Teoria de Rough Sets Aplicada à Data Mining

Pila, A. D. e Monard, M. C.
Av. do Trabalhador Sãocarlense, 400
São Carlos/SP - CEP: 13560-970

Resumo

The Rough Sets Theory is a mathematical approach developded by Zdzislaw Pawlak in early 80's. The main concept of this theory is the *indiscerbility relation* which express the inability to discern examples based on some features. This concept is used to define the *reducts* which is a minimal subset of features that preserve the same indiscernibility relation of the entire set. This theory have been proposed as a method for the purpose of Data Mining. This paper shows the basic concept of the Rough Sets Theory and the Data Mining framework as well as some questions related to the effective application of Rough Sets in Data Mining.

1 Introdução

Previsões apontam que a quantidade de dados que serão armazenados nos computadores nos próximos cinco anos será maior que a quantidade armazenada nos últimos trinta anos. Essas previsões baseiam-se no crescimento exponencial do uso da maior rede de computadores do mundo, a Internet. Com isso, a maior parte dos dados das empresas estarão armazenados nos computadores e uma manipulação adequada desses dados faz-se necessária.

A principal preocupação está em como gerenciar essa crescente quantidade de dados. Essa preocupação fundamenta-se na premissa que os dados de uma empresa é um de seus maiores patrimônios. Na verdade, os dados armazenados durante anos de trabalho possuem implicitamente boa parte da memória corporativa da empresa. Por exemplo, se um analista financeiro de um banco trabalhou durante anos na concessão de empréstimos, os dados registrados por ele devem conter implicitamente as direções que o levaram a tomada da decisão (conceder ou não o empréstimo). Portanto, uma transformação desses dados em conhecimento pode proporcionar um auxílio inteligente à tomada de decisões dentro da empresa.

A transformação de dados em conhecimento vem sendo estudada ao longo de vários anos e pode ser feita de duas formas: manual ou automática. A transformação manual dos dados em conhecimento [Rezende and Pugliesi, 1998] é um trabalho caro e lento, e com a crescente quantidade de dados sendo armazenados, esse método esta se tornando inviável. A transformação automática de dados em conhecimento é um método que vêm crescendo a cada dia, principalmente com o surgimento da área de pesquisa nomeada Data Mining — DM [Fayyad et al., 1996b, Fayyad et al., 1996c, Fayyad et al., 1996d, Fayyad et al., 1996a].

Essa transformação automática de dados em conhecimento pode ser dividida em dois casos. O primeiro caso é utilizar o conhecimento extraído para a predição de novos casos, *i.e.* no exemplo do analista financeiro, se um novo cliente requisitar empréstimo, basta que os dados desse cliente seja submetido ao conhecimento extraído para obter a resposta da concessão (o conhecimento extraído funciona como uma "caixa preta"). No segundo caso o conhecimento

extraído é utilizado para explicar as tomadas de decisões, ou ainda, mostrar padrões de comportamento nos dados. Nesse segundo caso o conhecimento extraído deve ser inteligível para o ser humano, para que este consiga entender o que o conhecimento extraído representa em relação aos dados.

Desde o aparecimento da Teoria de Rough Sets sua aplicação vem sendo proposta em várias áreas da Inteligência Artificial, tais como modelagem de conjuntos, tratamento de incerteza, lógica aplicada, etc. Uma das aplicações mais recomendadas em artigos relacionados à Rough Sets é em Data Mining. Neste artigo é apresentada a Teoria de Rough Sets e alguns obstáculos que o processo de Data Mining deve superar. É apresentada ainda uma breve discussão e um posicionamento sobre a aplicação de Rough Sets em Data Mining.

2 Teoria de Rough Sets

A Teoria Rough Sets — RS — é motivada pela necessidade prática de interpretar, caracterizar, representar e processar o não-discernimento entre elementos [Pawlak, 1982]. A premissa central na filosofia de Rough Sets é que o conhecimento consiste na habilidade de classificar objetos [Slowinski, 1995]. Por exemplo, se um grupo de pacientes é descrito utilizando vários sintomas, então muitos pacientes compartilham os mesmos sintomas e, portanto, são indistinguíveis em relação a esses sintomas. Na Teoria de Rough Sets os exemplos do mundo real são expressos por um sistema de decisão formado por atributos, exemplos e a decisão associada.

2.1 Sistemas de Informação

A forma mais comum para representação dos dados na abordagem RS é um *sistema de informação*, o qual contém um conjunto de *objetos* descritos através de *atributos* e *valores* relacionados a cada um desses atributos.

Definição 2.1 (Sistema de Decisão). *Um* Sistema de Decisão *— SD — é um par ordenado* $\mathcal{A} = (U, A \cup \{d\})$ *onde* U *é um conjunto finito e não-vazio de objetos chamado de* Universo, A *é um conjunto finito e não-vazio de elementos chamado de* Atributos *e* $d \notin A$ *é o* atributo de decisão. *Os elementos do Universo serão referenciados como* objetos. *Cada atributo* $a \in A$ *é uma função total* $a : U \rightarrow V_a$*, onde* V_a *é o conjunto dos valores permitidos para o atributo* a *(sua faixa de valores). Os elementos de A são chamados de* atributos condicionais *ou simplesmente* condições (por exemplo, ver Tabela 1)

	Atributos			**Decisão**
Exemplos	*Estudos*	*Educação*	*Trabalha*	*Renda*
e_1	não	boa	sim	alta
e_2	não	boa	sim	alta
e_3	sim	boa	sim	nenhuma
e_4	não	pobre	não	baixa
e_5	não	pobre	não	média

Tabela 1: Sistema de Decisão

2.2 Distinguindo Objetos

A próxima definição introduz o conceito de *relação de não-discernimento*. Se tal relação existe entre dois objetos, isso significa que todos os valores de seus atributos são idênticos com respeito aos atributos sendo considerados, portanto não podem ser distinguidos entre si considerando esses atributos.

2.2.1 Relação de Não-Discernimento

Uma relação binária $R \subseteq X \times X$, a qual é reflexiva (*i.e.*, um objeto está relacionado com ele próprio xRx), simétrica (se xRy então yRx) e transitiva (se xRy e yRx então xRz), é chamada de *relação de equivalência*. A *classe de equivalência* de um elemento $x \in X$ consiste de todos os objetos $y \in X$ para os quais xRy.

Definição 2.2 (Relação de Não-Discernimento). *Para cada subconjunto de atributos $B \subseteq A$ no SI $\mathcal{A} = (U, A)$, uma relação de equivalência IND$_\mathcal{A}$(B) é associada, chamada de* Relação de Não-Discernimento*, e é definida como segue:*

$$IND_\mathcal{A}(B) = \{(x, y) \in U^2 \mid \forall a \in B, a(x) = a(y)\}$$

na qual IND$_\mathcal{A}$(B) é chamado de relação de não-discernimento-B. O conjunto de todas as classes de equivalência na relação IND$_\mathcal{A}$(B) é denotado por U/IND$_\mathcal{A}$(B).

Para o sistema de decisão representado na Tabela 1, os possíveis subconjuntos não-vazios dos atributos condicionais são: {*Estudos*}, {*Educação*}, {*Trabalha*}, {*Estudos, Educação*}, {*Estudos, Trabalha*}, {*Educação, Trabalha*} e {*Estudos, Educação, Trabalha*}. Considerando por exemplo o subconjunto {*Educação*}, os objetos e_1, e_2 e e_3 estão na mesma classe de equivalência e são não-discerníveis, assim como os objetos e_4 e e_5.

Pode-se notar que para cada subconjunto de atributos os objetos são agrupados e os grupos consistem de objetos que não podem ser discernidos entre si quando utilizado esse subconjunto de atributos. Segundo a Teoria de RS, cada um desses grupos é uma classe. Por exemplo, as classes para o subconjunto {*Estudos, Educação, Trabalha*} estão representadas na Tabela 2. A classe E_1 originou-se dos objetos e_1 e e_2, a classe E_2 originou-se do objeto e_3 e a classe E_3 originou-se dos objetos e_4 e e_5. Note ainda que a classe E_3 possui dois objetos com diferentes valores no atributo de decisão.

	Atributos		
Classes	*Estudos*	*Educação*	*Trabalha*
E_1	não	boa	sim
E_2	sim	boa	sim
E_3	não	pobre	não

Tabela 2: Classes para B={*Estudos, Educação, Trabalha*}

2.2.2 Matriz de Discernimento

Uma *Matriz de Discernimento* é uma matriz na qual as classes são índices e os atributos condicionais que podem ser utilizados para distinguir entre as classes são inseridos na linha e coluna correspondente às classes a serem discernidas.

Definição 2.3 (Matriz de Discernimento). *Para um conjunto de atributos $B \subseteq A$ em $\mathcal{A} = (U, A)$, a* Matriz de Discernimento *é dada por $M_D(B) = \{m_D(i,j)\}_{n \times n}, 1 \leq i, j \leq n, com\ n = |U/IND(B)|$, onde*

$$m_D(i,j) = \{a \in B \mid a(E_i) \neq a(E_j)\}\ para\ i, j = 1, 2, ..., n$$

O elemento $m_D(i,j)$ na matriz de discernimento é o conjunto de atributos de B que discerne (distingue) as classes de objetos $E_i, E_j \in U/IND(B)$.

Para a Tabela 2, pode-se observar que o único atributo com valor diferente para as classes E_1 e E_2 é *Estudos*. No caso das classes E_1 e E_3 são dois os atributos com valores diferentes, *i.e.* $\{$*Educação, Trabalha*$\}$.

2.2.3 Função de Discernimento

Definição 2.4 (Função de Discernimento). *A* Função de Discernimento *f(B) de um conjunto de atributos $B \subseteq A$ de um sistema de decisão é a função booleana*

$$f(B) = \bigwedge_{i,j \in \{1,...,n\}} \bigvee \overline{m}_D(E_i, E_j)$$

onde n = |U/IND(B)|, e $\bigvee \overline{m}_D(E_i, E_j)$ é a disjunção sobre o conjunto de variáveis booleanas $\overline{m}_D(E_i, E_j)$ que correspondem ao elemento $m_D(i,j)$ da matriz de discernimento.

Isso implica que a função de discernimento $f(B)$ computa o conjunto mínimo de atributos necessários para discernir qualquer classe de equivalência de todas as demais.

Definição 2.5 (Dispensável). *Um atributo a é* dispensável *ou* supérfluo *ou* redundante *em $B \subseteq A$ se IND(B) = IND(B – $\{a\}$), caso contrário é* indispensável *em B. Se todos os atributos $a \in B$ são indispensáveis em B, então B é chamado* ortogonal.

Considerando o exemplo da Tabela 1, no qual B=$\{$*Estudos, Educação, Trabalha*$\}$, tem-se que IND(B) = IND(*B* – $\{$*Trabalha*$\}$) = IND(*B* – $\{$*Educação*$\}$).

2.2.4 Redução da Representação

Os dados em um sistema de informação podem ser utilizados para discernir classes somente até um certo grau. Contudo, nem todos os atributos podem ser necessários para desempenhar essa tarefa. Em razão desse fato, a próxima definição é importante.

Definição 2.6 (Reduto). *Um* Reduto *de B é um conjunto de atributos $B' \subseteq B$ tal que todos os atributos $a \in B - B'$ são dispensáveis e IND(B') = IND(B). O termo RED(B) é utilizado para denotar a família de redutos de B. O conjunto de primos implicantes da função de discernimento f(B) determina os redutos de B.*

Para encontrar os redutos do exemplo considerado, a função de discernimento é utilizada. A função é minimizada no formato de soma de produtos, como mostrado na Equação 1. Isso resulta em redutos mínimos porque cada função de discernimento foi minimizada. Um reduto mínimo é portanto um reduto no qual nenhum dos atributos pode ser removido sem modificar as propriedades do reduto [Solheim and Aasheim, 1996]. Na Equação 1 existem dois redutos unidos na forma de uma disjunção.

$$f(E) = (Estudos \ \wedge \ Educação) \ \vee \ (Estudos \ \wedge \ Trabalha) \quad (1)$$

Deve ser observado que computar classes de equivalência é um processo simples. Entretanto, encontrar redutos mínimos, ou seja, redutos que têm cardinalidade mínima entre todos os redutos, é um problema *NP-hard*. Na realidade, o cálculo de redutos é considerado o maior problema na abordagem de Rough Sets. Afortunadamente, existem algumas heurísticas que permitem computar um número suficiente de redutos em tempo aceitável, sempre que o número de atributos não for muito grande [Komorowski et al., 1999].

2.3 De Redutos para Regras

Quando os redutos são encontrados, o trabalho de se definir regras para os valores de decisão com base nos atributos condicionais está praticamente feito, basta unir os valores dos atributos condicionais da classe de objetos da qual foi originado o reduto com os atributos correspondentes ao reduto. Então, para completar a regra, a decisão é adicionada ao final da regra. As regras para o exemplo da Equação 1 são:

$$\begin{array}{l} Estudos = não \wedge \textsf{Educação} = boa \ \longrightarrow \ Renda = alta \\ Estudos = não \ \wedge \ Trabalha = sim \longrightarrow \ Renda = alta \\ Estudos = sim \wedge \textsf{Educação} = boa \ \longrightarrow \ Renda = nenhuma \\ Estudos = sim \ \wedge \ Trabalha = sim \longrightarrow \ Renda = nenhuma \\ Estudos = não \wedge \textsf{Educação} = pobre \ \longrightarrow \ Renda = ? \\ Estudos = não \ \wedge \ Trabalha = não \longrightarrow \ Renda = ? \end{array}$$

As últimas duas regras não especificam o valor do atributo *Renda*, pois o valor desse atributo não é o mesmo para todos os objetos da classe. Ele pode ser chamado de categoria imprecisa. Uma forma melhor de apresentar esse tipo de regra sem utilizar um sinal de interrogação é dizer que quando a *Educação* é pobre, existe uma chance de 50% de que a *Renda* seja baixa, e existe uma chance de 50% de que a *Renda* seja média. Deve-se notar que para um conjunto de dados contendo cinco elementos, foram extraídas seis regras.

3 Obstáculos em Data Mining

O interesse pela descoberta de possíveis relações e padrões nas informações contidas em bases de dados tem motivado o fomento de grande quantidade de pesquisas na área de Data Mining. Essas pesquisas são financiadas em grande parte por empresas privadas que buscam desenvolver ferramentas capazes de automatizar o processo de Data Mining. Estima-se que a quantidade de informações armazenadas somente no ano de 1999 foi superior em 12% do que aquela armazenada em toda a história da humanidade [Lesk, 1997].

Assim, métodos computacionais capazes de lidar com *Gigabytes* de informação são um desafio a ser superado. Na verdade, é desejável que o conhecimento extraído desse montante de informações seja o mais simples possível do ponto de vista da compreensão. Por exemplo, a base de dados *EEG Data* presente no Repositório KDD da UCI [Bay et al., 2000] que possui aproximadamente 3G de informações, cujo objetivo é descobrir uma correlação entre dados provenientes de exames de eletroencefalograma (EEG) e a predisposição genética para alcoolismo. Embora essa base seja grande, o conhecimento extraído — relação entre EEG e

alcoolismo — deve ser inteligível ao humano, para que este possa compreender e avaliar essa possível relação.

Existem alguns métodos capazes de descobrir essa relações nas informações, tais como redes neurais, árvores de decisão e regras de produção. No entanto, a maior parte desses métodos é muito sensível a quantidade de dados a eles fornecidos. Essa sensibilidade em relação ao montante de informações resulta em processos computacionais não viáveis e conhecimento representado em hipóteses de difícil compreensão para o ser humano.

Outro problema ainda em aberto é a escalabilidade. Alguns métodos computacionalmente viáveis e que resultam em hipóteses compreensíveis, esbarram no problema de produzirem hipóteses cada vez mais complexas conforme o montante de dados aumenta. Têm sido propostos alguns métodos para superar o problema da escalabilidade, tal como relatado em [Diamantini and Panti, 2000, Garofalakis and Ratogi, 2000].

Na verdade existem muitos outros obstáculos a serem superados, como extrair conhecimento na presença de dados desbalanceados e de grande quantidade de valores ausentes [Batista et al., 1999, Batista, 2000], extrair conhecimento na presença de informações de baixa frequência [Zadrozny and Elkan, 2001], entre outros.

4 Rough Sets e Data Mining

Desde o aparecimento da Teoria de Rough Sets sua aplicação vem sendo proposta em várias áreas da Inteligência Artificial. Uma das aplicações mais recomendadas em artigos relacionados à Rough Sets é em Data Mining [Lin and Cercone, 1997], cujo objetivo principal é a extração de padrões presentes em grandes bases de dados. Levando em consideração os obstáculos brevemente descritos na seção anterior, foram feitos alguns experimentos sobre 9 conjuntos de dados provenientes do Repositório da UCI [Blake et al., 1998], os quais são brevemente descritos a seguir:

TA: Este conjunto de dados consiste em medidas da qualidade do ensino num período de três semestres regulares e dois semestres de verão. As medidas são relativas a 151 professores assistentes do Departamento de Estatística da Universidade de Wisconsin – Madison.

Bupa: Este conjunto de dados consiste em predições de quando um paciente tem ou não desordens no fígado com base em vários testes sanguíneos e no consumo de álcool.

Pima: Neste conjunto de dados todos os pacientes são mulheres com idade mínima de 21 anos e pertencentes à linhagem de Índios Pima que vivem próximos a Phoenix, Arizona, USA. O problema é predizer quando uma paciente terá resultado positivo para o teste de diabetes.

Breast-cancer2: Este conjunto de dados é um dos conjuntos nomeados Breast Cancer que estão na UCI, no qual o problema é predizer sobre a recorrência de cancer de mama.

CMC: Os exemplos presentes neste conjunto de dados são relativos a mulheres casadas que não estavam grávidas ou não sabiam se estavam grávidas no momento da entrevista. O problema consiste em predizer o método contraceptivo escolhido por cada mulher (nenhum, método a curto prazo, método a longo prazo) com base nas características demográficas e sócio-econômicas de cada uma delas.

Breast-cancer: Neste conjunto de dados o problema é predizer quando uma amostra de tecido da mama extraído de uma paciente possui tumor benigno ou maligno.

Smoke: Este conjunto de dados está relacionado ao problema de predizer atitudes resultantes da restrição ao fumo em locais de trabalho (proibição, restrição, sem restrição) com base em leis, ambiente ou variáveis sócio-econômicas.

Hungarian: Neste conjunto de dados os exemplos são relativos a diagnósticos de doenças cardíacas.

Hepatitis: O conteúdo deste conjunto de dados esta relacionado a predição da expectativa de vida de pacientes com hepatite.

Na Tabela 3 é apresentado um resumo das principais características de cada um dos conjuntos de dados utilizados neste trabalho. É mostrado, o número de exemplos (#Exemplos), número e percentual de exemplos duplicados (aparecem mais que uma vez) ou conflitantes (possuem o mesmo conjunto atributo-valor mas diferente classe de decisão), número de atributos (#Atributos) contínuos e nominais, o erro majoritário e se o conjunto de dados tem ao menos um valor ausente.

Conjuntos de dados	#Exemplos	Dupl. ou Confl. (%)	#Atributos (cont.,nom.)	Classe	%Classe	Erro Majoritário	Vrs. Aus.
ta	151	45 (39.13%)	5 (1,4)	1 2 3	32.45% 33.11% 34.44%	65.56% 3	N
bupa	345	4 (1.16%)	6 (6,0)	1 2	42.03% 57.97%	42.03% 2	N
pima	769	1 (0.13%)	8 (8,0)	0 1	65.02% 34.98%	34.98% 0	N
breast-cancer2	285	2 (0.7%)	9 (4,5)	recurrence no-recurrence	29.47% 70.53%	29.47% no-recurrence	S
cmc	1473	115 (7.81%)	9 (2,7)	1 2 3	42.70% 22.61% 34.69%	57.30% 1	N
breast-cancer	699	8 (1.15%)	9 (9,0)	2 4	65.52% 34.48%	34.48% 2	S
smoke	2855	29 (1.02%)	13 (2,11)	0 1 2	5.29% 25.18% 69.53%	30.47% 2	N
hungarian	294	1 (0.34%)	13 (13,0)	presence absence	36.05% 63.95%	36.05% absence	S
hepatitis	155	0 (0%)	19 (6,13)	die live	20.65% 79.35%	20.65% live	S

Tabela 3: Características dos Conjuntos de Dados

É possível notar através da Tabela 3 que os conjuntos de dados utilizados nos experimentos não podem ser classificados como pertencentes a tarefa de Data Mining, pois conjuntos de dados como o EEG possuem milhões de registros, enquanto que os anteriormente descritos possuem algumas centenas. Contudo, esses conjuntos de dados são úteis para demonstrar o comportamento da Teoria de Rough Sets na tarefa de extração de conhecimento.

Os experimentos foram feitos sem que os algoritmos sofressem quaisquer modificações em suas configurações padrão. Para a execução dos experimentos foram utilizados os indutores $\mathcal{CN}2$ e $\mathcal{C}4.5$-rules pertencentes à Biblioteca $\mathcal{MLC}^{++}$ [Kohavi et al., 1996], bem como a ferramenta Rosetta [Øhrn, 1999] que implementa os conceitos referentes à Teoria de Rough Sets.

Cada conjunto de dados foi submetido aos indutores descritos anteriormente para que estes pudessem induzir o conhecimento no formato de regras. O número de regras geradas utilizando cada um dos indutores está na Tabela 4. Deve-se notar que o número de regras geradas pelo Rosetta supera a quantidade gerada por qualquer um dos outros dois indutores. Pode-se notar que na maior parte dos casos o número de regras induzidas pelo Rosetta é pelo menos duas vezes maior que o número de regras induzidas pelos outros métodos.

Conjunto de Dados	$\mathcal{CN}2$	$\mathcal{C}4.5$-rules	Rosetta
ta	61	17	103
bupa	34	11	110
pima	56	6	114
breast cancer2	40	12	275
cmc	174	36	1345
breast cancer	18	8	30
smoke	426	22	2826
hungarian	25	11	287
hepatitis	19	10	77

Tabela 4: Número de Regras

Contudo, analisar somente o número de regras induzidas pode levar a conclusões precipitadas a respeito da hipótese. Essa medida deve ser analisada em conjunto com com o erro na classificação. O erro na classificação utilizando *10-fold-cross-validation* é uma estimativa do erro esperado na classificação de novos exemplos, o qual é apresentado na Tabela 5. Em alguns casos, apesar da hipótese ser complexa (muitas regras) o erro na classificação é baixo. Dessa forma, embora a hipótese induzida não possa ser utilizada de forma inteligível, esta pode ser utilizada como preditor. No caso do Rosetta, além de gerar hipóteses muito complexas, o erro na classificação foi muito alto. Na verdade, em alguns casos o erro obtido foi maior que o erro da classe majoritária. Nesses casos um preditor que simplesmente "apostasse" que um novo exemplo pertence à classe majoritária sairia-se melhor.

Conjunto de Dados	$\mathcal{CN}2$	$\mathcal{C}4.5$-rules	Rosetta
ta	51.67±3.42	53.58±6.00	53.58±14.20
bupa	35.35±2.01	34.13±2.85	50.00±9.45
pima	25.12±1.97	25.87±1.07	45.78±6.34
breast cancer2	27.03±2.29	27.71±1.73	56.78±12.64
cmc	49.64±1.01	45.90±1.38	72.50±4.45
breast cancer	4.87±0.77	4.29±0.60	8.58±4.14
smoke	32.18±0.64	32.54±0.68	71.82±6.80
hungarian	21.44±2.19	20.05±2.90	47.30±9.47
hepatitis	16.18±1.80	20.54±3.02	34.20±14.82

Tabela 5: Erro na Classificação Utilizando *10-fold-cross-validation*

Na Tabela 5 pode-se notar que, em todos os casos, o erro obtido pelo Rosetta é superior àquele obtido por qualquer um dos outros dois indutores. Em alguns casos, o erro chega a

ser maior que o dobro daquele obtido por qualquer um dos outros dois indutores. O erro na classificação e o número de regras induzidas são duas fortes considerações a serem feitas na escolha de um método para extração de conhecimento.

O maior problema na Teoria de Rough Sets é a admitir que as informações são dadas por atributos categóricos. No entanto, sabe-se que o contrário é o mais comum, ou seja, a maior parte das informações existentes nas bases de dados são constituídas de atributos numéricos. Nesse caso, para aplicar a Teoria de Rough Sets é necessário que o conjunto de dados passe por um processo de discretização que não faz parte da metodologia proposta surgida com RS. Sabe-se que no processo de discretização sempre há perda de informação, pois os valores dos atributos passam a estar relacionados a intervalos determinados pelos pontos de "corte" [Félix et al., 2000]. Assim, valores originalmente distintos, passam a ser tratados como iguais após o processo de discretização. Logicamente, nesse caso, o conhecimento implícito é perdido.

Na verdade, o número de regras induzidas pelo Rosetta esta diretamente ligado ao processo de discretização. No ponto extremo, a indução das regras ocorre sem que o conjunto de dados tenha sido discretizado, e neste caso o Rosetta induz uma quantidade de regra equivalente ao número de exemplos no conjunto de dados, ou seja, um *look-up*.

Outro problema ainda relacionado às regras é a forma de induzi-las. Na Teoria de Rough Sets a tarefa primordial do aprendizado — a generalização — é negligenciada. A indução das regras é feita de forma que após calcular os redutos, estes são simplesmente sobrepostos sobre o conjunto de dados que foi utilizado para calcula-los. Assim, a indução de regras se resume em mapear os exemplos do conjunto de dados em regras, cujo antecedente é formado pela disjunção de atributos e valores e o conseqüente é o valor da decisão para aquele exemplo em particular. Na verdade, o classificador possui bom desempenho somente se o conjunto de dados contiver alguns exemplos capazes de representar todos os demais, caso contrário as regras geradas não são capazes de representar o conhecimento de forma generalizada, como é o caso das regras induzidas pelos indutores $\mathcal{CN}2$ e $\mathcal{C}4.5$-rules .

Com relação ao erro na classificação de novos exemplos, os percentuais encontrados para a Teoria de Rough Sets justificam-se pelo fato do conhecimento extraído ter, embora com grande quantidade de regras, baixo poder de generalização. Assim, quando novos exemplos são submetidos ao conhecimento extraído, este possui baixa taxa de predição da classe correta.

No entanto, existem outros obstáculos em Data Mining para os quais a Teoria de Rough Sets é uma possível escolha. Em [Pila and Monard, 2001, Pila, 2001] são relatados experimentos utilizando a noção de redutos da Teoria de Rough Sets como método para filtrar atributos relevantes. Nesses trabalhos estão relatados que os redutos podem ser utilizados como forma para selecionar atributos relevantes, uma vez que essa abordagem acabou por selecionar a menor quantidade de atributos, quando comparada com outros métodos.

5 Conclusões

A Teoria de Rough Sets é uma teoria que surgiu no início da década de 80, cujo idealizador é o pesquisador polonês Zdzislaw Pawlak. Essa teoria esta matematicamente muito bem fundamentada e pode ser aplicada em muitas outras áreas da computação [Pawlak et al., 1995]. Essa teoria possui duas principais funcionalidades para a área de DM: calcular os atributos relevantes segundo a relevância presente no conceito da relação de não-discernimento e induzir

as regras utilizando os redutos.

O problema principal desta teoria é admitir que as informações presentes nos conjuntos de dados devam ser categóricas. Isso faz com que o processo do calculo dos redutos e a posterior indução das regras fique isolado do passo de discretização, não ocorrendo em conjunto como nos indutores $\mathcal{CN}2$ e $\mathcal{C}4.5$-rules que geram árvore de decisão e regras de produção, respectivamente. Sabe-se que nenhum método de discretização é capaz de manter o mesmo conhecimento implícito que os dados em sua forma original. Assim, quando do cálculo dos redutos ou da indução das regras, algum conhecimento que estava presente nos dados já se perdeu e o conhecimento extraído no final não retrata bem o que os dados contém.

Outro problema é a indução das regras que é feita apenas sobrepondo o reduto no conjunto de dados original. Existem outros métodos capazes de construir hipóteses muitos mais simples que aquelas construídas utilizando a Teoria de Rough Sets, e assim, do ponto de vista de DM, os primeiros são preferíveis.

Referências

[Batista, 2000] Batista, G. E. A. P. A. (2000). Pré-processamento de dados em aprendizado de máquina supervisionado. Minidissertação para Qualificação de Doutorado, ICMC-USP.

[Batista et al., 1999] Batista, G. E. A. P. A., Carvalho, A. C. P. L., and Monard, M. C. (1999). Aplicando seleção unilateral em conjuntos de exemplos desbalanceados: Resultados iniciais. In *Anais II Encontro Nacional de Inteligência Artificial - ENIA 99*, pages 327–340.

[Bay et al., 2000] Bay, S. D., Kibler, D., Pazzani, M. J., and Smyth, P. (2000). The uci kdd archive of large data sets for data mining: Research and experimentation. *SIGKDD Explorations*, pages 81–85.

[Blake et al., 1998] Blake, C., Keogh, E., and Merz, C. (1998). Uci irvine repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[Diamantini and Panti, 2000] Diamantini, C. and Panti, M. (2000). An efficient and scalable data compression approach to the classification task. *SIGKDD Explorations*, pages 49–55.

[Øhrn, 1999] Øhrn, A. (1999). Rosetta: Technical reference manual. Technical report, Knowledge System Group, Norwegian University on Science and Technology, NO.

[Fayyad et al., 1996a] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI Magazine*, Fall:37–54.

[Fayyad et al., 1996b] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996b). Advances in knowledge discovery and data mining. *American Association for Artificial Intelligence*.

[Fayyad et al., 1996c] Fayyad, U. M., Platestsky-Shapiro, G., and Smyth, P. (1996c). *From Data Mining to Knowledge Discovery: An Overview*, pages 1–30. In [Fayyad et al., 1996b].

[Fayyad et al., 1996d] Fayyad, U. M., Platestsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996d). *Advances in Knowledge Discovery and Data Mining*. In [Fayyad et al., 1996b].

[Félix et al., 2000] Félix, L. C. M., Rezende, S. O., Monard, M. C., and Caulkins, C. W. (2000). Transforming a regression problem into a classification problem using hybrid discretization. *Computación y Sistemas. Special issue in Artificial Intelligence*, pages 44–52.

[Garofalakis and Ratogi, 2000] Garofalakis, M. and Ratogi, R. (2000). Scalable data mining with model constraints. *SIGKDD Explorations*, pages 41–48.

[Kohavi et al., 1996] Kohavi, R., Sommerfield, D., and Dougherty, J. (1996). Data mining using $\mathcal{MLC}^{++}$: A machine learning library in C^{++} . *Tools with IA*, pages 234–245.

[Komorowski et al., 1999] Komorowski, J., Pawlak, Z., Polkowski, L., and Skowron, A. (1999). Rough sets: A tutorial. Technical report, Warsaw University.

[Lesk, 1997] Lesk, M. (1997). How much information is there in the world? Technical report. http://www.lesk.com/mlesk/ksg97/ksg.html.

[Lin and Cercone, 1997] Lin, T. Y. and Cercone, N. (1997). *Rough Sets and Data Mining: Analysis of Imprecise Data*. Kluwer Academic Publishers.

[Pawlak, 1982] Pawlak, Z. (1982). Rough sets. *International Jornal of Computer and Information Sciences*, pages 341–356.

[Pawlak et al., 1995] Pawlak, Z., Grzymala-Busse, J., Slowinski, R., and Ziarko, W. (1995). Rough sets. *Comunications of the ACM*, pages 89–95.

[Pila, 2001] Pila, A. D. (2001). Seleção de atributos relevantes para aprendizado de máquina utilizando a abordagem de rough sets. Dissertação de Mestrado, ICMC-USP.

[Pila and Monard, 2001] Pila, A. D. and Monard, M. C. (2001). Rough sets reducts as a filter approach for feature subset selection: An empirical comparison with wrapper and other filters. Technical Report 134, ICMC-USP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_134.ps.zip.

[Rezende and Pugliesi, 1998] Rezende, S. O. and Pugliesi, J. B. (1998). Aquisição de conhecimento explícito ou manual. Technical Report 37, ICMC-USP. http://labic.icmc.sc.usp.br/didatico/PostScript/rt_ac.ps.zip.

[Slowinski, 1995] Slowinski, R. (1995). Rough set approach to decision analysis. *AI Expert*, March:19–25.

[Solheim and Aasheim, 1996] Solheim, H. G. and Aasheim, O. T. (1996). Rough sets as a framework for data mining. Technical report, The Norwegian University of Science and Technology, NO.

[Zadrozny and Elkan, 2001] Zadrozny, B. and Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown. Technical Report CS2001-0664, UCSD.