Seleção de Atributos no Escopo da Teoria de Conjuntos Aproximativos

Alex Sandro Aguiar Pessoa[1], Stephan Stephany[2]

[1]Programa de Mestrado ou Doutorado em Computação Aplicada – CAP
Instituto Nacional de Pesquisas Espaciais – INPE

[2]Laboratório Associado de Computação e Matemática Aplicada – LAC
Instituto Nacional de Pesquisas Espaciais – INPE

asapessoa@gmail.com, stephan@lac.inpe.br

Abstract. *This paper uses the Rough Set Theory (RST) in problem solving of Feature Selecting. RST is used mainly in data mining, in the treatment of uncertainty and vagueness of data. But implicitly performs a Feature Selection by means of so-called reductions.*

Resumo. *Este trabalho trata da aplicação da Teoria dos Conjuntos Aproximativos (TCA) na solução do problema de Seleção de Atributos. A TCA é empregada principalmente em mineração de dados, no tratamento de incerteza e imprecisões nos dados. Porém implicitamente realiza uma Seleção de Atributos, por meio das chamadas reduções.*

Palavras-chave: *seleção de atributos, teoria dos conjuntos aproximativos, metas-heurísticas.*

1. Introdução

Este trabalho propõe a aplicação da Teoria dos Conjuntos Aproximativos (TCA) como técnica de seleção de atributos. A TCA tem uma característica bem proeminente a compactação de bases de dados, que é obtida por meio de dois conceitos: a *relação de indiscernibilidade* e a *redução*. Embora sejam definições distintas estas estão intimamente ligadas, no que diz respeito ao tratamento de incertezas e compactação de bases de dados. Enquanto a indiscernibilidade reduz uma base de dados com relação ao número de objetos, a redução promove a compactação da base de dados, sob o ponto de vista da eliminação de atributos/variáveis redundantes. Esta busca pelos atributos mais "significativos" é conhecida como um problema de *Seleção de Atributos*, que é uma área de aprendizado de máquinas e estatística, responsável pela seleção de um subconjunto de atributos relevantes para a construção de modelos de aprendizado robustos.

Tendo em vista que o grande gargalo da TCA é o cálculo das reduções, este trabalho, propõe o uso de três metas-heurísticas para busca de reduções, sendo estas: *Variable Neighborhood Search* (VNS*), Variable Neighborhood Decent* (VND) e *Decrescent Cardinality Search* (DCS). As duas primeiras metas-heurísticas são bem conhecidas na literatura [Mladenovic et. al 1997] e a última foi proposta neste trabalho, onde é uma variação da VNS, porém mais agressiva, onde o intuito é sempre forçar a busca por soluções com menor cardinalidade. Os resultados são comparados com os obtidos e mostrados em Hedar et, al (2006) para os algoritmos de Busca Tabu, Colônia de Formigas, Recozimento Simulado e Algoritmo Genético.

2. Teoria dos Conjuntos Aproximativos

O conhecimento em TCA é representado na forma tabular, representado pelo par ordenado $S = (U; A)$, chamado de *sistema de informação*, onde U é um conjunto finito não-vazio de objetos chamado de universo e A é um conjunto finito não-vazio de atributos condicionais ou condições, tal que a: $U \rightarrow V_a$ para todo $a \in A$. O conjunto V_a é chamado de conjuntos de valores de a. Uma forma particular do sistema de informação é adicionando um atributo distinto dos atributos condicionais, com o objetivo de criar classes. Essa forma do sistema de informação é chamada de *sistema de decisão*, onde o $S = (U; A \cup \{d\})$, onde $d \notin A$ é o atributo de decisão.

A relação de indiscernibilidade diz que dois elementos são indiscerníveis se são iguais segundo um conjunto de atributos $B \subseteq A$. Evidentemente esta relação forma classes de equivalências, denotadas por $[x]_B$ para $x \in X$ e $X \subseteq U$. Essas classes de equivalência induzem um particionamento do conjunto U. A aproximação inferior, que é uma partição de U, onde é constituída por objetos que certamente pertencem a uma determinada classe. Seja $B \subseteq A$, $x \in X$ e $X \subseteq U$, a aproximação inferior é dada por:

$$\underline{B}X = \{x \mid [x]_B \subseteq X\} \quad (2)$$

Com base na aproximação inferior é possível mensurar a dependência entre dois subconjuntos de atributos. Essa medida é conhecida como *grau de dependência*. Sejam $IND(B)$ e $IND(C)$ relações de indiscernibilidade em U e os conjuntos $B \subseteq A$ e $C \subseteq A$, o grau de dependência entre B e C é dado por:

$$\gamma_B(C) = \frac{|POS_B(C)|}{|U|} \quad (3)$$

onde $POS_B(C) = \bigcup_{X_i \in U/IND(C)} \underline{B}X_i$ e i = 1..|C| é conhecida como região positiva. Se $\gamma_B(C) = 1$ então C é totalmente dependente de B. Se $\gamma_B(C) < 1$ diz-se que C e parcialmente dependente de B.

3. Metas-Heurísticas

Uma meta-heurística é um método heurístico para resolver de forma genérica problemas de busca e otimização, sendo sua aplicação geralmente associada a problemas dos quais não se conhece a solução, como no caso dos problemas NP-difíceis.

As soluções codificadas, neste trabalho, para as metas-heurísticas são representadas na forma de uma cadeia de zeros e uns, onde cada posição desta cadeia é relativa a um atributo. O zero representa a ausência do atributo na solução e o um é o caso contrário.

O método adotado para a busca local é simples, consistindo na busca da melhor solução, por meio da varredura dos vizinhos mais próximos da solução candidata e a função usada para avaliar as soluções é o grau de dependência.

3.1. VNS, VND e DCS

O VNS ou busca com vizinhança variável, é uma meta-heurística baseada num procedimento iterativo que compreende a geração aleatória de um novo vizinho a partir de uma solução inicial ou corrente e a busca local que corresponde à exploração da vizinhança deste novo vizinho, cuja melhor solução passa a ser a solução corrente. Isto permite explorar gradativamente vizinhanças mais distantes [Mladenovic et. al 1997],

[Hansen et al. 2003].

O VND ou busca com vizinhança variável em descida [Mladenovic et.al 1997] é uma variante do VNS, ou, mais especificamente, um caso particular. Este algoritmo explora a vizinhança de uma solução inicial, buscando o mínimo local através do gradiente.

O DCS ou busca de cardinalidade decrescente, é uma nova meta-heurística, proposta neste trabalho e derivada do VNS, cujo enfoque, diferentemente das outras metas-heurísticas aqui abordadas, que buscam soluções melhores na vizinhança próxima, é buscar uma nova solução candidata (s') qualquer que tenha necessariamente menor cardinalidade do que a solução corrente.

4. Metodologia

Foram analisados 5 conjuntos de dados, conforme a lista abaixo, onde $|A|$ denota a cardinalidade do conjunto de atributos e $|U|$ o número de objetos/elementos na base de dados.

- **Vote**: $|A| = 16$ e $|U| = 300$
- **Credit**: $|A| = 20$ e $|U| = 1000$
- **Mushroom**: $|A| = 22$ e $|U| = 8124$
- **Derm**: $|A| = 34$ e $|U| = 366$
- **Lung**: $|A| = 56$ e $|U| = 32$

Cada algoritmo foi executado 20 vezes com soluções iniciais diferentes que foram geradas aleatoriamente. A máquina utilizada possui um processador AMD® Athlon II X4 - 3GHz e 8 GB de memória RAM.

5. Resultados

Na Tabela 1 são mostrados os resultados para o problema de seleção de atributos utilizando a TCA com as diversas metas-heurísticas abordadas: colônia de formigas (Ant), recozimento simulado (SA), algoritmo genético (AG), Busca Tabu (TS), busca com vizinhança variável (VNS), busca com vizinhança variável em descida (VND), e busca com cardinalidade decrescente (DCS). Os resultados de Ant, SA, AG e TS são provenientes de Hedar, et. al (2006) e serviram de referência para aqueles obtidos pelas metas-heurísticas propostas.

Tabela 1. Resultados

Dados	$\|A\|$	TCA – Seleção de Atributos						
		Ant	**SA**	**AG**	**TS**	**VNS**	**VND**	**DCS**
Vote	16	$8^{(20)}$	$8^{(15)}9^{(5)}$	$8^{(2)}9^{(18)}$	$8^{(20)}$	$8^{(20)}$	$5^{(1)*}6^{(1)*}7^{(2)*}$ $8^{(3)}9^{(6)}$ $10^{(4)}11^{(3)}$	$8^{(3)}9^{(13)}$ $10^{(4)}$
Credit	20	$8^{(12)}9^{(4)}$ $10^{(4)}$	$8^{(18)}9^{(1)}$ $11^{(1)}$	$10^{(6)}$ $11^{(14)}$	$8^{(13)}9^{(5)}$ $10^{(2)}$	$7^{(17)}8^{(3)}$	$5^{(1)*}9^{(4)}$ $10^{(9)}11^{(5)}$ $12^{(1)}$	$7^{(5)}8^{(15)}$
Mushroom	22	$4^{(20)}$	$4^{(20)}$	$5^{(1)}6^{(5)}$ $7^{(14)}$	$4^{(17)}5^{(3)}$	$3^{(4)}4^{(15)}$ $5^{(1)}$	$8^{(9)}9^{(6)}10^{(4)}$ $14^{(1)}$	$3^{(14)}4^{(6)}$
Derm	34	$6^{(17)}7^{(3)}$	$6^{(12)}7^{(8)}$	$10^{(6)}$ $11^{(14)}$	$6^{(14)}7^{(6)}$	$10^{(1)}$ $11^{(19)}$	$11^{(1)}14^{(3)}15^{(5)}$ $16^{(8)}17^{(2)}18^{(1)}$	$10^{(5)}11^{(12)}$ $12^{(3)}$
Lung	56	$4^{(20)}$	$4^{(7)}5^{(12)}$ $6^{(1)}$	$6^{(8)}7^{(12)}$	$4^{(6)}5^{(13)}$ $6^{(1)}$	$4^{(10)}5^{(10)}$	$18^{(2)}19^{(3)}21^{(2)}$ $22^{(6)}24^{(2)}25^{(3)}$ $27^{(1)}33^{(1)}$	$3^{(7)}4^{(12)}$ $5^{(1)}$

onde $X^{(Y)}$, X denota a cardinalidade da solução e Y número de vezes que soluções com tal cardinalidade foi encontrada. * refere-se a soluções com graus de dependência menor do que 1 ($\gamma_B < 1$).

Na Tabela 2 são mostrados os resultados referente ao tempo médio gasto pelos algoritmos propostos neste trabalho.

Tabela 2. Tempos Médios de Execução (segundos)

Dados	$\|A\|$	VNS	VND	DCS
Vote	16	495	0,5	21
Credit	20	1825	4	448
Mushroom	22	559	6	1458
Derm	34	298	1	298
Lung	56	473	0,7	55

Em geral os resultados obtidos pelos métodos VNS, VND e DCS são equiparáveis aos resultados mostrados em Hedar et. Al (2006) e em algumas situações são melhores. Um exemplo é para base de dados *mushroom*, onde o melhor resultado do trabalho de Hedar et. Al (2006) é como o algoritmo de colônia de formigas (Ant) de valor $|A| = 4$. Já neste trabalho foi encontrado, para o algoritmo DCS a cardinalidade 3 em 14 situações, contra a cardinalidade 3 encontrada pelo VNS 4 vezes. Já os tempos mostram o VNS com um melhor desempenho neste caso.

6. Conclusões

Este trabalho mostrou o uso da Teoria dos Conjuntos Aproximativos aplicado ao problema de seleção de atributos. Para isso foram implementadas três metas-heurísticas, de forma inovadora na TCA, sendo estas o VNS, VND e DCS. Os resultados mostram a viabilidade do uso de tais métodos na busca de solução para reduções e consequentemente para a Seleção de Atributos.

Referências

Hansen, P.; Mladenovic, N (2003). A Tutorial on Variable Neighborhood Search, Les Cahiers du GERAD, HEC Montreal and GERAD.

Hedar, A.; Wang J.; Fukushima, M (2006). Tabu search for attribute reduction in rough set theory, Technical Report 2006-008, Department of Applied Mathematics and Physics, Kyoto University. 2006

Komorowski, J.; Polkowski, L.; Skowron (1999). A. Rough sets: A tutorial. in: S.K. Pal and A. Skowron (eds.), Rough fuzzy hybridization: A new trend in decision-making, Springer-Verlag, Singapore, 1999.

Mladenovic, N.; Hansen, P (1997). Variable Neighborhood Search. Computers and Operations Research, 24:1097-1100.

Øhrn, A (1999). Discernibility and Rough Sets in Medicine: Tools and Applications. Tese (Doutorado). Norwegian University of Science and Technology, Department of Computer and Information Science, NTNU.

Pawlak, Z (1982). Rough sets. International Journal of Computer and Information Sciences, vol. 11. p. 341-356..