% da energia comprada pelas distribuidoras [1]. Nesses 15% estão incluídas as chamadas perdas técnicas, que correspondem às perdas intrínsecas ao transporte de energia no sistema elétrico, e as perdas não-técnicas, causadas por furtos e fraudes por parte dos consumidores.

As perdas não-técnicas refletem diretamente no aumento da tarifa [2] como forma de compensar, principalmente, o montante desviado por fraudadores. As fraudes estão comumente associadas a ligações irregulares, clandestinas e alterações das características dos medidores instalados nas unidades consumidoras. Assim, existe a necessidade do aprimoramento de procedimentos e soluções tecnológicas que minimizem esta ação por parte dos usuários finais.

A Cemig é uma das distribuidoras de energia elétrica com menores índices de perdas não-técnicas do Brasil [3], comparáveis aos das melhores empresas do mundo. Atualmente, a perda não-técnica da empresa encontra-se em torno de 2,78% do montante de energia ingressada no sistema de distribuição, enquanto a média nacional situa-se em torno de 6%. No entanto, desde 2003, o volume de energia associado a perdas não-técnicas na empresa vem apresentando uma taxa de crescimento em torno de 6% ao ano, tendo totalizado 147,8 GWh em 2007, o que corresponde a cerca de R$ 108,7 milhões.

Visando coibir esse aumento da prática de irregularidades, foram implementadas várias ações, dentre as quais destaca-se o desenvolvimento de um Sistema Inteligente de Detecção de Irregularidades (SIDI), que permite à empresa identificar as unidades consumidoras com maior probabilidade de possuir irregularidades no consumo.

Este trabalho descreve o processo de concepção e implementação do SIDI junto à Cemig, financiado pelo programa de P&D da Agência Nacional de Energia Elétrica (ANEEL), cujo principal objetivo é aumentar a taxa de acerto do procedimento atualmente utilizado pela Cemig para identificação de unidades consumidoras com suspeita de irregularidade no consumo de energia, que varia aproximadamente entre 38% no interior do Estado e 48% na região metropolitana de Belo Horizonte. O sistema proposto utiliza uma rede neural artificial para gerar a probabilidade de determinada unidade consumidora possuir algum tipo de irregularidade, e conta com uma fase de pré-processamento da base de dados da Cemig, cuidadosamente implementada junto a especialistas da empresa no domínio do problema. Atualmente, o sistema inspeciona apenas unidades consumidoras atendidas em baixa tensão de distribuição, mas sua maior contribuição está na forma com que os históricos de consumo são utilizados para detectar mudanças de comportamento do consumidor juntamente com os outros atributos utilizados.

O sistema foi concebido utilizando um repositório de dados contendo informações provenientes das seguintes fontes: (i) dados cadastrais das unidades consumidoras, (ii) histórico de solicitações de serviços pelos consumidores através de canais de atendimento ao cliente, (iii) histórico de consumo em kW/h de energia elétrica de todos os clientes da concessionária e (iv) histórico de inspeções executadas pela empresa em unidades consumidoras com suspeitas de irregularidades indicadas pelos procedimentos atualmente adotados pela empresa.

A melhoria do processo de seleção de unidades consumidoras a serem inspecionadas resulta na redução das perdas comerciais e dos custos operacionais com inspeções ineficientes, uma vez que os índices de acerto na seleção de unidades consumidoras com suspeita de irregularidades obtidos pelos atuais procedimentos e sistemas adotados pelas distribuidoras brasileiras são baixos [4].

O sistema foi validado utilizando métricas para análise de desempenho de classificadores de padrões, assim como inspeções feitas em campo por técnicos da própria Cemig. Os resultados mostraram um ganho de 50% em relação à metodologia utilizada atualmente pela empresa na identificação de consumidores com suspeita de irregularidades. Isso equivale a uma redução aproximada de 150 mil inspeções/ano ineficientes, representando uma economia operacional de R$ 1,2 milhões/ano com a eliminação de serviços de campo. Outros benefícios, embora não calculados neste projeto, são o aumento da quantidade de consumidores regularizados e aumento da energia agregada ao sistema.

O restante desse artigo está organizado da seguinte forma. A Seção 2 apresenta um panorama de como o problema de detecção de irregularidades é tratado atualmente pela Cemig. A Seção 3 aborda vários estudos relacionados ao tema publicados na literatura nos últimos anos. A Seção 4 trata da base de dados utlizada no trabalho. A Seção 5 descreve o sistema desenvolvido e seus principais módulos. A Seção 6 apresenta os experimentos computacionais realizados e uma discussão sobre os resultados alcançados. Finalmente, as conclusões e direções para trabalhos futuros são descritas na Seção 7.

2. DETECÇÃO DE UNIDADES CONSUMIDORAS IRREGULARES PELA CEMIG

Como mencionado anteriormente, esse trabalho trata da identificação de unidades consumidoras que contribuem para as perdas não técnicas da CEMIG, principalmente por furto ou fraude. Atualmente, grande parte das empresas de energia elétrica apresentam métodos para seleção das unidades consumidoras baseados em conhecimentos de *experts*. Portanto, gerenciar conhecimento de especialistas em perdas não-técnicas é um processo extremamente importante para as distribuidoras de energia elétrica. O desafio inicial está na gestão do volume de dados pertinente para a tomada de decisão. Coexistem dentro das distribuidoras um grande volume de dados em múltiplos repositórios que crescem continuamente, e são muitas vezes formatados para atender requisitos específicos de cada setor da empresa: comercial (faturamento), relacionamento com cliente, sistemas de medição de consumo, entre outros.

Dessa forma, os dados contidos nesses repositórios podem ser considerados pouco confiáveis, redundantes e inconsistentes, mas são coletados em planilhas eletrônicas e formatados de acordo com quantidade e qualidade disponível. A partir dessas planilhas, os dados são tratados com o uso do conhecimento de especialistas em perdas não técnicas, de operadores matemáticos e de regras do tipo IF-THEN-ELSE, constituindo mini-sistemas especialistas, guiando a decisão sobre as unidades consumidoras a serem inspecionadas.

Essa seção apresenta a metodologia atualmente utilizada pela CEMIG para selecionar unidades consumidoras para inspeção e verificação de irregularidade. Um requerimento de inspeção pode ser gerado a qualquer momento de acordo com três mecanismos: (i) denúncia externa, (ii) denúncia de leiturista ou (iii) ativação das regras criadas por um especialista. A denúncia externa se refere ao recebimento de informações provenientes de fontes externas à distribuidora, através de um telefonema anônimo ou reclamação pelos canais de atendimento da empresa. A denúncia de leiturista é aquela efetuada pelo profissional da empresa responsável por efetuar a leitura do medidor de energia, e que no momento da leitura mensal para faturamento constatou no local alguma irregularidade que deva ser investigada por profissional habilitado para tal serviço. Já os motivos de suspeita de irregularidades gerados pela ativação de regras são aqueles que representam uma mudança no padrão de comportamento do cliente, comportamento esse pré-definido por especialistas no domínio do problema.

Essas regras são ativadas quando há uma violação da condição normal esperada para determinado parâmetro de um consumidor, caracterizando algum tipo de suspeita de irregularidade. Geralmente faz-se necessária a violação de mais de uma regra para caracterizar um comportamento anômalo de uma unidade consumidora, que leva à geração de uma inspeção. Portanto, uma ordem de inspeção pode conter um ou mais motivos, segundo critério adotado pelo especialista. Este critério pode obedecer, por exemplo, relação com a região geográfica, nível de consumo ou potencial de recuperação de receita.

Uma vez gerada a ordem de inspeção, esta é mantida como pendente, juntamente com os motivos que a geraram, até que seja executada a inspeção. O resultado de cada inspeção indica duas situações possíveis: presença de irregularidade ou ausência de irregularidade. Além disso, o resultado de uma inspeção é associado a uma série de códigos de fechamento, que definem tipos de irregularidades, situações encontradas ou serviços executados naquela unidade consumidora. Cabe ressaltar que a inspeção de uma unidade consumidora regular é mais dispendiosa financeiramente para a distribuidora de energia elétrica, pois além do gasto desnecessário em inspecionar uma unidade consumidora normal, deixou-se de inspecionar outra que pudesse apresentar irregularidades.

Hoje, ao seguir essa metodologia, a CEMIG apresenta uma taxa de acerto que varia de 38 a 48%, dependendo da região geográfica considerada. Embora esse número pareça baixo, o retorno financeiro que ele traz é enorme. Porém, um melhor gerenciamento dos dados disponíveis pode levar a ganhos ainda maiores, como mostrado nos próximas seções.

3 TRABALHOS RELACIONADOS

Diversas propostas já foram consideradas para lidar com o problema de decidir os alvos de inspeção considerando um grande volume de dados com incerteza. Como regra geral, utilizam-se técnicas de Inteligência Computacional para esse fim, como por exemplo, algoritmos de classificação, regressão linear e agrupamento [5,6]. Redes neurais estão entre os métodos mais utilizados, dada a maneira eficaz com que lidam com problemas de incerteza e não linearidade dos dados.

Além de diversos algoritmos, a seleção dos dados a serem utilizados também é outro ponto importante. Diversos trabalhos no passado consideraram conjuntos diferentes de dados obtendo taxas de sucesso variadas. Os próximos paragráfos descrevem alguns métodos e conjuntos de dados utilizados, e os respectivos resultados alcançados.

Em [7] foram analisadas as técnicas de *Two Step Cluster* e Critério de Informação de Bayes para a criação de 14 grupos, entre eles o de fraudador. Este é caracterizado por um perfil de cliente que realiza chamadas ao Call Center para serviços e informações, que apresenta serviço não regulado e que possui débito automático. Já em [8], dois algoritmos de classificação foram aplicados na detecção de consumidores fraudadores de energia: o Naïve Bayes e árvores de decisão. Estes consistem na extração de padrões de consumo em kWh de dados históricos e arranjo dos dados de várias formas, calculando a média anual, mensal, semanal e diária. Foram testadas também duas arquiteturas de Rede Neural do tipo Multi-Layer Perceptron com algoritmo de treinamento Backpropagation, onde a saída de cada rede é uma variável binária, indicando se o consumidor analisado é suspeito de fraude ou não.

Eller [9] também trata o problema da fraude utilizando redes neurais. Dois modelos de redes são propostos para classes diferentes de consumidores. Enquanto uma rede *Multi-Layer Perceptron* com o algoritmo de Backpropagation é utilizada para classificação de consumidores fraudulentos residenciais/comerciais, uma rede *Self Organizing Maps* é responsável por segmentar consumidores industriais. São utilizadas bases de dados reais da distribuidora de energia Celesc, constituindo-se de informações relativas ao histórico de consumo de seus consumidores residenciais, comerciais e industriais, além de variáveis como tipo do medidor, demanda e consumo por período.

Já Patrício [10], em sua dissertação de mestrado, apresenta uma metodologia que define perfis de comportamentos diários de unidades consumidoras de energia elétrica ligadas em alta tensão, com a finalidade de detectar fraudes ou erros de medição. A partir desta metodologia, construiu-se um sistema baseado em regras, utilizando informações estáticas (dados cadastrais) e dinâmicas (memória de massa dos medidores obtida de forma on-line através de sistemas de telemedição instalados nos clientes) sobre o consumidor. Os resultados foram considerados satisfatórios, uma vez que a taxa de acerto na identificação de fraude obtida pelo sistema, utilizando-se análise semanal na unidade consumidora, a partir da pré-seleção dos consumidores com suspeitas de fraude foi de 64,7%.

Finalmente, Todesco et al. [11] apresenta um sistema para identificação de possíveis fraudadores de energia elétrica utilizando somente informações sobre o consumo para calcular a diferença entre o consumo do mês atual e o consumo do mesmo mês no ano anterior para 12 meses, sendo acumulado no último mês. Após ajuste do sistema, a taxa de acerto para o grupo de consumidores residenciais foi de 64% e a taxa média de acerto para o grupo de consumidores comerciais (padarias, lanchonetes e postos de gasolina) foi de 80%.

4 BASE DE DADOS

Antes de descrever o sistema inteligente proposto neste trabalho, é interessante fazer uma breve descrição dos dados, já que grande parte do pré-processamento levou em consideração as características da base de dados e o conhecimento dos *experts*. Como visto na seção anterior, grande parcela dos dados utilizados nos trabalhos verificados na literatura são provenientes de históricos sobre consumo de energia elétrica mensal e inspeções executadas no passado, estas últimas trazendo também informações cadastrais tais como localização geográfica da unidade consumidora, atividade econômica, classe de consumo, número de fases e nível de tensão em que é atendida pela distribuidora. Enquanto a primeira não traz informação *a priori* quanto a presença ou ausência de irregularidade, a segunda contém uma série de informações sobre inspeções executadas e os respectivos resultados encontrados, ou seja, a presença ou ausência de irregularidade.

No estudo proposto, os dados são formados por quatro bases distintas: (i) dados cadastrais das unidades consumidoras, (ii) histórico de solicitações de serviços pelos consumidores através de canais de auto-atendimento, (iii) histórico de consumo em kW/h de energia elétrica de todos os clientes da concessionária e (iv) histórico de inspeções executadas pela empresa em unidades consumidoras com suspeita de irregularidades, e seu respectivo resultado (i.e., se a unidade inspecionada era realmente fraudadora ou não).

Base de Dados	Atributo	Descrição	Selecionado
Cadastro	Instalação	Identificação da unidade consumidora (*chave primária*)	x
	Ramo de Atividade	Ramo de atividade econômica da unidade consumidora	x
	Número de Fases	Número de fases da instalação da unidade consumidora, podendo ser monofásica, bifásica ou trifásica	x
	Classe	Faixa de consumo médio da instalação da unidade consumidora	x
	Gerência	Filial da distribuidora de energia elétrica em que a unidade consumidora está registrada	x
	Cidade	Cidade da unidade consumidora	x
	Local	Micro região geográfica que define um conjunto de unidades consumidoras	x
	Razão	Conjunto de caminhos percorridos pelo leiturista	x
	Rota	Caminho percorrido pelo leiturista	x
	Objeto de Ligação	Edificação ou propriedade conectada a rede elétrica	
Histórico de Consumo	Instalação	Identificação da unidade consumidora (*chave primária*)	x
	Consumo Mensal	60 inteiros representando o consumo mensal em KWh da unidade consumidora nos anos de 2006 a 2010	x
Histórico de Solicitações de Serviços	Instalação	Identificação da unidade consumidora (*chave primária*)	x
	Tipo de Nota	Código da nota de serviço	
	Tipo da Solicitação	Código do tipo de serviço solicitado	x
	Data	Data em que o serviço foi executado	x
Histórico de Inspeções	Instalação	Identificação da unidade consumidora (*chave primária*)	x
	Nota	Código para identificar um procedimento realizado em campo (ex: uma inspeção)	
	Data da Inspeção	Data em que foi realizada a inspeção	
	Denúncia de Leiturista	Indica se houve ou não denúncia de um leiturista quanto à detecção de uma irregularidade aparente na instalação da unidade consumidora no momento da leitura	x
	Denúncia Externa	Indica se houve ou não denúncia de irregularidade na instalação da unidade consumidora feita por fontes externas à empresa (ex: central de atendimento, ligação anônima, denúncia policial)	x
	Motivo	Código gerado pelo sistema atual de geração de motivos referente ao tipo e motivo para inspeção	x
	Data do Motivo	Data em que o motivo para a inspeção foi gerado pelo sistema atual de geração de motivos	
	Motivo Mais Antigo	Motivo mais antigo gerado pelo sistema para inspeção	x
	Código de Fechamento	Código usado no sistema para fechamento da nota de inspeção, o qual identifica a situação encontrada e/ou serviço executado na inspeção	
	Responsável pelo Cálculo	Em caso de comprovação da irregularidade, pessoa responsável pela negociação do débito junto ao consumidor	
	Mês de Referência	Mês em que foi tomada a decisão de inspecionar a unidade consumidora, sendo este o mês utilizado como referência para a derivação de novos atributos. No caso em que o sistema é utilizado para classificar unidades consumidoras ainda não inspecionadas, esse atributo deve conter a data corrente, ou seja, a data em que o sistema é executado.	x
	Resultado da Inspeção	Indica se foi encontrada fraude ou não na instalação da unidade consumidora	x
	Executor da Inspeção	Pessoa responsável pela inspeção	
	Gr Exec Inspeção	Data em que a inspeção foi disponibilizada para ser executada	

Tabela 1: Bases de dados disponibilizadas pela CEMIG

As bases de dados disponibilizadas pela Cemig contêm dados de unidades consumidoras e seus históricos de consumo e solicitações de serviços de 2006 a 2010, além do histórico de 361.980 inspeções realizadas nos anos de 2009 e 2010. Os atributos dessas bases são de natureza heterogênea, sendo representados por variáveis categóricas, numéricas e binárias. A Tabela 4 apresenta os atributos das bases de dados disponibilizadas pela distribuidora.

5. UM SISTEMA INTELIGENTE PARA DETECÇÃO DE IRREGULARIDADES

Com o objetivo de melhorar o processo de seleção de unidades consumidoras de energia elétrica atendidas em baixa tensão de distribuição com suspeita de irregularidades no consumo de energia, o presente trabalho propõe um sistema baseado em técnicas de Inteligência Computacional capaz de indicar a probabilidade de determinado consumidor da distribuidora de energia possuir ou não algum tipo de irregularidade no consumo.

A melhoria do processo de seleção de unidades consumidoras para serem inspecionadas resulta na redução das perdas comerciais e dos custos operacionais com inspeções ineficientes, haja visto os baixos índices de acerto na seleção de unidades consumidoras com suspeita de irregularidades obtidos pelos atuais procedimentos e sistemas adotados pelas distribuidoras brasileiras [4].

O sistema proposto é formado por três módulos: (i) Pré-processamento, (ii) Treinamento e Validação e (iii) Classificação. No módulo Pré-processamento, os atributos das bases de dados são pré-selecionados, normalizados e codificados, e novos atributos são derivados. No módulo Treinamento e Validação, os dados obtidos no pré-processamento são utilizados para treinar uma rede neural artificial [12], encarregada de modelar o perfil de unidades consumidoras com suspeita de irregularidade no consumo. No módulo Classificação, o melhor modelo classificador obtido no módulo Treinamento é utilizado para indicar a probabilidade de irregularidade e a classe (Normal ou Fraudador) de cada cliente da empresa que ainda não foi inspecionado.

5.1 Módulo Pré-processamento

O módulo Pré-processamento corresponde ao módulo mais importante do sistema. A literatura de aprendizado de máquina e mineração de dados deixam clara a noção de que, para gerar um bom modelo de classificação, é necessário partir de um conjunto de dados consistente e confiável [13]. O módulo de pré-processamento aqui proposto envolve cinco tarefas essenciais: seleção de atributos, limpeza de dados, extração de novos atributos, normalização e codificação binária.

5.1.1 Seleção de atributos

Na etapa de seleção de atributos, foram selecionados da base de dados completa da empresa os atributos considerados relevantes para a tarefa de classificação dos consumidores da distribuidora. A primeira fase desse processo utilizou o método χ^2 [14], e foi validada pelos especialistas da empresa no domínio do problema. A última coluna da Tabela 1 apresenta os atributos selecionados nesta fase.

5.1.2 Limpeza de dados

Na etapa de limpeza de dados são eliminados dados duplicados e tratados problemas de dados faltantes e inconsistentes. Unidades consumidoras com campos nulos ou inválidos nos atributos de cadastro ou em meses específicos do histórico de consumo (meses estes determinados em função do atributo "Mês de Referência" da base de inspeções, como explicado adiante), são eliminados.

5.1.3 Extração de Novos Atributos - Tratamento do Histórico de Consumo e Solicitações de Serviços

A etapa de extração de novos atributos envolve a derivação de atributos a partir dos dados do histórico de solicitações de serviços e do histórico de consumo. Em relação ao histórico de solicitações de serviços, para cada unidade consumidora é calculado o número total de solicitações realizadas até o mês anterior ao "Mês de Referência". Como existem mais do que 70 tipos distintos de serviços que podem ser solicitados por um consumidor da empresa, considerar um atributo para representar se houve ou não uma solicitação para cada um desses tipos acaba por tornar o processo de treinamento do modelo classificador muito lento. No sistema proposto, optamos por utilizar apenas a informação agregada do número total de solicitações de serviços realizadas até a data de referência.

Apesar da grande maioria dos sistema propostos para detecção de fraude na rede elétrica utilizar dados de históricos de consumo utilizando uma série temporal, o sistema pode ganhar muita informação interessante se dermos a essas séries temporais de consumo um tratamento mais sofisticado. A abordagem proposta aqui baseia-se na mudança de comportamento de um usuário em relação a um grupo de usuários similares. Assim, a partir do tipo da unidade consumidora (região geográfica e classe de consumo, tais como residencial, padaria ou supermercado), e sazonalidade (o padrão de consumo do inverno é diferente do verão), mudanças de comportamento são detectadas. Consideramos essa uma das grandes contribuições desse trabalho.

O método proposto inicialmente agrupava as unidades consumidoras de acordo com os atributos "Ramo de Atividade", "Número de Fases", "Local" e "Rota", de modo que duas unidades consumidoras pertencem ao mesmo grupo se e somente se possuem esses quatro atributos iguais. Esse agrupamento tem como objetivo dividir unidades consumidoras em grupos que tendem a apresentar um comportamento semelhante em relação ao consumo de energia elétrica. Mostrou-se que essa etapa de pré-processamento tem grande influência na taxa de acerto do sistema.

Em seguida, para considerar os aspectos de sazonalidade das séries temporais mensais de consumo, são calculadas quatro novas séries temporais trimestrais para cada unidade consumidora u, denotadas por c_u, m_u, pc_u e pm_u, em que $c_u[t]$ e $m_u[t]$ representam, respectivamente, o consumo médio mensal e uma média móvel exponencial [15] de ordem 4 do consumo médio mensal da unidade consumidora u no t-ésimo trimestre a partir do primeiro trimestre de 2006, inclusive.

Para definir pc_u e pm_u, vamos denotar o grupo ao qual pertence a unidade consumidora u por $cluster(u)$ e definir os multiconjuntos $C_u(t)$ e $M_u(t)$ como $C_u(t) = \{c_v[t] : v \in cluster(u)\}$, para $t = \{1, 2, \ldots, 20\}$, ou seja, $C_u(t)$ é o multiconjunto formado pelos valores dos consumos médios mensais do trimestre t de todas as unidades consumidoras pertencentes ao mesmo grupo da unidade consumidora u, e $M_u(t) = \{m_v[t] : v \in cluster(u)\}$, para $t = \{1, 2, \ldots, 20\}$. Com isso, podemos definir $pc_u[t]$ como o percentil correspondente ao valor $c_u[t]$ em relação ao multiconjunto $C_u(t)$ e $pm_u[t]$ como o percentil correspondente ao valor $m_u[t]$ em relação ao multiconjunto $M_u(t)$, para $t = \{1, 2, \ldots, 20\}$. Se $pc_u[t] \leq 10$, classificamos o consumo da unidade consumidora u no trimestre t como consumo *baixo*, se $10 < pc_u[t] \leq 90$ classificamos como consumo *médio* e se $pc_u[t] \geq 90$ classificamos como consumo *alto*. Os valores 10 e 90 foram determinados em experimentos preliminares.

Após essa classificação, podemos analisar a mudança no comportamento de consumo da unidade consumidora u no último ano em relação ao comportamento das unidades consumidoras pertencentes ao $cluster(u)$, comparando os valores de $pc_u[t]$ e $pc_u[t-4]$, em que t é o trimestre anterior ao trimestre do atributo "Mês de Referência". Dessa forma, definimos 9 novos atributos binários para cada unidade consumidora u, cada um representando uma das possíveis transições da classificação de $pc_u[t-4]$ (baixo, médio, alto) para a classificação de $pc_u[t]$ (baixo, médio, alto). O mesmo processamento é feito considerando os percentis da série temporal pm_u, gerando 9 novos atributos binários. Apenas os 18 atributos binários derivados no processo explicado acima são utilizados como entrada relacionada ao histórico de consumo no modelo classificador explicado na próxima seção.

5.1.4 Normalização e Codificação

Nas últimas etapas do pré-processamento, os dados numéricos das bases da dados são *normalizados* no intervalo fechado $[0, 1]$ utilizando a técnica de normalização *min-max* [14] e os dados categórigos são *codificados* em representação binária.

Como resultado do pré-processamento, obtém-se uma base de dados validada e consistente, adequada para ser utilizada nos módulos Treinamento e Validação e Classficação. Os dados da base da empresa associados a unidades consumidoras com registro no histórico de inspeções são utilizados para gerar a base de dados para o módulo Treinamento e Validação, enquanto os dados associados a unidades consumidoras não inspecionadas formam a base de dados para o módulo Classificação.

5.2 Módulo Treinamento e Validação

Como modelo de classificação das unidades consumidoras, utilizamos uma rede neural perceptron multicamada (MLP - *Multi Layer Perceptron*) [12], treinada com o algoritmo de aprendizagem *backpropagation* [16].

A rede possui três camadas e recebe como entrada os dados obtidos no pré-processamento, com valores normalizados no intervalo $[0, 1]$ para os atributos numéricos e valores binários codificados a partir dos atributos categóricos. No total, 91 neurônios são utilizados. A camada de saída possui exatamente um neurônio, que indica a probabilidade da unidade consumidora associada ao padrão de entrada ser fraudadora. O número de neurônios da camada intermediária é igual à média geométrica do número de neurônios da camada de entrada e da camada de saída [17].

Foi utilizada a função de ativação logística, taxa de aprendizado igual a 0.01 e momentum igual a 0.5, com a rede sendo treinada por 1000 épocas. Esses parâmetros foram obtidos em um conjunto de experimentos preliminares. Após o término do treinamento, os pesos da rede são salvos para utilização no módulo Classificação.

5.3 Módulo Classificação

O módulo Classificação é responsável por indicar a probabilidade de cada consumidor atendido em baixa tensão pela distribuidora de energia elétrica ser fraudador. Para tanto, os pesos da rede neural salvos no módulo Treinamento e Validação são carregados e a probabilidade de fraude de cada consumidor é calculada.

É importante destacar que o módulo Classificação é capaz de indicar a probabilidade de qualquer unidade consumidora ser fraudadora, incluindo aquelas que não tiveram nenhum motivo apontado pelos sistemas já utilizados pela distribuidora. Neste caso, utiliza-se um valor *default* para todos os campos relacionados às inspeções que estiverem nulos, como por exemplo, os campos "Motivo" e "Motivo Mais Antigo".

6. EXPERIMENTOS COMPUTACIONAIS E RESULTADOS

Como mencionado anteriormente, as bases de dados cedidas pela CEMIG e utilizadas nos experimentos computacionais do sistema proposto possuem 361.980 registros de inspeções realizadas nos anos de 2009 e 2010. Após a execução do pré-processamento, restaram 328.877 registros de inspeções realizadas em unidades consumidoras com dados válidos de cadastro e histórico de consumo, dos quais 253.759 (77,16%) correspondem a unidades consumidoras normais e 75.118 (22,84%) a unidades consumidoras com fraude. Desses registros, 295.975 foram separados para treinamento e validação e 32.902 para teste.

A avaliação do desempenho do modelo de classificação foi feita seguindo duas abordagens, ambas com a análise de métricas importantes para problemas de detecção de fraude derivadas de uma matriz de confusão. Na primeira abordagem, a matriz de confusão foi construída a partir dos resultados alcançados pelo modelo na classificação das unidades consumidoras do *conjunto de teste*. A segunda abordagem foi realizar *inspeções reais em campo* em unidades consumidoras ainda não inspecionadas.

A Tabela 2 ilustra a matriz de confusão usualmente utilizada em problemas de detecção de fraudes com duas classes, denotadas na matriz por F (Fraudador) e N (Normal). A entrada q_{ff} representa a quantidade de padrões fraudadores classificados corretamente como fraudadores. A entrada q_{fn} representa a quantidade de padrões fraudadores classificados incorretamente como normais. A entrada q_{nf} representa a quantidade de padrões normais classificados incorretamente como fraudadores. A entrada q_{nn} representa a quantidade de padrões normais classificados corretamente como normais.

		Classe Predita	
		F	N
Classe Real	F	q_{ff}	q_{fn}
	N	q_{nf}	q_{nn}

Tabela 2: Matriz de confusão usualmente utilizada em problemas de detecção de fraudes com duas classes

A métrica mais utilizada na avaliação de classificadores é a taxa de acerto (a), definida como $a = (q_{ff} + q_{nn})/(q_{ff} + q_{fn} + q_{nf} + q_{nn})$. No entanto, para problemas de detecção de fraudes, devido ao desbalancemaento das classes - existem muito mais unidades consumidoras que não representam fraudes do que o contrário, duas outras métricas são mais importantes e utilizadas [18].

A primeira delas é a precisão (p), definida como $p = q_{ff}/(q_{ff} + q_{nf})$, ou seja, a razão entre o número de fraudadores corretamente classificados e o número total de consumidores que foram classificados como fraudadores. Essa métrica é importante já que fornece uma noção de confiança do modelo classificador, apontando o percentual de acertos na identificação dos clientes fraudadores.

A segunda métrica é a revocação (r), definida como $r = q_{ff}/(q_{ff} + q_{fn})$, ou seja, a razão entre o número de fraudadores corretamente classificados e o número total de consumidores que são fraudadores. Essa métrica é importante já que fornece uma noção de cobertura do modelo classificador, apontando o percentual do conjunto dos consumidores fraudadores que está sendo identificado pelo sistema.

Para a definição do modelo de classificação foi utilizado o método de validação cruzada com 10 partições. Nesse método, a base de dados é dividida em 10 partes de mesmo tamanho, e dez modelos de classificação são treinados e validados. Para cada

		Classe Predita	
		F	N
Classe Real	F	5499	3894
	N	3252	20257

Tabela 3: Matriz de confusão da classificação para a base de teste, com $\alpha = \beta = 0,5$

		Classe Predita	
		F	N
Classe Real	F	888	557
	N	166	7512

Tabela 4: Matriz de confusão da classificação para a base de teste, com $\alpha = 0,1$ e $\beta = 0,9$

		Classe Predita	
		F	N
Classe Real	F	376	170
	N	224	565

Tabela 5: Matriz de confusão da classificação para as inspeções de campo

um desses modelos, uma das 10 partes é utilizada como conjunto de validação enquanto as outras nove partes são utilizadas como conjunto de treinamento. O modelo que apresentou o melhor desempenho na classificação do conjunto de validação foi utilizado para indicar a probabilidade de fraude das unidades consumidoras do conjunto de testes e das unidades consumidoras que ainda não haviam sido inspecionadas.

A partir das probabilidades de fraude indicadas para cada consumidor do conjunto de teste, podemos classificá-los como fraudador ou normal. Para tanto, é necessário definir dois valores limiares α e β, indicando que consumidores associados a uma probabilidade inferior a α devem ser classificados como normais e consumidores associados a uma probabilidade superior a β devem ser classificados como fraudadores.

A abordagem usualmente utilizada em problemas de detecção de fraudes consiste em assumir $\alpha = \beta = 0,5$. A Tabela 3 apresenta a matriz de confusão obtida com estes valores de α e β para o conjunto de teste com 32.902 registros de inspeções, dos quais 23.509 são registros de inspeções realizadas em unidades consumidoras normais e 9.393 em unidades consumidoras fraudadoras. A taxa de acerto foi igual a $78,28\%$, a precisão igual a $62,84\%$ e a revocação igual a $58,54\%$.

Apesar desde já ser um resultado bastante satisfatório, superior aos alcançados atualmente pelos sistemas da distribuidora de energia elétrica, que apresentam a precisão próxima de 40%, ele é pessimista em relação ao processo real de realização de inspeções da companhia. Isso porque estamos considerando o resultado que seria obtido se todos os clientes com probabilidade acima de $0,5$ fossem inspecionados. No entanto, considerando o grande número de consumidores da companhia, a limitação do número de equipes disponíveis para realizar inspeções e o elevado custo operacional destas, só é possível inspecionar um subconjunto bem restrito dos consumidores da empresa. Dessa forma, uma abordagem mais apropriada para a classificação consiste em considerar como fraudadores (selecionar para inspeção) apenas aqueles consumidores com probabilidade de irregularidade próxima de um, e como normais apenas aqueles consumidores com probabilidade próxima de 0, desconsiderando na análise unidaedes consumidoras com probabilidade entre $0,1$ e $0,9$.

A Tabela 4 mostra a matriz de confusão obtida com os valores de $\alpha = 0,1$ e $\beta = 0,9$ para os registros do conjunto de teste. O número de unidades consumidoras normais e fraudadoras consideradas na análise foi igual a 7.678 ($32,66\%$ do valor original) e 1.445 ($15,38\%$), respectivamente, totalizando 9.123 ($27,73\%$) consumidores. A taxa de acerto foi igual a $92,07\%$ e a precisão igual a $84,25\%$, valores significativamente superiores aos obtidos com a primeira abordagem. A revocação obtida foi de $61,45\%$, considerando apenas as unidades consumidoras mantidas sob análise. No entanto, observamos que apenas $9,45\%$ das 9.393 unidades consumidoras fraudadoras da base de teste foram detectadas por essa abordagem, uma vez que mais de 84% delas foram desconsideradas da análise.

A segunda avaliação consistiu nas inspeções de campo, realizadas no primeiro trimestre de 2010 em 1.335 unidades consumidoras. A Cemig decidiu inspecionar tanto unidades classificadas como fraudadoras quanto unidades classificadas como normais, com o intuito de avaliar o desempenho do modelo na caracterização das duas classes. Para tanto, foram selecionadas regiões geográficas onde poderiam ser realizadas inspeções, e as unidades consumidoras dessas regiões com maior e menor probabilidade de irregularidade foram inspecionadas de acordo com as restrições do tamanho das equipes de inspeção.

No total, foram inspecionadas 600 unidades consumidoras classificadas pelo modelo como fraudadoras e 735 unidades consumidoras classificadas como normais. A Tabela 5 mostra a matriz de confusão obtida a partir das inspeções de campo. A taxa de acerto foi igual a $70,49\%$, a revocação igual a $68,86\%$ e a precisão igual a $62,67\%$, representando uma melhora na efetividade das inspeções realizadas pela distribuidora de energia elétrica superior a 50%, uma vez que a confiabilidade positiva do sistema atualmente utilizado pela companhia para este mesmo conjunto de inspeções foi igual a $40,90\%$.

7. CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho descreveu o processo de implementação de um Sistema Inteligente de Detecção de Irregularidades junto à Cemig. O principal objetivo do sistema é melhorar o índice de acerto do processo de seleção de alvos de potenciais consumidores fraudadores. A robustez do sistema evita que a empresa envie desnecessariamente a campo profissionais para fazer uma verificação presencial de fraude nos medidores de energia.

O sistema proposto trabalha em três fases, sendo o módulo de pré-processamento o mais elaborado. Em especial, a fase de pré-processamento de históricos de consumo, permitindo a identificação de mudanças de padrão de comportamento de unidades consumidoras similares, é uma das grandes contribuições desse artigo. Após pré-processados, os dados são utilizados para treinar e validar uma rede neural artificial, que gera a probabilidade de uma unidade consumidora ser fraudadora.

O método utiliza dados contendo informações de quatros tipos, incluindo dados cadastrais, histórico de solicitações de serviços, histórico de consumo e histórico de inspeções executadas pela empresa em unidades consumidoras com suspeita de

irregularidades. As probabilidade geradas pela rede neural foram transformadas em fraude ou não fraude, e os resultados avaliados utilizando métricas de aprendizado de máquina e verificação em campo. Os resultados mostraram um ganho de 50% em relação à metodologia atualmente utilizada pela empresa na identificação de consumidores fraudadores, gerando uma enorme economia de recursos para a empresa.

O sistema foi projetado para indicar inspeções apenas em unidades consumidoras de baixa tensão. No futuro, esse método pode ser estendido para unidades consumidores de média ou alta tensão. Além disso, pesquisas relativas ao balanceamento de classes de consumidores não-fraudadores e fraudadores precisam ser melhor estudadas, para melhorar os resultados da classificação. Como a diferença do número de exemplos em cada classe é grande, sistemas de aprendizado podem encontrar dificuldades em induzir o conceito relacionado à classe minoritária.

Agradecimentos. Os autores agradecem a ANEEL e a CEMIG pela ajuda financeira e disponibilidade dos técnicos Luiz R. F. Rios e Leandro L. G. Ribeiro. Agradecem também a empresa PROATIVA durante as etapas de desenvolvimento, implantação e testes em campo do sistema computacional.

REFERÊNCIAS

[1] Agência Nacional de Energia Elétrica, "Nota Técnica 290/2008-SRE/ANEEL", 2009.

[2] Agência Nacional de Energia Elétrica, "Resolução ANEEL no. 456 - Condições Gerais de Fornecimento de Energia Elétrica", 2000.

[3] Associação Brasileira das Concessionárias de Energia, "Contribuições para o processo de revisão tarifária da Cemig", CP 004/2009, 2009.

[4] P. E. M. Almeida, R. L. Durães and B. L. A. Alberto. "Inteligência Computacional nas distribuidoras de energia elétrica: evolução tecnológica, aplicações e impactos na redução das perdas não-técnicas". *XVIII Seminário Nacional de Distribuição de Energia Elétrica*, 2008.

[5] C. Tahan. "Desenvolvimento de Sistema de Estimativa de Consumo para Recuperação de Receitas". *IV CITENEL - Congresso de Inovação Tecnológica em Energia Elétrica*, 2007.

[6] A. P. Braga, A. C. P. L. F. Carvalho and T. Ludermir. *Redes Neurais Artificiais: teoria e aplicações*. Livros Técnicos e Científicos (LTC), 2a. Edição, 2007.

[7] E. Francisco, A. Petrielli and C. Reina. "Segmentação comportamental de clientes para o setor elétrico". *Congresso anual de tecnologias da informação da Fundação Getúlio Vargas*, 2006.

[8] A. Souza, P. da Silva, A. Oltremari, M. Zago, F. Amaral and P. C. Jr. "Desenvolvimento de um Sistema Especialista para Detecção de Pontos Potenciais de Perdas Comerciais". *IV Congresso de Inovação Tecnológica em Energia Elétrica - CITENEL*, 2007.

[9] N. A. Eller. "Arquitetura de informação para o gerenciamento de perdas comerciais de energia elétrica". *Tese de Doutorado, Universidade Federal de Santa Catarina, Departamento de Engenharia de Produção e Sistemas*, 2003.

[10] M. Patrício. "Detecção de fraude ou erro de medição em grandes consumidores de energia elétrica utilizando Rough Sets baseado em dados históricos e em dados em tempo real". *Dissertação de mestrado, Universidade Federal de Mato Grosso do Sul*, 2005.

[11] J. Todesco, L. Garbeloto, A. Morales, E. Athayde and S. Rautenberg. "Aplicação de técnicas de Mineração de Dados para detecção de fraudes de energia". *Anais do IV CITENEL Congresso de Inovação Tecnológica em Engenharia Elétrica*, pp. 1–8, 2007.

[12] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1998.

[13] D. Pyle. *Data preparation for data mining*. Morgan Kaufmann, 1999.

[14] P. Tan, M. Steinbach and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 2005.

[15] Y. S. Moon and J. Kim. "Efficient moving average transform-based subsequence matching algorithms in time-series databases". *Information Sciences*, vol. 177, no. 23, pp. 5415–5431, 2007.

[16] S. Russel and P. Norvig. "Artificial Intelligence: A Modern Aproach." 2009.

[17] D. Bourg and G. Seemann. *AI for Game Developers*. O'Reilly Media, 2004.

[18] R. C. Prati, G. E. A. P. A. Batista and M. C. Monard. "Uma experiência no balanceamento artificial de conjuntos de dados para aprendizado com classes desbalanceadas utilizando análise ROC". *Anais da IV Jornada Chilena de Computação, IV Workshop de Inteligência Artificial*, pp. 1–10, 2003.