Mineração de Dados Espaço-Temporal Aplicada a Previsão Climática Utilizando a Teoria dos Conjuntos Aproximativos

Alex Sandro Aguiar Pessoa
Instituto Nacional de Pesquisas Espaciais
Laboratório Associado de Matemática e Computação Aplicada
asapessoa@lac.inpe.br

José Demisio da Silva Simões
Instituto Nacional de Pesquisas Espaciais
Laboratório Associado de Matemática e Computação Aplicada
demisio@lac.inpe.br

Resumo

Apesar do comportamento caótico atmosférico, este artigo visa estabelecer relações, que busquem simplificar a extração de informações, entre o comportamento atmosférico real e o modelo de previsão climática utilizado pelo CPTEC por meio do processo de descoberta de conhecimento (KDD). A ferramenta matemática utilizada para extrair informações entre os dados é a Teoria dos Conjuntos Aproximativos (TCA), que trata basicamente da manipulação de informações incertas e redução de dados.

1. Introdução

A previsão climática no CPTEC é realizada através do modelo de circulação geral atmosférico (MCGA) por ensemble [1], que simplificadamente processa um conjunto de condições iniciais, através de modelo baseado em leis físicas, e obtém um resultado constituído por um número de membros igual ao número de condições iniciais, ou seja, para cada condição inicial há um membro como saída.

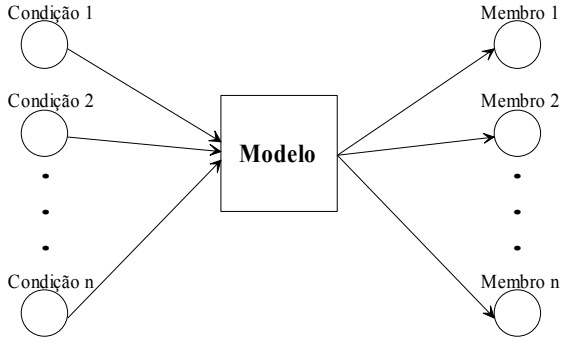

Figura 1 – Modelo de Previsão Climática

Então este trabalho tem como objetivo estabelecer quais membros são necessários para chegar ao melhor resultado (mais próximo do real), por meio de um processo de descoberta de conhecimento, mais conhecido como *KDD*.

O processo de KDD, dentre muitas etapas, tem a chamada mineração de dados, que como o próprio nome sugere é a tarefa responsável pela descoberta de padrões escondidos entre os dados. Existem muitas técnicas para fazer mineração de dados, mas como o enfoque deste trabalho é sobre a Teoria dos Conjuntos Aproximativos (TCA), ("*Rough Sets Theory*", do inglês) daremos mais ênfase a tal assunto.

2. Meteorologia

A meteorologia é definida como a ciência que estuda os fenômenos que ocorrem na atmosfera, e está relacionada ao estado físico, dinâmico, químico da atmosfera e as interações entre elas e a superfície terrestre subjacente.

Em meteorologia há uma distinção entre tempo e clima que são conceitos usados para se entender o comportamento da atmosfera em diferentes "intervalos de tempo". Assim, o clima representa uma generalização, enquanto o tempo lida com eventos específicos.

2.1. Previsão climática

A previsão climática é uma estimativa do comportamento médio da atmosfera com alguns meses de antecedência. Por exemplo, pode-se prever se o próximo verão será mais quente ou mais frio que o normal, ou ainda, mais ou menos chuvoso. Todavia, tal estimativa não pode dizer exatamente qual será a quantidade de chuvas ou quantos graus a temperatura estará mais ou menos elevada.

Para previsão climática, no CPTEC-INPE são utilizados modelos numéricos, alguns em caráter experimental, pois no Brasil e no mundo, essa é uma área

que está em constante evolução com o propósito de torná-la mais confiável.

Dentre os modelos numéricos temos o chamado modelo de circulação geral atmosférico (MCGA), que tem sido utilizado para estudar variabilidade e mudanças climáticas, e/ou predição sazonais no CPTEC, na qual emprega uma técnica para o tratamento do comportamento caótico da atmosfera denominada de ensemble, que é uma ferramenta necessária para reduzir os efeitos das condições iniciais [1].

A previsão por ensemble surgiu com a finalidade de aumentar os prazos de previsões de tempo e a previsibilidade dos modelos dinâmicos, através da suposição de que os modelos sejam perfeitos e, assim, considerando apenas a incerteza na condição inicial, busca-se, através de alguma técnica específica, estimar os erros associados às observações para criar um conjunto de condições iniciais perturbadas. Este método veio para solucionar o problema da previsibilidade numérica de forma determinística, pois Lorenz (1963, 1965, 1969) observou que a solução de sistemas de equações semelhantes às que governam os movimentos atmosféricos apresentam dependência sensível em relação às condições iniciais fornecidas no início da integração, ou seja, ele notou que partindo de condições ligeiramente perturbadas, após algum tempo de integração, as soluções podem ser completamente diferentes. Isto se deve ao fato de erros inerentes as observações utilizadas no momento de geração da condição inicial, onde uma previsão poderia não ser verificada depois de alguns dias.

3. KDD

A descoberta de conhecimento em banco de dados (KDD – *Knowledge Discovery in Database*) é o processo de encontrar em um banco de dados padrões que estejam escondidos, sem uma idéia pré-determinada ou hipótese sobre o que são estes padrões [2]. Além de prover esta descoberta de informações a KDD também é composta pelas etapas de: definição dos objetivos, limpeza dos dados, transformação dos dados, mineração de dados e interpretação dos resultados (Figura 2).

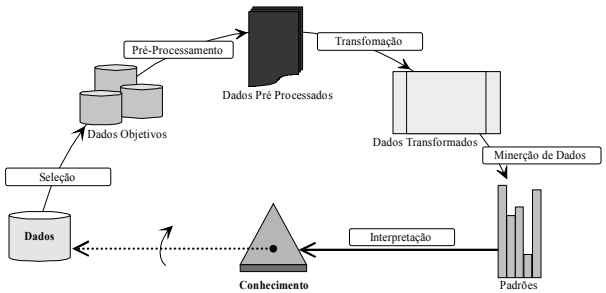

Figura 2: Etapas do processo de KDD

A pesquisa em KDD tem crescido e atraído esforços, baseada na disseminação da tecnologia de bancos de dados e na premissa de que as grandes coleções de dados hoje existentes podem ser fontes de conhecimento útil, que está implicitamente representado e pode ser extraído. No sentido de viabilizar esta tecnologia a KDD se vale, entre outras coisas, de técnicas de aprendizado de máquina, área da inteligência artificial, e de conceitos estatísticos para lidar com a incerteza relacionada às descobertas.

A etapa de maior interesse no processo de descoberta de conhecimento é a mineração de dados, cujo enfoque é a exploração e análise de grandes quantidades de dados para descobrir significativamente modelos e regras. Existem diversas técnicas de mineração de dados, podendo destacar entre tantas: a teoria dos conjuntos aproximativos (*rough sets theory*), teoria dos conjuntos nebulosos (*fuzzy sets theory*), redes neurais artificiais, indução de regras e árvores de decisão.

4. Teoria dos Conjuntos Aproximativos

A TCA é uma extensão da teoria dos conjuntos, cujo enfoque é o tratamento de vagueza e incerteza em dados. Foi inicialmente desenvolvida por *Zdzislaw Pawlak* [9] no início da década de 80, década esta cujo preço e o desempenho dos computadores propiciaram o crescimento e surgimento de novas extensões para a TCA.

4.1. Sistemas de informação

Para o estudo da TCA se faz necessário introduzir alguns conceitos básicos que serão descritos a seguir, começando pela noção de Sistemas de Informação (SI). Um SI nada mais é que um conjunto de dados, sendo representado por uma tabela onde cada linha representa um caso, um evento, um paciente, ou simplesmente um objeto. Toda coluna representa um atributo (uma variável, uma observação, uma propriedade, etc.). Esta tabela, então, é chamada de sistema de informação. Mais formalmente, é um par S = (U; A), onde U é um conjunto finito não-vazio de objetos chamado de universo e A é um conjunto finito não-vazio de atributos tal que a: $U \rightarrow Va$ para todo $a \in A$. O conjunto Va é chamado de conjuntos de valores de Va.

Em muitas aplicações, para fins classificatórios, um certo atributo é distinguido e é denominado atributo de decisão. Os sistemas de informação deste tipo são chamados Sistemas de Decisão (SD). Um SD é sistema qualquer de informação na forma S = (U; $A \cup \{d\}$), onde $d \notin A$ é o atributo de decisão. Os elementos de A são chamados atributos condicionais ou simplesmente condições [7].

Um sistema de decisão pode ser resumido com regras, como por exemplo, algo da forma:

*"**Se** a ='x' e b ='y' **então** d* = Sim*";*
*"**Se** a* = 0.25 ***então*** *d_1 é* 0 *ou d_2 é* 1*";*
*"**Se** a =[155; 159] **então** d é* Não*";*

4.2. Indiscernibilidade

Um SD (no formato de uma tabela, por exemplo) pode ser desnecessariamente grande, em pelo menos dois modos: quando elementos "iguais" são representados muitas vezes ou/e quando alguns atributos são supérfluos.

Com relação aos objetos que são representados muitas vezes, existe uma relação de equivalência que tem a capacidade de "tratar" estes problemas de modo que apenas um objeto represente toda uma classe. Esta relação será formalizada abaixo.

Dado S = (U; A) como sistema de informação, então com qualquer $B \subseteq A$ existe uma relação de equivalência $\mathrm{IND}_A(B)$:

$$IND_A(B) = \{(x, x') \in U \mid \forall a \in B, a(x) = a(x')\} \quad (1)$$

$\mathrm{IND}_A(B)$ é chamada de relação de B-indiscernibilidade. Se $(x, x') \in IND_A(B)$, então objetos x e x' são indiscerníveis relativamente a qualquer atributo de B. A classe de equivalência da relação determinada por x pertencente a X é denotado $[x]_B$ [7].

4.3. Aproximação dos Conjuntos

Claramente, a idéia de relação de equivalência induz a partição do Universo. Estas partições podem ser usadas para construir novos subconjuntos do universo. Subconjuntos que são freqüentemente de interesse têm o mesmo valor do atributo de resultado. Porém, pode acontecer que um conceito não seja definido claramente devido aos elementos serem indiscerníveis e terem valores de decisões contraditórias. Em outras palavras, não é possível induzir uma descrição precisa de tais pacientes em uma tabela. Ou seja, surge aqui a noção de conjunto aproximativo (rough set). Os elementos são divididos então em três classes: os que podem certamente ser classificados pertencentes a uma desejada classe, os elementos que não podem ser classificados e os elementos que não pertencem a classe desejada. Se existem elementos que não podem ser classificados, o conjunto é dito *aproximativo*. Estas noções são formalmente expressadas como segue [7].

Dado S = (U; A), sendo um sistema de informação e $B \subseteq A$ e $X \subseteq U$. Nós podemos aproximar X usando somente as informações contidas em B construindo as aproximações B-inferiores e B-superiores de X, denotados respectivamente $\underline{B}X$ e $\overline{B}X$, onde:

$$\underline{B}X = \{x \mid [x]_B \subseteq X\} \text{ e } \overline{B}X = \{x \mid [x]_B \cap X \neq \varnothing\} \quad (2) \text{ e } (3)$$

Os objetos em $\underline{B}X$ podem ser certamente classificados como membros de X na base de conhecimento B, enquanto os objetos em $\overline{B}X$ podem somente serem classificados como possíveis membros de X na base de conhecimento B. O conjunto $\mathrm{BN_B}(X) = \overline{B}X - \underline{B}X$ é chamada de região de B-fronteira de X, sendo que estes objetos não podem ser classificados pertencentes a X na base de conhecimento B com absoluta certeza. O conjunto $\mathrm{E_B}(X) = U - \overline{B}X$ é então chamado de B-região externa de X, e estes objetos certamente podem ser classificados como não pertencentes a X (na base de conhecimento B).

Portanto um conjunto é dito aproximativo se a região da fronteira não é vazia, e será chamado de crisp (preciso) caso contrário.

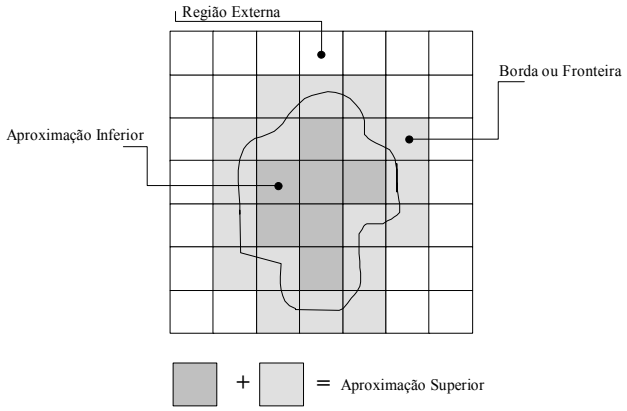

Figura 3. Aproximações dos Conjuntos

4.4. Reduções

Anteriormente um artifício natural de reduzir dados foi visto, que é de identificar as classes de equivalência, por exemplo, objetos que são indiscerníveis usando os atributos disponíveis. Deste modo será feita uma otimização, pois é necessário utilizar somente um elemento da classe de equivalência para representar a classe inteira. Um outro artifício para redução é manter somente os atributos que preservam a relação de indiscernibilidade. Os atributos restantes são redundantes desde que suas remoções mantenham a mesma a classificação. Normalmente existem vários subconjuntos de atributos e esses que são mínimos são usualmente chamados de reduções. A determinação das reduções é

um problema "*NP-hard*" [13]. O número de reduções de um SI com m atributos podem ser iguais a:

$$\binom{m}{\lfloor m/2 \rfloor} \qquad (4)$$

O cálculo das reduções é uma tarefa cujo custo computacional é muito alto.

Dado o sistema de informação S = (*U*, *A*) as definições das noções apresentadas até então são mostradas a seguir:

Uma redução de S é um conjunto mínimo de atributos $B \subseteq A$ tal que $IND_S(B) = IND_S(A)$. Em outras palavras, uma redução (RED(*B*)) é o conjunto mínimo de atributos de *A* que preserva o particionamento do universo realizado pela ralação de indiscernibilidade, e conseqüentemente a habilidade para executar classificações assim como o conjunto de atributos (completo) a faz. O núcleo pode ser dado por $N(B) = \cap RED_i(B)$, i = 1...n.

Sendo S um SI com n objetos, a matriz de discernibilidade de S é uma matriz simétrica n x n com entradas c_{ij}, que é dado na equação abaixo. Cada entrada consiste em um conjunto de atributos que difere os objetos x_i e x_j.

$$c_{ij} = \{a \in A \mid a(x_i) \neq a(x_j)\} \text{ para } i,j = 1, ..., n \quad (5)$$

A função de discernibilidade f_A para um sistema de informação S é uma função de m variáveis Booleanas: $a_1^*,...,a_m^*$ (correspondente aos atributos a1,..., am), cuja definição é dada abaixo e onde $c_{ij}^* = \{a^* \mid a \in c_{ij}\}$.

$$f_A(a_1^*,...a_n^*) = \wedge\{\vee\ c_{ij}^* \mid 1 \leq j \leq i \leq n,\ \ c_{ij} \neq \varnothing\} \quad (6)$$

O conjunto de todos implicantes primos de f_A determina o conjunto de todas as reduções de *A*.

Cada linha da função de discernibilidade corresponde a uma coluna da matriz de discernibilidade. Esta matriz é simétrica e com a diagonal vazia.

Podemos construir uma função Booleana somente considerando a coluna *k* (variável relativa a um objeto especifico) da matriz de discernibilidade, ao invés de todas as colunas, para então obtermos a função de discernibilidade *k*-relativa. O conjunto de todos implicantes desta função determina o conjunto de todas reduções *k*-relativas. Estas reduções revelam a quantidade mínima de informações necessárias para discernir $x_k \in U$ (ou mais precisamente, $[x_k] \subseteq U$) de todos os outros objetos [7].

Uma maneira de aferir as aproximações em um conjunto *B* pode ser através dos seguintes coeficientes,

$$\alpha_B = \frac{|\cup \underline{B}(X_i)|}{|\cup \overline{B}(X_i)|} \qquad (7)$$

$$\beta_B = \frac{|\cup \underline{B}(X_i)|}{|U|} \qquad (8)$$

onde α_B é chamado de *acurácia de aproximação* e β_B é chamado de *qualidade de aproximação*. Obviamente $0 \leq \alpha_B \leq 1$ e $0 \leq \beta_B \leq 1$. Se $\alpha_B = 1$, X_i é dito *preciso* (*crisp)* em relação a *B*, caso contrário, isto é, se $\alpha_B < 1$, então *X* é *aproximado* em relação a *B*.

4.5. Função de Pertinência Aproximativa

Em TCA a noção de função de pertinência é diferente, pois a função de pertinência aproximativa $\mu_X^B : U \rightarrow [0,1]$ quantifica o relativo grau de sobreposição entre o conjunto *X* e a classe de equivalência [x] para cada x. A definição é a seguinte [7], [11], [14]:

$$\mu_X^B(x) = \frac{|[x] \cap X|}{|[x]_B|} \qquad (9)$$

Podemos então definir os conjuntos aproximativos utilizando a função de pertinência aproximativa através das seguintes expressões:

$$\underline{B}(X) = \{x \in U : \mu_X^B(x) = 1\} \quad (10)$$

$$\overline{B}(X) = \{x \in U : \mu_X^B(x) > 0\} \quad (11)$$

$$BN(X) = \{x \in U : 0 < \mu_X^B(x) < 1\} \quad (12)$$

5. ROSETTA

O sistema ROSETTA (*Rough Set Toolkit for Analysis of Data*) é uma ferramenta baseada na TCA, que cobre as diversas etapas do processo de KDD [8], o que possibilita diferentes abordagens e modelagens. Mas uma vantagem de extrema importância do ROSETTA é a possibilidade de inclusão de novos algoritmos, caso haja necessidade, pois este é um sistema modular, ou seja, construído em blocos.

6. Estudo de Caso

O estudo de caso escolhido para análise de dados utilizando a TCA está relacionado com a previsão climática. Neste trabalho procura-se estabelecer relações entre lugares e certas precipitações, períodos de tempo, como por exemplo Dezembro-Janeiro-Fevereiro, e a

precipitação, além de apontar os membros que mais se aproximam, ou em outras palavras, são mais similares ao observado. Quanto aos membros a idéia central é montar uma base de dados, que ao analisar na previsão climática em determinados períodos, forneça a informação de quais os membros são mais aptos para representar dada situação, não necessitando assim de todos os nove membros para tal representação.

6.1. Modelagem dos dados

Os dados colhidos no CPTEC serão dispostos em uma forma tabular para análise. O método utilizado para transformar os dados em tabela é de "*snapshot*", onde cada variável é subdividida em mapas para cada tempo, e então cada registro da tabela é montado a partir do valor da variável no tempo *t* e posição *x*. Isto fica mais evidente visualizando a Figura 4.

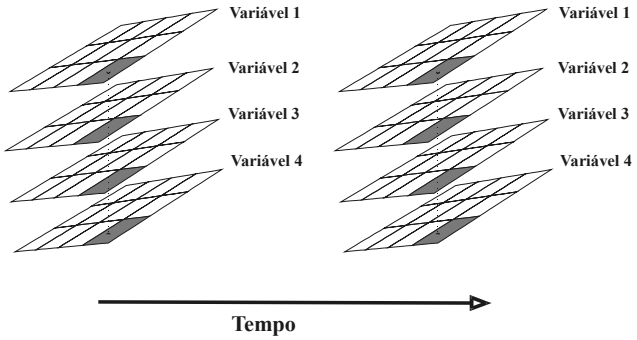

Figura 4: Representação Espaço-Temporal

Na Tabela 1 é mostrado como os dados estão dispostos de modo a serem analisados pelos algoritmos de mineração de dados. Só para ilustrar (Figura 5), no estudo de caso, a variável temporal seria o mês correspondente, no período de 10 anos (*t*=1...120), as variáveis espaciais seriam a latitude e a longitude da região a ser analisada, as variáveis condicionais seriam os membros resultantes do modelo, e por fim, o atributo de decisão seria a variável observada ou real.

Tabela 1. Representação Tabular

VARIÁVEL TEMPORAL	VARIÁVEL ESPACIAL	VAR 1	VAR 2	...	VAR D (DECISÃO)
1	l_1	a_1	b_1	...	d_1
2	l_2	a_2	b_2	...	d_2
3	l_3	a_3	b_3	...	d_3
4	l_4	a_4	b_4	...	d_4
5	l_5	a_5	b_5	...	d_5
6	l_6	a_6	b_6	...	d_6
. . .	. . .	. . .	. . .	. . .	. . .
n	l_n	a_n	b_n	...	d_n

	lon	lat	tempo	prec1	prec2	prec3	prec4	prec5	prec6	prec7	prec8	prec9	cmap
1	-43.12	-7.49	1	1	1	1	1	1	1	1	1	1	[6, 12)
2	-43.12	-7.49	2	663	739	758	713	749	653	751	502	718	[*, 6)
3	-43.12	-7.49	3	678	668	436	649	31	743	610	628	636	[*, 6)

Figura 5. Dados de Precipitação

6.2. Processo de KDD

O modo idealizado para gerar tais conhecimentos, é o processo de descoberta de conhecimento, que já fora abordado anteriormente. Isto porque, com o uso de tal método, consegue-se cobrir todas as necessidades que surgem neste tipo de análise, tais como, limpeza dos dados, discretização e a etapa mais importante de todos, o núcleo da descoberta de conhecimento a chamada mineração de dados.

A primeira etapa do processo de KDD consiste em selecionar os dados para análise. Os dados do problema selecionados são:

Intervalo de Tempo
JAN82 – DEZ91 (120 tempos)

Coordenadas:
LAT 10N 35S
LON 80W 30W
Obs.: Referente à América do Sul;

Simulação (9 membros)
PREC Precipitação Total (Kg m**-2 Day**-1)

CMAP
PREC Precipitação (Kg m**-2 Day**-1)

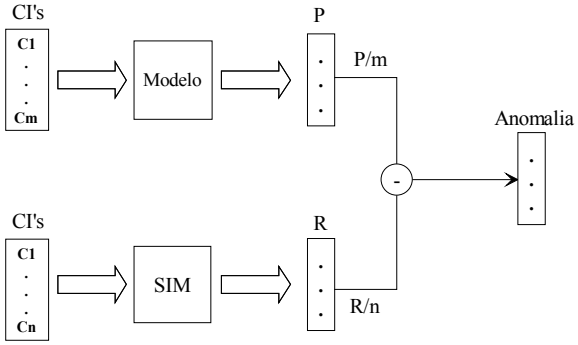

Figura 6. Previsão Climática

Os dados colhidos até então no CPTEC são referentes a uma simulação realizada para um período de dez anos, de janeiro de 1982 a dezembro de 1991, com o objetivo de construir a climatologia do modelo, baseando-se nas observações do mundo real.

Essa simulação é o resultado da integração de nove condições iniciais (CI), 11 a 19 de novembro de 1981, para dez anos no modo ensemble. Os resultados do

modelo, chamados de membros, são confrontados com dados observados.

Na previsão climática atualmente estes dados são utilizados para uma comparação com os resultados do modelo, na qual resulta a chamada **anomalia** (Figura 6).

Existem também dados colhidos que são de outra espécie, tal como dados do CMAP que são dados observados ou reais.

Os dados a serem analisados a priori são os relativos à precipitação ($prec_1$ a $prec_9$ são membros do modelo, *cmap* é a precipitação observada).

Na etapa de pré-processamento o objetivo é eliminar os ruídos ou ausência de valores existentes na base de dados. No caso da Precipitação e CMAP existem valores indefinidos que podem ser substituídos pela média dos valores para o dado atributo.

Antes da etapa mais importante do KDD, há de se fazer uma transformação nos dados, visando transformar atributos não-categóricos em atributos categóricos. Isto pode se dar através de uma discretização. Este passo é muito importante pois uma das premissas da TCA é reduzir a precisão dos dados revelando suas regularidades [10].

Na mineração de dados será utilizada a ferramenta matemática da TCA, visando além de induzir conhecimento, reduzir a quantidade de informações analisadas para geração das regras. Nesta etapa, basicamente usam-se algoritmos de reduções e geração de regras.

Na interpretação dos dados ainda pode-se gerar um classificador, que busca através do conjunto de regras gerado na mineração de dados, uma série de métricas para validação das mesmas.

7. Conclusão

A TCA é uma ferramenta matemática de grande poder no tratamento de informações. Muitas são as vantagens da utilização da TCA, como por exemplo, a sua simplicidade, versatilidade, a redução do número de variáveis redundantes no processo e redução do volume de dados, o que acarreta uma compactação no banco de dados; a boa fundamentação matemática e a possibilidade de modelagem dos dados por meio de regras, permitindo a construção de softwares e por fim, a não utilização de informações adicionais para análise de dados.

Os maiores problemas da TCA, são

- O cálculo das reduções, que podem ser caros computacionalmente, e quando realizados podem gerar muitas regras;
- Dificuldade de trabalhar com dados contínuos.

Quando o número de regras for elevado devido ao tipo de reduções efetuadas, como por exemplo, na redução *k*-relativa é calculada a redução de cada elemento em relação ao todo, resultando em um grande número de reduções e conseqüentemente muitas regras. Pode-se executar alguns tipos de filtros para reduzir este número de regras.

Todas as regras calculadas pelo sistema, podem ainda serem exportadas para uma linguagem como Prolog, ou C++, para a construção de um simples classificador ou de algo maior, como um sistema especialista, por exemplo.

Referências

[1] Cavalcanti, I. F. A. et al. *Global Climatological Features in a Simulation Using the CPTEC-COLA AGCM*. Jounal of Climate, vol. 15, n 27, p. 2965-2988. 2002

[2] Chen, Z. *Data Mining and Uncertain Reasoning: an Integrated Approach***.** New York, John Wiley & Sons, 2001.

[3] Holsheimer M. & Siebes A. *Data mining: the search for knowledge in databases*. Report CSR9406. Amsterdam, the Netherlands: CWI, Jan. 1994.

[4] Lorenz, E. N. *Deterministic non-periodic flow*. J. Atmos. Sci., v. 20, p. 130-141, 1963.

[5] Lorenz, E. N. *A study of the predictability of a 28-variable atmospheric model*. Tellus, v. 17, p. 321-333, 1965.

[6] Lorenz, E. N. *The predictability of a flow which possesses many scales of motion*. Tellus, v. 21, p. 289-307,1969.

[7] Komorowski, J. *et al*. *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.

[8] Øhrn, A. *Discernibility and Rough Sets in Medicine: Tools and Applications.* PhD thesis, Norwegian University of Science and Technology, Department of Computer and Information Science, NTNU. 1999.

[9] Pawlak Z. *Rough sets*. International Journal of Computer and Information Sciences, 11:341--356, 1982.

[10] Pawlak, Z. *Rough sets – Theorical aspects of reasoning about data*, Kluver Academic Publishers, Dordrecht, 1991.

[11] Pawlak, Z.; SKOWRON, A. *Rough membership functions*. In: R. Yager, M. Fedrizzi J. Kacprzyk (Eds.), Advances in the Dempster-Shafer Theory of Evidence, Wiley, NewYork, p. 251-271, 1994.

[12] Røed G. *Knowledge Extraction from Process Data: A Rough Set Approach to Data Mining on Time Series*, "citeseer.nj.nec.com/119626.html", 1999

[13] Skowron, A. & Rauszer C. *The Discernibility Matrices and Functions in Information Systems*. In: Slowinski, p. 331 – 362, 1992.

[14] Wong, S. K. M. & Ziarko, W. *Comparison of the probabilistic approximate classification and the fuzzy set model*. Fuzzy Sets and Systems 21, p. 357-362, 1986.