Aplicação dos conceitos da Teoria dos Conjuntos Aproximados no tratamento da indiscernibilidade

Use of the Rough Set Theory concepts in the treatment of indiscernibility

Renato José Sassi
Doutor em Engenharia Elétrica – POLI – USP;
Professor e Pesquisador do Curso de Mestrado em Engenharia de Produção – Uninove;
Pesquisador Associado do Grupo de Inteligência Computacional, Modelagem e Neurocomputação – ICONE-LSI – USP.
São Paulo, SP [Brasil]
sassi@uninove.br

A Teoria dos Conjuntos Aproximados (TCA), ou Teoria dos *Rough Sets* (RS) foi proposta por Pawlak como um modelo matemático para representação do conhecimento e tratamento de incerteza. Conjuntos Aproximados podem ser considerados conjuntos com fronteiras nebulosas, ou seja, conjuntos que não podem ser caracterizados precisamente, utilizando-se dos atributos disponíveis. RS trata de um tipo fundamental de incerteza, a indiscernibilidade, a qual surge quando não é possível distinguir elementos de um mesmo conjunto, e representa a situação em que esses componentes parecem todos ser um único elemento. O objetivo neste trabalho é a aplicação dos conceitos fundamentais da teoria para reduzir a incerteza contida em uma base de dados. Os resultados obtidos são satisfatórios, o que confirma RS como uma teoria que apresenta uma matemática consistente com um bom embasamento teórico e que pode ser utilizada no tratamento da incerteza.

Palavras-chave: Conjuntos aproximados. Incerteza. Indiscernibilidade. *Rough Sets*.

The Rough Set Theory was proposed by Pawlak, in 1982, as a mathematical model to represent knowledge and to treat uncertainty. The rough sets can be considered as sets with hazy frontiers, i.e., which can not be precisely characterized, using the available features. Rough set is a fundamental type of uncertainty, the indiscernibility. Indiscernibility arises when it is not possible to distinguish elements from the same set, and represents the situation in which these components seem to be only one element. The objective of this paper was the application of the theory's main concepts to reduce the uncertainty in a database. The achieved results are satisfactory, which confirms that the rough set is a theory with a consistent mathematics of good theoretical basis and that can be used in the uncertainty treatment.

Key words: Indiscernibility. Rough Sets. Uncertainty.

1 Introdução

A Teoria dos *Rough Sets* (RS), ou Teoria dos Conjuntos Aproximados (TCA), foi proposta por Pawlak (1982) como um modelo matemático para representação do conhecimento e tratamento de incerteza. Conjuntos aproximados podem ser considerados conjuntos com fronteiras nebulosas, ou seja, que não podem ser caracterizados precisamente utilizando-se dos atributos disponíveis (PAWLAK, 1991).

A incerteza pode se manifestar de diversas formas, tais como imprecisão, incompletude, inconsistência, etc. RS trata de um tipo fundamental de incerteza, a indiscernibilidade que surge quando não é possível distinguir elementos de um mesmo conjunto e representa a situação em que esses componentes parecem todos ser um único elemento (UCHÔA, 1998).

Uma das principais vantagens da teoria dos Rs é que ela não necessita de informações preliminares ou adicionais sobre os dados, tais como a distribuição de probabilidade em estatística, atribuição de probabilidades básicas na Teoria de Dempster-Shafer, ou mesmo os graus de pertinência na Teoria dos Conjuntos *Fuzzy* (UCHÔA, 1998), porque esta técnica utiliza única e exclusivamente a estrutura interna dos dados analisados para modelar o conhecimento. Isso pode ser verificado em Ziarko (1994).

O conceito dos Rs relaciona-se, de alguma maneira, com outras teorias matemáticas desenvolvidas para manipular incerteza, particularmente com a Teoria da Evidência de Dempster-Shafer. A principal diferença é que a teoria proposta por Dempster-Shafer utiliza a função de crença como ferramenta principal, enquanto a teoria dos Rs faz uso de Aproximações Inferior e Superior. Existe também uma relação entre a teoria dos Rs e a teoria dos *Fuzzy Sets*, as quais são frequentemente comparadas e, até mesmo, confundidas, pois tratam da incerteza. Um estudo comparativo mais aprofundado entre RS e *Fuzzy Sets* pode ser encontrado em Yao (1998).

Szladow e Ziarko (1993) utilizam um exemplo da área de processamento de imagens para explicar a diferença entre RS e *Fuzzy Sets*. Enquanto *Fuzzy Sets* aborda a existência de mais de um nível de cinza nos *pixels*, RS cuida do tamanho desses *pixels*. Assim, *Fuzzy Sets* trata da relação entre intensidades de elementos dentro da mesma classe, e o RS, da relação entre grupos de elementos em diferentes classes. Entretanto, a teoria dos RS não compete com a dos *Fuzzy* Sets, mas a complementa. Adjei et al. (2001) apresentam um trabalho que combina ambas as teorias. Na realidade, a teoria dos Rs e a dos *Fuzzy Sets* são duas abordagens independentes para o tratamento de conhecimento impreciso.

Os conceitos de RS têm-se mostrado muito úteis, quando aplicados a problemas do tipo redução de atributos, descoberta de dependência entre eles e de padrões entre os dados (PAWLAK, 1996). A redução de atributos realizada pelos Rs é feita por meio dos chamados redutos, que são subconjuntos de atributos capazes de representar o conhecimento da base de dados com todos os seus atributos iniciais. Este procedimento de eliminação daqueles irrelevantes é uma das características da teoria.

Devido às características comentadas, a crescente utilização de RS pode ser comprovada pelo número de aplicações e publicações científicas nas seguintes áreas: KDD e *Data Mining* (XIAHOUA e CERCONE, 1996), medicina (KOMOROVSKI, 1999), negócios (NARAKESARI e ZAK, 2004), engenharia (BONALDI, SILVA, LAMBERT-TORRES et al., 2002); reconhecimento de padrões (ZHANG e YAO, 2004), *Text Mining* (YU, SHOUYANG e LAI, 2005), dentre outras.

Os principais conceitos dos Rs são: Espaços Aproximados, Aproximação Inferior

(AI), Aproximação Superior (AS), Sistema de Informação (S), Sistema de Decisão (SD) e Indiscernibilidade (IND).

Este trabalho não tem como finalidade o aprofundamento no formalismo matemático dos Rs por ser extenso, embora considerado um aspecto importante da teoria. Para isto, recomenda-se o trabalho de Uchôa (1998). Esse formalismo matemático não é sistematicamente abordado, não é padronizado e tampouco muito explorado. Uma das razões disso se deve ao fato dos Rs ser uma teoria razoavelmente recente. Ainda não existe uma padronização da notação matemática; assim, em alguns casos, foi necessário adotar uma notação própria, com o objetivo de torná-la mais clara, como: Aproximação Inferior (AI) $\underline{B}$, Aproximação Superior (AS) $\overline{B}$, Sistema de Informação (S), atributos condicionais (C), atributo de decisão (d), Sistema de Decisão (SD), Região de Fronteira RF(X), Região Negativa RN(X) e Reduto (RED). Deve-se entender também que a palavra elementos é tratada como sinônimo de casos, exemplos ou registros que compõem uma base de dados. Assim, o objetivo neste estudo é a aplicação dos conceitos fundamentais da teoria dos para reduzir a incerteza (indiscernibilidade) contida em uma base de dados.

O restante do trabalho está organizado da seguinte forma: na seção 2 discute-se o conceito de Espaços Aproximados e de Aproximação de Conjuntos; na seção 3, as Medidas de Qualidade das Aproximações são apresentadas; na seção 4, analisa-se a Dependência de Atributos e, na seção 5, o trabalho é encerrado com a conclusão.

2 Espaços Aproximados

Um espaço aproximado é um par ordenado A = (U,R), em que: U é um conjunto não vazio, denominado conjunto universo, e R é uma relação de equivalência sobre U, denominada Relação de Indiscernibilidade. Uma relação binária $R \subseteq X \times X$, a qual é reflexiva (um elemento está relacionado com ele próprio xRx), simétrica (se xRy então yRx) e transitiva (se xRy e yRz então xRz), é chamada de relação de equivalência. Dados os elementos x, y $\in$ U, se xRy então x e y são indiscerníveis em A, ou seja, a classe de equivalência definida por x é a mesma que a definida por y, i.e., [x]R = [y]R.

A classe de equivalência de um elemento x $\in$ X consiste em todos os elementos y $\in$ X para os quais xRy. Os elementos que são indiscerníveis formam os chamados conjuntos elementares. Dessa forma, pode-se dizer que as classes de equivalência de R são os conjuntos elementares de A. Na Figura 1, pode-se visualizar o espaço aproximado A = (U,R).

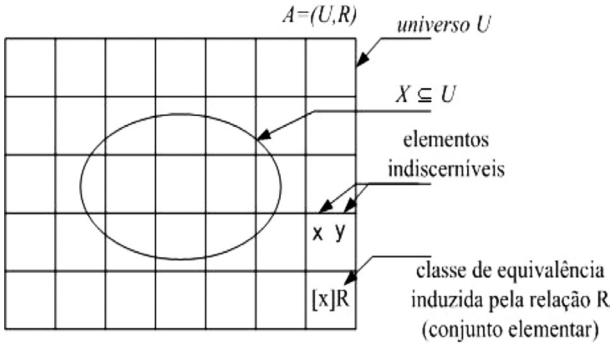

Figura 1: Espaço Aproximado A = (U,R) e X ⊆ U
Fonte: Adaptado de Pawlak (1991).

A forma mais comum para representação dos dados em RS é por um Sistema de Informação (S) que contém um conjunto de elementos, em que cada elemento tem uma quantidade de atributos condicionais. Esses atributos são os mesmos para cada um dos elementos, mas os seus valores nominais podem diferir (Tabela 1).

Dessa forma, um Sistema de Informação é um par ordenado S = (U, C), em que U é um conjunto finito e não vazio de elementos chamado de

universo (Figura 1), e C é um conjunto finito e não vazio formado pelos atributos. Cada atributo a ∈ C é uma função a: U → Va, em que Va é o conjunto dos valores permitidos para o atributo a (sua faixa de valores). Na Tabela 1, apresenta-se o sistema de informação S, e pode-se observar os principais conceitos de RS: o espaço aproximado *A* = (U,R); o universo U formado pelos elementos e1; e2; e3; e4; e5, e6 e os atributos (C) Experiência do Vendedor, Qualidade do Produto e Boa Localização, e R a relação de equivalência sobre U.

Tabela 1: Sistema de Informação (S)

LOJA	Experiência do Vendedor	Qualidade do produto	Boa localização
e1	Alta	Boa	Não
e2	Média	Boa	Não
e3	Média	Boa	Não
e4	Baixa	Média	Não
e5	Média	Média	Sim
e6	Alta	Média	Sim

Fonte: Adaptado de Pawlak (1991).

O principal conceito envolvido em RS é a Relação de Indiscernibilidade (PAWLAK, 1996), a qual normalmente está associada a um conjunto de atributos. Se tal relação existe entre dois elementos, isso significa que todos os valores nominais dos seus atributos são idênticos com respeito aos atributos considerados; portanto, não podem ser discernidos (distinguidos) entre si. Para cada subconjunto de atributos B ⊆ C no sistema de informação S = (U, C), uma relação de equivalência INDs(B) é associada, chamada de Relação de Indiscernibilidade definida como:

$$\mathrm{IND}s(B) = \{(x,y) \in U^2 \mid \forall a \in B,\ a(x) = a(y)\} \quad (1)$$

O conjunto de todas as classes de equivalência na relação INDs(B) é representado por U/INDs(B), denominado quociente de U pela relação INDs(B). Em muitos casos, é importante a classificação dos elementos considerando um atributo de decisão que informa a decisão a ser tomada. Dessa forma, um SI que apresenta um atributo de decisão é denominado Sistema de Decisão (SD). Um SD pode ser representado por SD = (U, C ∩ {d}), em que d ∉ C é o atributo de decisão. A Tabela 2 mostra um SD obtido a partir do Sistema de Informação S da Tabela 1, destacando os atributos condicionais (Experiência do Vendedor, Qualidade do Produto e Boa Localização) e o atributo de decisão (Retorno).

Tabela 2: Sistema de Decisão (Sistema de Informação com o atributo de decisão Retorno)

Loja	Atributos condicionais			Atributo de decisão
	Experiência do vendedor	Qualidade do produto	Boa localização	Retorno
e1	Alta	Boa	Não	Lucro
e2	Média	Boa	Não	Prejuízo
e3	Média	Boa	Não	Lucro
e4	Baixa	Média	Não	Prejuízo
e5	Média	Média	Sim	Prejuízo
e6	Alta	Média	Sim	Lucro

Fonte: Adaptado de Pawlak (1991).

Os valores dos atributos são chamados de valores nominais e estão expressos como: Experiência do Vendedor {Alta, Média, Baixa}; Qualidade do Produto {Boa, Média}; Boa Localização {Não, Sim} e Retorno {Lucro, Prejuízo}. Considerando cada atributo condicional de forma independente, a relação de equivalência do Sistema de Informação S (Tabela 1) forma os seguintes conjuntos elementares: experiência do vendedor Alta {e1, e6}; Média {e2, e3, e5}; Baixa {e4}; Qualidade do Produto: Boa {e1, e2, e3}; Média {e4, e5, e6} e Boa Localização: Não {e1, e2, e3, e4}; Sim {e5, e6}.

Ao utilizar todos os atributos condicionais do Sistema de Informação S da Tabela 1, obtêm-se os seguintes conjuntos elementares: {e1}, {e2, e3}, {e4}, {e5} e {e6}. Observando a Tabela 3, pode-se

perceber que existem 2 elementos (casos) {e2} e {e3} iguais (destacados em negrito), no que se refere a valores de atributos condicionais.

Tabela 3: Sistema de Decisão com os elementos e2 e e3 indiscerníveis, com relação aos atributos condicionais

Loja	Experiência do vendedor	Qualidade do produto	Boa localização	Retorno
e1	Alta	Boa	Não	Lucro
e2	Média	Boa	Não	Prejuízo
e3	Média	Boa	Não	Lucro
e4	Baixa	Média	Não	Prejuízo
e5	Média	Média	Sim	Prejuízo
e6	Alta	Média	Sim	Lucro

Fonte: Adaptado de Pawlak (1991).

Existindo a Relação de Indiscernibilidade entre os elementos {e2} e {e3} como mostrado na Tabela 3, significa que todos os valores nominais de seus atributos são idênticos com relação ao subconjunto de atributos B (B ⊆ S) considerado, ou seja, não podem ser diferenciados entre si.

2.1 Aproximação de Conjuntos

A Tabela 3 apresenta os elementos do Sistema de Informação S segundo as características do atributo de decisão. Pode-se então, fazer a seguinte pergunta: quais características dos atributos condicionais definem o retorno da loja como lucro ou prejuízo? Note-se que não há uma resposta única para essa pergunta, pois as lojas {e2} e {e3} apresentam as mesmas características dos atributos condicionais, mas se diferenciam no atributo de decisão.

Pode-se dizer com certeza, conforme a Tabela 3, que qualquer loja com características iguais às das lojas {e1} e {e6} terão lucro, assim como qualquer loja com características iguais às das lojas {e4} ou {e5} terá prejuízo, porém, nada se pode afirmar para lojas com características iguais às das lojas {e2} e {e3}; pois, apesar de apresentarem atributos condicionais com as mesmas características, possuem atributos de decisão diferentes. São nesses casos que RS pode ser aplicado.

Um conjunto definível em S é qualquer união finita de conjuntos elementares. Para cada conceito X – que é o conjunto de elementos com respeito a B, ou seja, X é obtido pelas informações dos atributos de B –, são computados o maior conjunto definível contido em X, e o menor, que contém X. O primeiro conjunto é chamado de Aproximação Inferior (AI) de X, e o segundo, de Aproximação Superior (AS) de X (PAWLAK, 1982). A Aproximação Inferior $\underline{B}(X)$ e a Aproximação Superior $\overline{B}(X)$ de um conjunto de elementos $X \subseteq U$ com respeito a um conjunto de atributos $B \subseteq S$ (definindo uma relação de equivalência em U) pode ser definida em termos das classes na relação de equivalência, da seguinte forma:

$$\underline{B}(X) = \{x \in U \mid U/\ INDs(B) \subseteq X\} \quad (2)$$

$$\overline{B}(X) = \{x \in U \mid U/\ INDs(B) \cap X \neq \emptyset\} \quad (3)$$

Os elementos da Aproximação Inferior $\underline{B}(X)$ são classificados com certeza como membros de X, utilizando o conjunto de atributos B, e os elementos da Aproximação Superior $\overline{B}(X)$ podem ser identificados como possíveis membros de X, utilizando o mesmo conjunto de B. Os elementos que correspondem a lucro podem ser tomados como exemplos do conceito X. Com base na Tabela 3 pode-se observar que existem três elementos que possuem como atributo de decisão lucro {e1, e3, e6}, porém, existe Relação de Indiscernibilidade entre os elementos {e2} e {e3}, impedindo que {e3} seja considerado com certeza como membro de X. Assim, somente os elementos {e1, e6} podem ser classificados como membros de X e elementos da Aproximação Inferior $\underline{B}(X)$. A Figura 2 ilustra a Aproximação Inferior $\underline{B}(X)$ destacada na cor cinza.

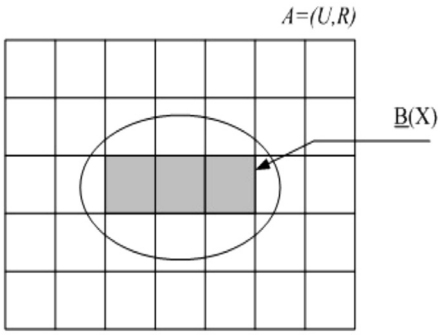

Figura 2: Aproximação Inferior de X
Fonte: Adaptado de Pawlak (1991).

Na Figura 2, nos quadrados de cor cinza, estão contidos os elementos que correspondem a Aproximação Inferior $\underline{B}(X)$, os quadrados em branco tocados pela elipse (X) estão os elementos que correspondem a Aproximação Superior $\overline{B}(X)$ e os quadrados em branco que não são tocados pela elipse correspondem com certeza aos elementos que não pertencem a $\overline{B}(X)$ (Região Negativa). Na Aproximação Superior $\overline{B}(X)$ são classificados os elementos que são possíveis membros de X. Dessa forma, ela reúne os elementos com atributo de decisão igual a lucro {e1, e3, e6}, e também, os igual a prejuízo, desde que exista uma Relação de Indiscernibilidade entre eles, como ocorre entre o elemento {e2} e o {e3}. A Figura 3 ilustra a Aproximação Superior $\overline{B}(X)$.

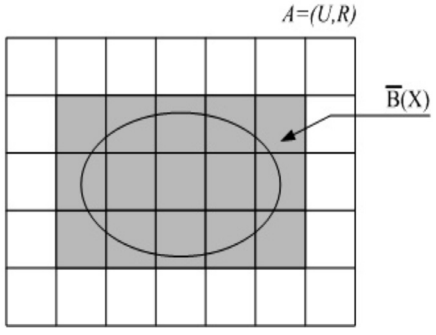

Figura 3: Aproximação Superior de X
Fonte: Adaptado de Pawlak (1991).

Portanto, obtém-se uma Região de Fronteira de X pela diferença de $\overline{B}X$ para $\underline{B}(X)$, representada por RF(X), ou seja, $RF(X) = \overline{B}(X) - \underline{B}(X)$. A Região de Fronteira, também chamada de Duvidosa, possui somente os elementos que não podem ser classificados com certeza como pertencentes em X, utilizando o conjunto de atributos B. É a região formada pelos elementos de U que pertencem à Aproximação Superior, mas que não pertencem à Aproximação Inferior.

Um conjunto X é definido como *rough* (impreciso) se a sua Região de Fronteira é diferente do conjunto vazio ($RF(X) \neq 0$), e como *crisp* (preciso), se o conjunto for vazio($RF(X)=0$). Na Tabela 3, pode-se observar que os elementos {e2, e3} fazem parte da Região de Fronteira.

A Região Negativa é dada pela diferença dos elementos de U para $\overline{B}(X)$, representada por $RN(X) = U - \overline{B}(X)$. A Região Negativa possui somente os elementos que com certeza não podem ser classificados como pertencentes à Aproximação Superior $\overline{B}(X)$, utilizando o conjunto de atributos B.

As regiões de X ficaram assim: Aproximação Inferior: $\underline{B}(X)$ {e1, e6}; Aproximação Superior: $\overline{B}(X)$ {e1, e2, e3, e6}; Região de Fronteira (Duvidosa): RF(X) {e2, e3} e Região Negativa: RN(X) {e4, e5}. A Figura 4 ilustra todas as regiões de X em A.

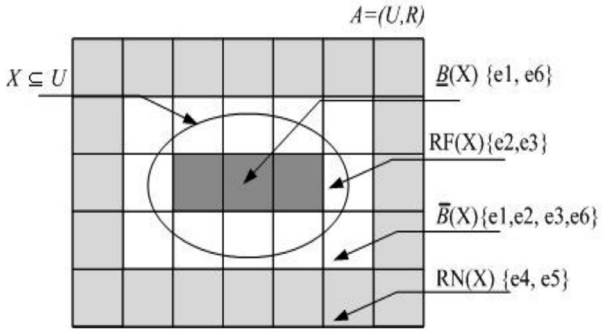

Figura 4: Todas as regiões de X em A
Fonte: Adaptado de Pawlak (1991).

Por meio dos conceitos de AS e AI, podem-se definir as quatro classes básicas de incerteza em RS (KOMOROVSKI, PAWLAK, POLKOWSKI et al., 1999).

Um conjunto definível em S é qualquer união finita de conjuntos elementares. O significado dessas classes, onde –X denota U – X (complemento) é o seguinte: X é *rough* B-definível, se e somente se $\underline{B}(X) \neq \emptyset$ e $\overline{B}(X) \neq U$.

Se X é *rough* B-definível, isso significa que é possível decidir para alguns elementos de U quando eles pertencem a X, e para alguns elementos de U, quando eles pertencem a –X, utilizando B; X é internamente B-indefinível, se e somente se $\underline{B}(X) = \emptyset$ e $\overline{B}(X) \neq U$.

Se X é internamente B-indefinível, isso significa que é possível decidir para alguns elementos de U quando eles pertencem a –X, mas não é possível decidir para nenhum elemento de U quando ele pertence a X, utilizando B; X é externamente B-indefinível, se e somente se $\underline{B}(X) \neq \emptyset$ e $\overline{B}(X) = U$.

Se X é externamente B-indefinível, isso significa que é possível decidir para alguns elementos de U quando eles pertencem a X, mas não é possível decidir para qualquer elemento de U quando ele pertence a – X, utilizando B; X é totalmente B-indefinível, se e somente se $\underline{B}(X) = \emptyset$ e $\overline{B}(X) = U$. Se X é totalmente B-indefinível, isso significa que não é possível decidir para qualquer elemento de U quando ele pertence a X ou a –X, utilizando B.

3 Medidas de Qualidade das Aproximações

A Qualidade das Aproximações obtidas pelas definições dadas previamente pode ser caracterizada numericamente a partir dos próprios elementos que as definem. O Coeficiente para medir essas Qualidades é representado por α_B (X), sendo X o conjunto de elementos com respeito a B, e podem ser realizadas de três formas (PAWLAK, 1982): Coeficiente de Incerteza; Coeficiente de Qualidade da Aproximação Superior e Coeficiente de Qualidade da Aproximação Inferior. As três medidas foram calculadas nos exemplos a seguir:

O Coeficiente de Incerteza $\alpha_B(X)$, pode ser entendido como a Qualidade da Aproximação de X, dado por:

$$\alpha_B(X) = \frac{|\overline{B}(X)|}{|\underline{B}(X)|} \quad (4)$$

em que $|\underline{B}(X)|$ e $|\overline{B}(X)|$ denotam a cardinalidade das Aproximações Inferior e Superior, respectivamente. Obviamente $0 \leq \alpha_B \leq 1$. Se $\alpha_B(X) = 1$, X é *crisp* (preciso) em relação ao conjunto de atributos B. Se $\alpha_B(X) < 1$, X é *rough* (impreciso), em relação ao conjunto de atributos B.

a) O coeficiente $\alpha_B(X)$ pode ser interpretado como o percentual de todos os elementos possivelmente classificados como pertencentes a X, dado por:

$$\alpha_B(X) = \frac{|\overline{B}(X)|}{|\underline{B}(X)|} = \frac{|\{e1, e6\}|}{|\{e1,e2,e3,e6\}|} = \frac{2}{4} = 0{,}5 \quad (5)$$

ou seja, com base nos elementos da Tabela 3, 50% de X é preciso com respeito a B.

b) Coeficiente de Qualidade da Aproximação Superior; $\alpha_B(\overline{B}(X))$, que pode ser interpretado como sendo o percentual de todos os elementos possivelmente classificados como pertencentes a X, dado por:

$$\alpha_B(\overline{B}(X)) = \frac{|\overline{B}(X)|}{|U|} = \frac{|\{e1, e6\}|}{|\{e1,e2,e3,e4,e5,e6\}|} = \frac{4}{6} = 0{,}66 \quad (6)$$

que significa dizer que, com base nos elementos da Tabela 3, 66% de U possivelmente pertence à X.

c) Coeficiente de Qualidade da Aproximação Inferior $\alpha_B(\underline{B}(X))$, que pode ser interpretado como sendo o percentual de todos os elementos certamente classificados como pertencentes a X, dado por:

$$\alpha_B(\underline{B}(X)) = \frac{|\underline{B}(X)|}{|U|} = \frac{|\{e1, e6\}|}{|\{e1,e2,e3,e4,e5,e6\}|} = \frac{2}{6} = 0{,}33 \quad (7)$$

isto é, com base nos elementos da Tabela 3, 33% de U certamente pertence à X.

4 Dependência de Atributos

RS pode ser utilizado para a análise de dependência entre atributos, visando principalmente identificar e eliminar os redundantes ou desnecessários. A eliminação dos redundantes permite que se encontre um subconjunto mínimo de atributos que possui o mesmo valor discriminatório do conjunto original (ZIARKO e KATZBERG, 1993).

Um conjunto de atributos D (decisão) depende totalmente de C (condicionais) (C ⇒ D), se todos os valores nominais de D forem univocamente determinados por valores nominais dos atributos de C. Em outras palavras, D depende totalmente de C se existir uma dependência funcional entre valores nominais de C e D. Se (C,D) = 1, diz-se que D depende totalmente de C; se (C,D) = 0, D não depende de C e se 0 < (C,D) < 1, D depende parcialmente de C. (ZIARKO e KATZBERG, 1993). Formalmente, a dependência pode ser assim definida: sejam C e D subconjuntos de S. O Grau de Dependência de D em relação a C é dado por:

$$\gamma(C,D) = \frac{|\underline{B}(X)|}{|U|} = \frac{|\{e1, e6\}|}{|\{e1,e2,e3,e4,e5,e6\}|} = \frac{2}{6} = 0{,}33 \quad (8)$$

Quando D depende parcialmente de C indica que não existe a necessidade da presença de todos os atributos condicionais de C para gerar os valores nominais do atributo de decisão D, abrindo espaço para uma redução de atributos. Do resultado de 0,33 (33%), deduz-se que dos três atributos condicionais pertencentes ao sistema de informação S (Experiência do Vendedor, Qualidade do Produto e Boa Localização), um deles (33%) pode ser reduzido sem que a base de dados perca a sua representatividade original e os outros dois (66%) apresentam dependência e, por isso, não podem ser reduzidos.

Constata-se que a redução de atributos pode ser realizada com RS e que auxilia na redução da incerteza, pois, reduz informação desnecessária que pode contribuir com a indiscernibilidade. Não é o objetivo, neste trabalho, discorrer-se sobre a redução de atributos e sim analisar o conceito de aproximação. Como foi visto, alguns atributos são mais significativos (relevantes) que outros impossibilitando a sua redução.

Sejam C e D conjuntos de atributos condicionais e de decisão, respectivamente, e seja **a** um atributo condicional pertencente a C. A Significância de a será calculada em razão da mudança do Grau de Dependência de D em relação a C (0,66) com a remoção de a, conforme a seguinte fórmula:

$$\sigma_{(P,O)}(a) = 1 - \left(\frac{(C - \{a\},D)}{(C,D)}\right) \quad (9)$$

Esse Coeficiente pode ser visto como o erro que ocorre na definição de D por C quando a é removido. Com base na Tabela 3 e utilizando os elementos com atributo de decisão igual a Lucro, temos: para X = {e1, e3, e6}, D = {Retorno}, C = {Experiência do Vendedor, Qualidade do Produto, Boa Localização}, o Grau de Significância do

atributo a é dado por 1 – (({Experiência do Vendedor, Qualidade do Produto})/({Experiência do Vendedor, Qualidade do Produto, Boa Localização})) = 1 – ((2/3)/(2/3)) = 0, significando que remover o atributo Boa Localização não afetará os resultados. O mesmo resultado seria encontrado ao se remover Qualidade do Produto de {Experiência do Vendedor, Qualidade do Produto, Boa Localização}.

Pode-se observar que a Aproximação de Conjuntos está diretamente relacionada com a Relação de Indiscernibilidade. A redução ou eliminação da Indiscernibilidade (incerteza) em uma base de dados gera regras de decisão mais confiáveis, além da diminuição do esforço computacional.

Outro trabalho será desenvolvido para tratar de forma específica e aprofundada a redução de atributos focalizando o conceito de redutos, de Matriz de Discernibilidade, de Função de Discernibilidade e de geração de regras de decisão.

5 Conclusão e contribuições

Neste trabalho, foram abordados os principais conceitos dos Rs e suas aplicações no tratamento de um tipo fundamental de incerteza, a Indiscernibilidade.

Inicialmente, por se tratar de temática emergente, esforços foram feitos nas revisões de literatura e, em particular, na formulação do texto, sendo necessária a padronização da notação matemática objetivando maior clareza.

Acredita-se que o estudo contribuiu para o maior conhecimento e maior difusão da Teoria dos Rs ao realizar uma revisão importante dos principais conceitos;

- ao padronizar o confuso formalismo matemático apresentado na bibliografia pesquisada;
- ao demonstrar a aplicação dos principais conceitos da teoria na redução da incerteza contida em uma base de dados;
- ao apresentar outra opção para o tratamento da incerteza, além das já tradicionalmente conhecidas como a Teoria dos *Fuzzy Sets*, a Distribuição de Probabilidade em Estatística e a Teoria da Evidência de Dempster-Shafer.

Em geral, quando se fala de redução de dados, sempre vem à mente um ganho em tempo de processamento. Entretanto, a redução de atributos realizada pelo RS faz com que a informação considerada desnecessária não seja utilizada, o que pode reduzir a incerteza contida na base de dados. Isso pode resultar numa geração de regras mais confiáveis por parte de um algoritmo de classificação ou agrupamentos mais coesos e confiáveis por parte de um algoritmo de clusterização.

RS também é utilizado em Arquiteturas Híbridas para pré-processar os dados antes de submetê-los a uma outra técnica inteligente como uma rede neural artificial.

A grande vantagem desse tipo de Arquitetura deve-se ao sinergismo obtido pela combinação de duas ou mais técnicas. Este sinergismo resulta na obtenção de um sistema mais poderoso (em termos de interpretação, de aprendizado, de estimativa de parâmetros, de generalização, dentre outros) e com menos deficiências.

Desde o aparecimento de RS, muitos trabalhos vêm sendo gerados para conhecer melhor a teoria e suas aplicações, tais como utilização dos conceitos de Aproximação Superior e de Aproximação Inferior para aproximar atributos desconhecidos com base em conhecidos, elaboração de relações e operações sobre conjuntos rough (imprecisos), similares àquelas utilizadas em conjuntos clássicos (BONIKOWSKI, 1998), medidas de incerteza baseadas em teoria da informação (BEAUBOFF e LANG, 1998) e aplicação de RS

em modelos de bancos de dados relacionais (HU e CERCONE, 1994).

Estudos mais avançados baseados nos principais conceitos dos Rs e suas aplicações como aqui apresentadas apontam para o desenvolvimento de trabalhos em Computação Granular (GULIATO e SANTOS, 2009) e em Mereologia (POLKOWSKI e ARTIEMJEW, 2009).

Conclui-se que RS é uma teoria que apresenta matemática consistente e confiável, além de um bom embasamento teórico e pode ser utilizada no tratamento da incerteza, na redução de atributos e em conjunto com outras técnicas formando Arquiteturas Híbridas. Quanto às dificuldades destaca-se que a documentação ainda é pequena e não existe padronização referente ao formalismo matemático. Assim, pode-se afirmar que existe muito a ser explorado no conteúdo da teoria e espaço para trabalhos de continuidade.

Referências

ADJEI, O. et al. A fuzzy search method for Rough Sets in data mining. *IFSA World Congress and 20th NAFIPS International Conference*, 2001, v. 2, p. 980-985.

BEAUBOUEF, T.; LANG, R. A Rough Sets Techniques for uncertainly management in automated story generation. *Communications of the ACM*, v. 4, p. 326-331, 1998.

BONIKOWSKI, Z. Extensions and intentions in the Rough Set Theory. *Journal of Information Sciences*, p. 149-167, 1998.

BONALDI, E. L. et al. Using Rough Sets techniques as a fault diagnosis classifier for induction motors. 28th Annual Conference of the Industrial Electronics Society, v. 4, p. 3383-3388, nov. 2002.

GULIATO, D.; SANTOS, J. C. S. Granular computing and Rough Sets to generate Fuzzy Rules. Proceedings of the 6th International Conference on Image Analysis and Recognition. Lecture Notes In Computer Science, v. 5627, p. 317-326, 2009.

HU, X.; CERCONE, N. Discovery of decision rules in relational databases: a Rough Set approach. *Conference on Information and Knowledge Management* (CIKM 19940, 1994.

KOMOROVSKI, J; Ohrn, A. Modelling prognostic power of cardiac tests using Rough Sets. *Artificial Intelligence in Medicine*, p. 167-191, 1999.

KOMOROWSKI, J.; PAWLAK, Z.; POLKOWSKI, L.; SKOWRON, A. *Rough Sets: a tutorial technical report*. Warsaw University, dec. 1999.

NARAKESARI, N.; ZAK, J. Application of Rough Sets methodology for data analysis in business marketing. *Harvard Business School*, 2004.

PAWLAK, Z. Rough Sets. *International Journal of Computer and information Sciences*, p. 341-356, 1982.

PAWLAK, Z. *Rough Sets: theoretical aspects of reasoning about data*. London, Kluwer, 1991.

PAWLAK, Z. Rough Sets, Rough relations and Rough functions. *Fundamenta Informaticae*, p. 103-108, 1996.

POLKOWSKI, L.; ARTIEMJEW, P. On knowledge granulation and applications to classifier induction in the framework of Rough Mereology. *International Journal of Computational Intelligence Systems (IJCIS)*, v. 2-4, p. 315-331, 2009.

SZLADOW, A.; ZIARKO, W. Rough Sets: working with imperfect data. *AI Expert*, july, p. 36-41, 1993.

UCHÔA, J. Q. *Representação e indução de conhecimento usando teoria de conjuntos aproximados*. 1998. Dissertação (Mestrado em Ciência da Computação)–Universidade Federal de São Carlos, 1998.

XIAOHUA, H.; CERCONE, N. Mining knowledge rules from databases: a Rough Set approach. *Proceedings of the Twelfth International Conference on Data Engineering*, p. 96-105, 1996.

YAO, Y.Y. A comparative study of Fuzzy Sets and Rough Sets. *Journal of Information Sciences*, p. 227-242, 1998.

YU, L; SHOUYANG, W.; LAI, K. K. A Rough-Set-Refined Text Mining Approach for Crude Oil Market Tendency Forecasting. *International Journal of Knowledge and Systems Sciences*, v. 2, n. 1, mar. 2005.

ZHANG, M.; YAO, J. T. A Rough Sets based approach to feature selection Processing. NAFIPS '04. *IEEE Annual Meeting of the Fuzzy Information*, v. 1, p. 27-30, p. 434-439, jun. 2004.

ZIARKO, W.; KATZBERG, J. D. Rough Sets approach to system modelling and control algorithm acquisition. WESCANEX 93. *Communications, Computers and Power in the Modern Environment.' Conference Proceedings., IEEE* , p. 154-164, may 1993.

ZIARKO, W. Rough Sets, Fuzzy Sets and Knowledge Discovery. 1.ed. New York: Springer-Verlag, 1994.

Recebido em 5 set. 2009 / aprovado em 9 fev. 2010

Para referenciar este texto
SASSI, R. J. Aplicação dos conceitos da Teoria dos Conjuntos Aproximados no tratamento da indiscernibilidade. *Exacta*, São Paulo, v. 8, n. 1, p. 89-98, 2010.