SENDI 2004
XVI SEMINÁRIO NACIONAL DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

Sistema de Detecção de Fraudes em Consumidores de Energia Elétrica Baseado em Rough Sets

José E. Cabral – UFMS, João Onofre Pereira Pinto – UFMS
José Reis Filho – ENERSUL/SA, Edgar M. Gontijo – ENERSUL/SA

E-mail: jcabral@nin.ufms.br

Palavras-chave – Detecção de fraudes, descoberta de conhecimento em banco de dados (KDD), computação flexível, rough sets.

Resumo – Este artigo descreve a teoria e a aplicação de rough Sets na detecção de fraudes em unidades consumidoras de energia elétrica, a partir de banco de dados. O conceito de reduto em rough sets foi usado para remover atributos condicionais e o algoritmo da decisão mínima (MDA) foi aplicado para remover valores insignificantes de atributos condicionais. O banco de dados minimizado aprendeu o comportamento dos consumidores, permitindo ao sistema de regras de classificação predizer perfis de consumidores fraudulentos. Os resultados obtidos foram bons o suficiente para demonstrar que rough sets é uma técnica poderosa para este tipo de problema.

Abstract – This article describes the theory and application of rough sets in fraud detection in electrical energy units consumers from databases. The rough sets concept of reduct was used to remove conditional attributes and the minimal decision algorithm (MDA) was used to remove insignificant classes of each conditional attribute. The minimized database approach the consumers behavior, allowing a classification rule system to predict fraud consumers profiles. The achieved results are good enough to demonstrate that rough sets is a very powerful technique for this type of problem.

1 INTRODUÇÃO

A recuperação de perdas de receitas ocasionadas por fraudes é essencial para manter o equilíbrio financeiro do caixa das empresas distribuidoras de energia elétrica. Porém, a identificação das unidades consumidoras com comportamento fraudulento é uma tarefa complexa. Normalmente, esta tarefa envolve inspeção *in loco*. Considerando-se o elevado número de unidades consumidoras e a não-linearidade do problema, os custos envolvidos assumem valores inviáveis. Esta inviabilidade ocorre porque geralmente tais inspeções são feitas aleatoriamente, ou a partir da experiência do responsável, ou seja, não existe nenhum sistema automático que possa indicar a probabilidade de um determinado consumidor estar fraudando. Como resultado disso, o número de fraudes detectadas na inspeção é muito baixo comparado com o número total de inspeções. O percentual de acerto, de maneira geral, chega a menos de 5%.

Por outro lado, é sabido que sistemas inteligentes de classificação, baseados em Computação Flexível (Soft-Computing), são empregados nas mais diversas áreas, comerciais e acadêmicas, na construção de sistemas de suporte a tomada de decisão. Os resultados oriundos de tais sistemas têm se mostrado bastante satisfatórios.

Na literatura foram reportados muitos trabalhos utilizando técnicas de computação flexível na detecção de fraudes em cartões de crédito. Dentre as técnicas utilizadas, destacam-se Redes Neurais Artificiais [1] e Lógica Nebulosa [2].

Rough sets é uma técnica emergente de computação flexível que vem sendo usada em muitas aplicações de descoberta de conhecimento em banco de dados, como por exemplo na determinação de regras de classificação [3]. No entanto, apesar do seu potencial, tal técnica tem sido preterida para problemas de detecção de fraude.

Este trabalho aborda a teoria e a aplicação de rough sets na detecção de fraudes em unidades consumidoras de energia elétrica, a partir de banco de dados. Inicialmente é feita um breve descrição da Teoria de rough sets, abordando os principais conceitos. Na seqüência, a solução na detecção de fraudes à partir de banco de dados é apresentada e finalmente são dados os resultados do sistema.

2 TEORIA DE ROUGH SETS

A teoria de rough sets foi proposta por Zdzislaw Pawlak na década de 80 [4]. Ela aborda basicamente a análise de tabelas (ou banco de dados) com o objetivo de aproximar conceitos e informações contidas nesses repositórios. Muitas vezes estas informações são imprecisas ou incertas, necessitando de métodos ou algoritmos para serem determinadas. Este motivo justifica a grande aplicabilidade da teoria de rough sets na descoberta de conhecimento em banco de dados. Alguns conceitos de rough sets devem ser apresentados para melhor consolidar sua teoria.

2.1 Sistema de Informação e Decisão

Um conjunto de dados é representado por uma tabela. As linhas representam os objetos (exemplos) e as colunas os atributos. Cada objeto caracteriza-se pelos valores de atributos que possui. Esta tabela é chamada sistema de informação [5]. Formalmente, o sistema de informação é definido por $\mathcal{A}=(U, A)$ onde U é um conjunto finito e não vazio de objetos e A é um conjunto finito e não vazio de atributos. Os sistemas de informação vêm geralmente acompanhados de outra informação, a classificação do objeto. Esta classificação é representada por outra coluna de atributo. O sistema de informação complementado com este atributo de classificação é chamado de sistema de decisão. Ele é definido por $\mathcal{A}=(U, A\cup\{d\})$, onde $d \notin A$ é o atributo de decisão. Os demais atributos de A são chamados de atributos condicionais. Um sistema de informação e decisão é ilustrado na Tabela 1.

2.2 Reduto

Considerando o conjunto A da Tabela 1, todos os elementos pertencentes a U são distintos. Ou seja, considerando os atributos *Tipo de Ligação, Classe e Média de Consumo*, o conjunto U é particionado nos subconjuntos elementares *{e1}, {e2}, {e3}, {e4}, {e5}, {e5} e {e6}*. Agora, considerando o subconjunto {*Tipo de Ligação,Classe*} de *A*, o conjunto U é particionado nos subconjuntos *{e1,e2,e3}, {e4,e6} e {e5}*, que são subconjuntos não-elementares. Sendo assim, somente os atributos *Tipo de*

Ligação e *Classe* não conseguem discernir todos exemplos da Tabela 1. Porém, o subconjunto {*Tipo de Ligação,Média de Consumo*} pode particionar o conjunto *U* em subconjuntos elementares. Somente os atributos *Tipo de Ligação* e *Média de Consumo* podem discernir todos exemplos da Tabela 1. Então, conclui-se que o atributo *Classe* é *redundante*. O conjunto *P*={*Tipo de Ligação,Média de Consumo*} não contém atributos redundantes e é chamado *reduto* do conjunto *A*.

Formalmente, o conjunto de atributos *P* é reduto de *A* se $P \subseteq A$ mantém as relações de discernibilidade de *A*. Em outras palavras, se *P* tem cardinalidade menos ou igual a *A* e pode representar todos elementos de um sistema de decisão, então *P* é um reduto de *A*. Considerando o reduto *P*={*Tipo de Ligação,Média de Consumo*}, um novo sistema de decisão é mostrado na Tabela 2. Embora a Tabela 2 mostre uma redução (à partir do reduto) do sistema de decisão da Tabela 1, redutos não são necessariamente únicos. Pode existir mais de um reduto para um dado sistema de informação qualquer[5].

Tabela 1 – Sistema de informação (cinza) e decisão (toda a tabela).

Cliente	Tipo de Ligação	Classe	Média de Consumo	Fraude
e1	1	1	Normal	Não
e2	1	1	Alta	Sim
e3	1	1	Baixa	Sim
e4	2	1	Normal	Não
e5	2	2	Alta	Não
e6	2	1	Baixa	Sim

Tabela 2 – Sistema de decisão considerando o reduto.

Cliente	Tipo de Ligação	Média de Consumo	Fraude
e1	1	Normal	Não
e2	1	Alta	Sim
e3	1	Baixa	Sim
e4	2	Normal	Não
e5	2	Alta	Não
e6	2	Baixa	Sim

Esta redução em sistemas de decisão é mais relevante quando o mesmo possui muitos atributos condicionais. Encontrar os redutos é um dos gargalos da teoria de rough sets. Porém, heurísticas baseadas em algoritmos genéticos computam os redutos com menor tempo computacional [5].

2.3 Aproximações

Analisando os atributos de decisão em um sistema de decisão, encontra-se o conjunto dos conceitos. Ele nada mais é que o conjunto dos possíveis valores de classificação que um elemento pode ter. Para o sistema de decisão das Tabelas 2, o conjunto de conceitos é *{Sim,Não}*, informando se o elemento é classificado como fraudador ou não. Considerando a Tabela 2, os elementos de *U* estão bem definidos. Para ilustrar uma situação problemática, será adicionado a Tabela 2 mais dois elementos, dando origem a Tabela 3. Os conceitos da Tabela 3 são representados pelos subconjuntos *{e1,e4,e5,e8}* e *{e2,e3,e6,e7}*. Porém, os elementos *e5* e *e7* têm classificação diferente e possuem os mesmos valores de atributos condicionais. O mesmo acontece com os exemplos *e6* e *e8*. Para tentar contornar esse problema, rough sets define três subconjuntos de *U*.

Seja *X* um conceito de um sistema de decisão. Pode ser encontrado um subconjunto de *X* com exemplos que *com certeza* pertençam ao conceito *X*. Este subconjunto é chamado *aproximação inferior* de *X*, ou simplesmente $\underline{X}$. Considerando a Tabela 3, se *X={e1,e4,e5,e8}*, então $\underline{X}$ *={e1,e4}*. Similarmente, se *X={e2,e3,e6,e7}*, então $\underline{X}$ *={e2,e3}*. Note que sempre $\underline{X} \subseteq X$.

A *aproximação superior* de *X*, ou simplesmente $\overline{X}$, corresponde ao subconjunto de *U* com exemplos que *podem* pertecer ao conceito *X*. Considerando a Tabela 3, se *X={e1,e4,e5,e8}*, então $\overline{X}$ *={e1,e4,e5,e6,e7,e8}*. Similarmente, se *X={e2,e3,e6,e7}*, então $\overline{X}$ *={e2,e3,e5,e6,e7,e8}*. Note que sempre $X \subseteq \overline{X}$.

A região de fronteira de *X*, ou simplesmente *BX*, corresponde a um subconjunto de *U* com exemplos que pertencem a $\overline{X}$, mas não pertencem a $\underline{X}$, ou seja, $BX = \overline{X} - \underline{X}$. Se *BX* é vazio, então $\overline{X}$ e $\underline{X}$ possuem os mesmo elementos; em outras palavras, o sistema de decisão, neste caso, não contém elementos inconsistentes. Consequentemente, quanto maior a cardinalidade de *BX*, maior é a indiscernibilidade entre os conceitos. A Figura 1 ilustra a distribuição das aproximações para o sistema de informação da Tabela 3.

Tabela 3 – Sistema de decisão inconsistente.

Cliente	Tipo de Ligação	Média de Consumo	Fraude
e1	1	Normal	Não
e2	1	Alta	Sim
e3	1	Baixa	Sim
e4	2	Normal	Não
e5	2	Alta	Não
e6	2	Baixa	Sim
e7	2	Alta	Sim
e8	2	Baixa	No

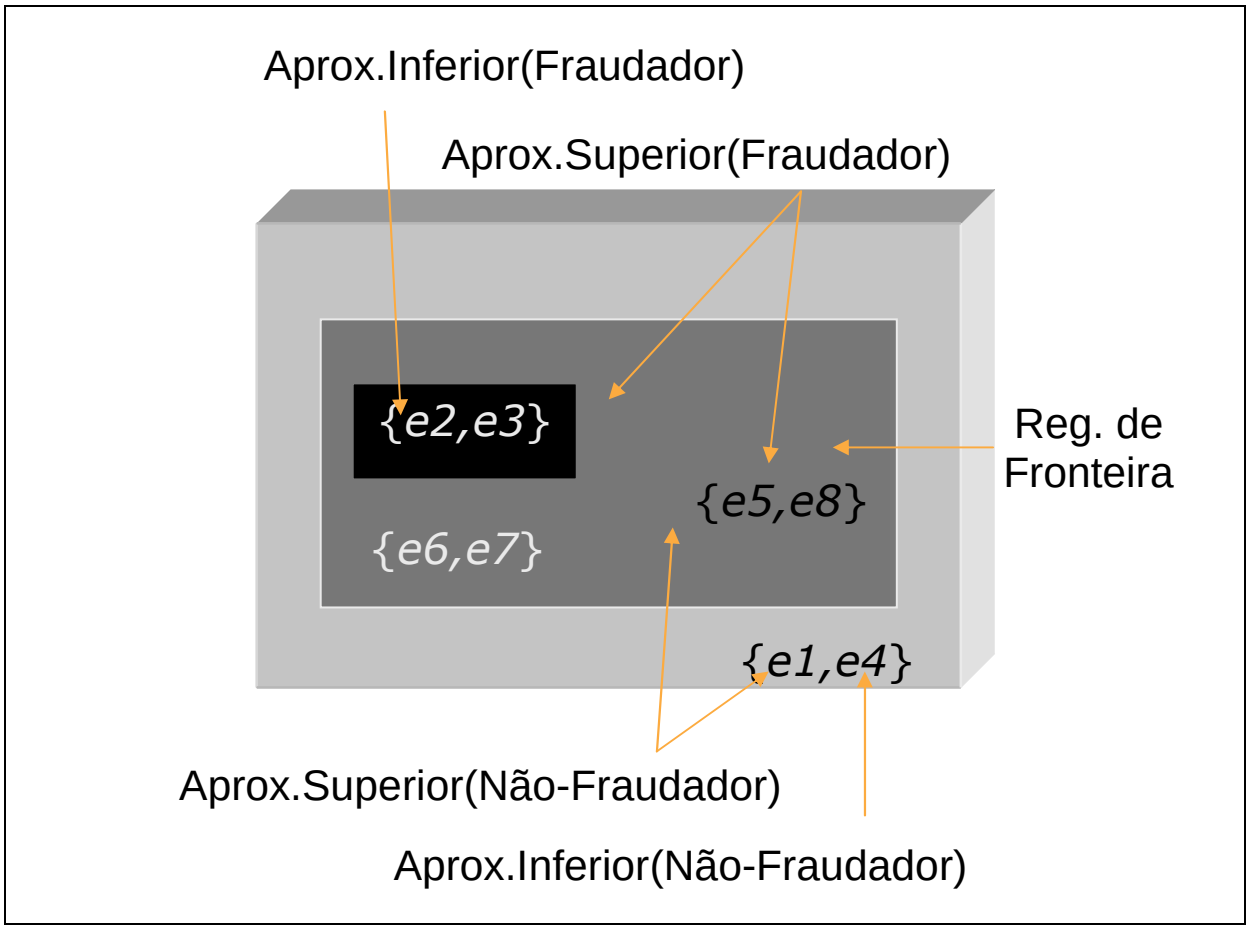

Figura 1 – Aproximação inferior, superior e região de fronteira.

2.4 Coeficiente de Incerteza

Pode ser interessante saber o quanto um conceito é bem definido ou não dentro de um sistema de decisão. Para tal, define-se o coeficiente de incerteza pela equação 1:

$$\alpha(X) = |\underline{X}| / |\overline{X}| \qquad (1)$$

O coeficiente de incerteza pode ser entendido como a qualidade da aproximação do conceito *X*. Ou seja, quanto mais $\alpha(X)$ se aproxima de 1, mais o conceito *X* é definido de forma exata (crips). E quanto mais $\alpha(X)$ se aproxima de 0, mais o conceito *X* é definido de forma imprecisa (rough).

2.5 Algoritmo da Decisão Mínima

O algoritmo da decisão mínima (MDA) é utilizado para reduções em sistemas de decisão ou banco de regras [6]. O MDA compara os valores dos atributos de um objeto com os demais objetos do sistema de decisão. Caso encontre valores de atributos que possam ser eliminados sem que dois objetos tornem-se contraditórios, o MDA retira este valor do objeto.

3 APLICAÇÃO

O software MATLAB [7] foi utilizado na implementação dos conceitos e algoritmos da teoria de rough sets, aproveitando sua facilidade de uso e portabilidade.

A solução aplicada consistiu em dividir o banco de dados em dados de treinamento e teste. Esta divisão é um procedimento típico do aprendizado supervisionado. O primeiro passo empregado foi eliminar registros repetidos, isto é, deixar os dados de treinamento somente com objetos distintos. No segundo passo, iniciou-se o uso dos conceitos de rough sets apresentados. Utilizou-se o software Rosetta [8] para determinar um reduto. A partir dele, os atributos linearmente dependentes foram eliminados dos dados de treinamento. Esta eliminação diminui diretamente a dimensão dos dados. Embora tenham sido eliminados atributos, alguns objetos também podem ser suprimidos. Isto porque novos objetos tornam-se idênticos com a retirada de atributos. A aproximação inferior para o conceito *Fraudador* = {*Sim*} foi encontrada e os demais registros eliminados. Conseqüentemente, não existiram mais objetos da região de fronteira. Garante-se assim que os objetos restantes nos dados de treinamento estão totalmente no conceito. A seguir, foi aplicado o MDA sobre os dados de treinamento reduzidos nos passos à cima. O algoritmo conseguiu minimizar significantemente os dados de treinamento. Novamente, após a aplicação do MDA, outros objetos idênticos surgiram e foram removidos.

Finalmente, para cada objeto dos dados de treinamento, uma regra de classificação foi derivada, a qual determina um perfil de fraudador. O conjunto de regras de classificação é chamado sistema de regras de classificação. Com o sistema de regras de classificação em mãos, bastou testar a qualidade das regras nos dados de teste.

4 RESULTADOS

O banco de dados utilizado possuía aproximadamente 40.600 registros (exemplos, objetos), dos quais 90% são classificados como Normal e 10% como Fraude. O conjunto de dados foi separado aleatoriamente em 20.300 registros para treinamento e 20.300 registros para teste. Tomando somente o conjunto de registros para treinamento, foi encontrado o reduto para este conjunto. A aproximação inferior, correspondendo aos exemplos fraudulentos, constou de 630 registros. Finalmente foi aplicado o MDA, resultando em 450 registros, sendo que estes registros resultaram em regras esparsas, i. e., nem todos os atributos foram utilizados em cada regras. Assim, de posse das regras geradas, o sistema foi submetido ao conjunto de teste com índice de acerto da ordem de 20 %, o que ficou abaixo do objetivo final que é de 30%.

5 Conclusões

- Rough sets é uma poderosa ferramenta de detecção de fraudes, principalmente quando não existem informações preliminares sobre o sistema, além do banco de dados;
- Apesar de alto custo computacional, os algoritmos de rough sets são de fácil implementação e compreensão;
- Um sistema híbrido, englobando conceitos de lógica fuzzy, pode ser a base de futuros trabalhos na detecção de fraudes usando rough sets.
- O sistema obtido resultou em acertos da ordem de 20%, o que ficou abaixo do objetivo final, que é de 30%.
- A melhoria está em desenvolvimento, porém constatou-se que o resultado obtido até o presente está abaixo do desejado principalmente devido à qualidade (fidelidade à realidade) dos dados.
- Novos estudos estão sendo feitos, utilizando-se das aproximações, no sentido de melhorar o pré-tratamento dos dados.

Referências Bibliográficas

[1] R.Brause, T. Langsdorf, M.Hepp: Neural data mining for card fraud detection. Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on, 9-11 Nov.1999, pp. 103 – 106.

[2] Deshmukh, A.; Talluru, T.L.N.: A rule based fuzzy reasoning system for assessing the risk of management fraud. Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation'., 1997 IEEE International Conference on , Volume: 1 , 12-15 Oct. 1997, pp. 669 - 673 vol.1.

[3] H. Furuta, M. Hirokane, Y. Mikumo: Extraction Method Based on Rough Sets Theory of Rule-Type Knowledge from Diagnostic Cases of Slope-Failure Danger Levels. Rough Sets in Knowledge Discovery 2 - Application, Case Studies and Software Systems, Part 1. Applications, Chapter 10, Pages: 178-192.

[4] Pawlak, Z. (1982). Rough Sets. International Jornal of Computer ans Information Sciences, pages 341-356.

[5] S. Pal, A. Skowron. Rough-fuzzy hybridization: a new trend in decision-marketing. Springer-Verlag Singapure Pte. Ldt. 1999.

[6] H. Furuta, M. Hirokane, Y. Mikumo: Extraction Method Based on Rough Sets Theory of Rule-Type Knowledge from Diagnostic Cases of Slope-Failure Danger Levels. Rough Sets in Knowledge Discovery 2 - Application, Case Studies and Software Systems, Part 1. Applications, Chapter 10, Pages: 178-192.

[7] MATLAB: The Language of technical Computing. Copyright 1984-2004 The Mathworks, Inc. http://www.mathworks.com

[8] Ohrn, A. Rosetta: Technical reference manual. Technical report, Knowledge System Group, Norwergian University on Science and Technology, NO. http://rosetta.lcb.uu.se/general/