USO DE TEORIA DE CONJUNTOS APROXIMATIVOS E REDES NEURAIS ARTIFICIAIS NO ESTUDO DE PADRÕES CLIMÁTICOS SAZONAIS

Juliana A. Anochi[1], José Demisio S. D. Silva[1]

[1]Instituto Nacional de Pesquisas Espaciais – Laboratório Associado de Computação e Matemática Aplicada

Av. dos Astronautas, 1758, São José dos Campos-SP, CEP:12227-010, Brasil

{juliana.anochi,demisio}@lac.inpe.br

Abstract. *This work describes an Artificial Intelligence based technique to prepare data for constructing a climate forecasting empirical model from reanalysis data in the South region of Brazil using Artificial Neural Network (ANN). The method uses Rough Sets Theory (TRS) to reduce the amount of variables in the forecasting task. The reduced data is used to train artificial neural networks for the specific task of climate forecasting. The input of ANN there is two kinds of data: the variables chosen by the TRS and full variables data to learn the seasonal behavior of the variable precipitation on the South region of Brazil.*

Resumo. *Este trabalho descreve o uso de técnicas baseada em Inteligência Artificial para o tratamento de dados de reanálise com a obtenção de construir um modelo empírico para a previsão climática sobre a região Sul do Brasil utilizando Redes Neurais Artificiais (RNAs). A técnica da Teoria dos Conjuntos Aproximativos (TCA) foi utilizada para reduzir o número de variáveis no processo de previsão. Como entrada as RNAs recebem dados de duas espécies: as variáveis reduzidas pela TCA e os dados completos para aprender o comportamento sazonal da variável de precipitação sobre a região Sul do Brasil.*

1. Introdução

Este trabalho apresenta um método de redução de atributos, baseado na Teoria dos Conjuntos Aproximativos, para a concepção de modelos de previsão climática sazonal sobre a região Sul do Brasil, utilizando Redes Neurais Artificiais, a partir de dados de reanálise.

O problema de previsão climática é dependente de informações de modelos que físico-matemáticos que descrevem o comportamento da atmosfera. Estes modelos são executados a partir de condições da atmosfera observadas, utilizando máquinas com grande poder de processamento. Entretanto a demanda por poder de processamento cresce com o aumento da disponibilidade de novos sensores para observação das condições atmosféricas uma vez que grandes volumes de dados são gerados. Por outro lado, a complexidade da modelagem da atmosfera e o fato de muitos processos

atmosféricos modelados serem de natureza caótica, implicam que a execução dos modelos demanda tempo podendo acarretar atrasos na emissão de boletins.

Buscando criar alternativas para concepção de modelos que permitam a aquisição de resultados em tempo hábil para o uso em processos de tomada de decisão, este trabalho busca, através da mineração de dados, analisar e compreender o comportamento atmosférico, com intuito de identificar informações relevantes que possam ser usadas em processos de previsão climática.

Como técnica de mineração de dados, o trabalho propõe o uso da Teoria dos Conjuntos Aproximativos (TCA) que trata informações incertas e imprecisas, por meio de aproximações de um conjunto de dados. Para a concepção dos modelos de previsão, o trabalho propõe o uso de Redes Neurais Artificiais (RNAs) devido à possibilidade de se conceber os modelos a partir do aprendizado do comportamento dos dados meteorológicos e para estimar a precipitação sobre a região Sul do Brasil.

2. Previsão Climática

Previsão climática consiste na estimativa do comportamento médio da atmosfera com alguns meses de antecedência. No processo de previsão climática em escala de tempo sazonal, pode-se prever, se o próximo inverno será mais frio que a média, ou ainda, se haverá mais ou menos chuva em relação a estação anterior. Cabe ainda à previsão climática, a partir da análise da friagem no inverno e das ondas de calor, prever as propriedades estatísticas do estado climático [Vianello 2000].

O Centro de Previsão de Tempo e Estudos Climáticos do Instituto Nacional de Pesquisas Espaciais (CPTEC-INPE) é um dos órgãos responsáveis pelos estudos climáticos no Brasil. Para isso dispõe de um supercomputador capaz de realizar estudos das condições do tempo e clima para a elaboração de cenários futuros de mudanças do clima com alta resolução. Estas atividades apóiam estudos de impactos e vulnerabilidade e permitem que se faça projeções dos extremos climáticos do estado atmosférico.

3. Descoberta de Conhecimento em Banco de Dados

A descoberta de conhecimento em banco de dados (Knowledge Discovery in Database KDD) é o processo que busca extrair padrões válidos e potencialmente úteis que estejam presentes nos dados, mas que não estão visíveis diretamente. O processo de KDD envolve uma sequência de tarefas que podem ser executadas diversas vezes, com possibilidade de modificação dos algoritmos e configurações até que se obtenha o melhor resultado [Fayyad et al 1996].

Para realizar uma tarefa de KDD é preciso: conhecer o domínio da aplicação; selecionar os dados; realizar o pré-processamento dos dados; limpar os dados; transformar os dados; escolher o algoritmo de mineração de dados; a interpretar os resultados obtidos; e utilizar o conhecimento descoberto [Fayyad et al 1996].

Entretanto, dentre as etapas incluídas no processo de KDD, a mineração de dados é o passo principal de todo o processo, pois consiste na aplicação de técnicas inteligentes com o propósito de extrair padrões interessantes em base de dados.

As técnicas empregadas em mineração de dados baseiam-se em diferentes princípios teóricos. Em termos de Inteligência Artificial, as técnicas mais citadas em trabalhos são: Teoria dos Conjuntos Aproximativos, Teoria dos Conjuntos Nebulosos, Redes Neurais Artificiais, Algoritmos Genéticos, Indução de Regras e Árvores de Decisão. Para o desenvolvimento deste trabalho, a mineração de dados é realizada utilizando a Teoria dos Conjuntos Aproximativos. As Redes Neurais Artificiais são utilizadas para a obtenção dos modelos de previsão de clima.

4. Teoria dos Conjuntos Aproximativos

A Teoria dos Conjuntos Aproximativos (TCA) foi introduzida pelo matemático polonês Zdzislaw Pawlak em 1982, como um formalismo matemático para tratar informações incertas e imprecisas, por meio de aproximações de um conjunto de dados [Pawlak 1982]. A TCA mede a similaridade entre objetos através da relação de indiscernibilidade. Os objetos são indiscerníveis se possuírem os mesmos valores para todos os atributos que os caracterizam.

4.1. Sistema de Informação

A representação dos dados na TCA é feita através de um Sistema de Informação (SI) em que os dados são representados no formato de tabela, onde cada linha representa um objeto e as colunas representam os atributos [Komorowski *et al* 1999].

Formalmente um *SI* é um par ordenado *SI* = (*U, A*) onde *U* é um conjunto finito e não vazio de elementos, chamado de universo, e *A* é um conjunto finito de elementos não vazio chamados atributos, tal que a: $U \rightarrow Va$ para todo $a \in A$. O conjunto *Va* é chamado de domínio de *a*, e os objetos a_n representam as ocorrências observadas.

Um *SI* é um Sistema de Decisão (SD) quando inclui um atributo de decisão *d*, que não pertence ao conjunto de atributos *A*. Formalmente, $SD = (U, A \cup \{d\})$, em que $d \notin A$ [Komorowski *et al* 1999]. A Tabela 4.1 apresenta um exemplo de um *SD*, em que existem cinco objetos {a_1, a_2, a_3, a_4, a_5}, três atributos condicionais {*Estação, Temperatura do Ar, Vento*} e um atributo de decisão {*Geada*}.

Tabela 4.1. Sistema de Decisão

U	**Atributos Condicionais**			**Decisão**
	Estação	Temperatura do ar	Vento	Geada
a_1	Verão	Alta	Fraco	Não
a_2	Outono	Média	Médio	Não
a_3	Inverno	Baixa	Forte	Sim
a_4	Primavera	Alta	Fraco	Não
a_5	Inverno	Baixa	Forte	Sim

4.2. Relação de Indiscernibilidade

A relação de indiscernibilidade é a similaridade entre dois ou mais objetos. Uma relação de equivalência tem a capacidade de tratar estes problemas de modo que apenas um objeto represente toda uma classe. Dado um *SI* = (*U, A*) como sistema de informação, então com para qualquer $B \subset A$ existe uma relação de equivalência $IND_A(B)$ (ou indiscernibilidade) definida pela expressão 4.1 [Komorowski *et al* 1999].

$$IND_A(B) = \{(x, x') \in U \mid \forall a \in B, a(x) = a(x')\} \qquad (4.1)$$

Na Equação 4.1, os dois objetos *x* e *x'* do conjunto *U*, são indiscerníveis para um subconjunto de atributos $B \subset A$, se para cada atributo *a* de *x* e *x'* em *B*, os valores forem iguais. A partir do *SD* apresentado na Tabela 4.1, os possíveis subconjuntos em relação aos atributos condicionais são mostrados na Tabela 4.2.

Tabela 4.2. Possíveis subconjuntos

IND_A ({Estação})	$\{\{a_1\}, \{a_2\}, \{a_3, a_5\}, \{a_4\}\}$
IND_A ({Temperatura do ar})	$\{\{a_1, a_4\}, \{a_2\}, \{a_3, a_5\}\}$
IND_A ({Vento})	$\{\{a_1, a_4\}, \{a_2\}, \{a_3, a_5\}\}$
IND_A ({Estação, Temperatura do ar})	$\{\{a_1\}, \{a_2\}, \{a_3, a_5\}, \{a_4\}\}$
IND_A ({Estação, Vento})	$\{\{a_1\}, \{a_2\}, \{a_3, a_5\}, \{a_4\}\}$
IND_A ({Temperatura do ar, Vento})	$\{\{a_1, a_4\}, \{a_2\}, \{a_3, a_5\}\}$
IND_A ({Estação, Temperatura do ar, Vento})	$\{\{a_1\}, \{a_2\}, \{a_3, a_5\}, \{a_4\}\}$

4.3. Redução

O processo de redução gera os chamados redutos (RED), que são subconjuntos de atributos com capacidade de representar o conhecimento da base de dados, ou seja, agrupar em classes os objetos que são indiscerníveis entre si [Komorowski *et al* 1999].

Os subconjuntos são agrupados e esses grupos são objetos que não podem ser discerníveis entre si. De acordo com a TCA esses grupos são considerados como classes. Por exemplo, na Tabela 4.2, observa-se que para o subconjunto $IND_A(B)$ = {*Estação, Temperatura do ar, Vento*}, os objetos a_3 e a_5 são indiscerníveis, portanto existe uma relação de indiscernibilidade entre eles, dessa forma é possível reduzi-los, formando a classe C_3, como mostrado na Tabela 4.3.

Tabela 4.3. Redução para o subconjunto {Estação, Temperatura do ar, Vento}

U	**Atributos Condicionais**		
	Estação	Temperatura do ar	Vento
C_1	Verão	Alta	Fraco
C_2	Outono	Média	Médio
C_3	Inverno	Baixa	Forte
C_4	Primavera	Alta	Fraco

Somente os atributos que preservam a relação de indiscernibilidade são mantidos na redução, levando em consideração as classes formadas pelo atributo de decisão. Os atributos restantes são redundantes, ou supérfluos, desde que suas remoções mantenham a mesma classificação [Komorowski *et al* 1999].

5. Redes Neurais Artificiais

Redes Neurais Artificiais são técnicas computacionais que apresentam um modelo matemático inspirado no cérebro. Elas adquirem conhecimento através de experiência [Haykin 2001]. As redes neurais apresentam como principais vantagens às características de adaptabilidade, generalização e tolerância a ruídos. Essas características parecem ser importantes na aplicação de redes neurais em problema de previsão climática, devido à complexidade de tal problema [Haykin 2001].

O modelo de rede utilizado para desempenhar a previsão de precipitação foi o Perceptron de Múltiplas Camadas (MLP), que utiliza o algoritmo de retropropagação do

erro. Este algoritmo é composto de dois passos: um passo para frente, a propagação e um passo para trás, a retro-propagação. Em um primeiro momento o sinal na rede neural se propaga da entrada para a saída. Na seqüência do treinamento o erro é calculado, pela comparação do resultado na saída e o desejado, e então este erro é propagado da saída até a camada de entrada, modificando os pesos de todas as camadas de acordo com o erro obtido. A métrica para quantificar o desempenho da previsão foi o erro quadrático médio E dado por:

$$E = \frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2 \tag{5.1}$$

em que N é o número de pontos da grade, y_k é o valor real no ponto de grade e $\hat{y}_k$ é a estimação produzida pela rede neural.

6. Metodologia e Resultados

A área de estudo para realização dos experimentos é a região Sul do Brasil como mostrado na Figura 6.1. As coordenadas geográficas estão compreendidas entre as longitudes [60°W, 45°W] e entre as latitudes [-35°S, -22.5°S], compreendendo 42 pontos de grade.

Figura 6.1 Região selecionada para análise

Os dados utilizados nos experimentos foram coletados da base de dados de reanálise do NCEP/NCAR (National Centers for Environmental Prediction / The National Center for Atmospheric Research) [http://ww.ncep.noaa.gov]. O período de tempo compreende uma janela de 21 anos entre janeiro de 1980 e dezembro de 2000, em uma área contida entre as latitudes [10°N, 35°S] e longitudes [80°W, 30°W], referente à América do Sul. A resolução espacial, em ambas as dimensões da grade, de 2.5° e resolução temporal (t) de 1 mês. As variáveis contidas na base de dados são: temperatura do ar (airt), divergência (div), precipitação (prec), umidade específica (shum), pressão da superfície (spres), componentes vento zonal hPa (u300), (u500) e (u850) e meridional hPa (v300), (v500) e (v850).

Do conjunto total de dados, foram selecionados 18 anos (janeiro de 1980 a dezembro de 1997) para o treinamento dos modelos de redes neurais e como entrada para o processamento pela TCA. Os demais 3 anos (janeiro de 1998 a dezembro de 2000) foram usados para a validação dos modelos.

A metodologia adotada neste trabalho consiste em duas diferentes abordagens para a realização do processo de previsão climática: na primeira uma rede MLP é treinada com todos os dados disponíveis, selecionados para treinamento; na segunda os dados são submetidos à TCA para reduzir o volume de dados, segundo a relação de

indiscernibilidade, formando os redutos para o treinamento da rede neural. Assim para efeito de comparação, as redes neurais treinadas com os dados completos são comparadas com a rede treinada com os dados reduzidos.

A topologia da rede MLP foi configurada durante testes preliminares, variando-se de maneira *ad hoc*, o número de neurônios nas camadas escondidas e o número máximo de épocas de treinamento. Os testes conduzidos levaram a uma arquitetura com uma camada escondida com 14 neurônios, submetida ao máximo de 10000 épocas, cada neurônio foi configurado com a função de ativação do tipo logística sigmoidal.

6.1. Resultados

Para a realização do processo de redução dos atributos, utilizou-se a ferramenta ROSETTA [Øhrn 1999] para calcular os redutos mínimos. Primeiramente os dados são discretizados, em seguida os dados são submetidos ao algoritmo de redução. A redução dos atributos mais relevantes, como aqueles com ocorrência igual ou superior a 70% (valor escolhido de forma *ad hoc*) de presença na função de indiscernimento são apresentados na Tabela 6.1.

Tabela 6.1. Variáveis extraídas pela TCA

Variáveis	%
u500	76%
u300	88%
v850	86%
v300	72%
spres	70%
div	71%

Na Tabela 6.2 é apresentado o desempenho em termos de erro quadrático médio para os dois modelos de redes neurais, para as quatro estações dos anos de 1998 a 2000.

Tabela 6.2. Erro Quadrático Médio

Erro quadrático médio			
Ano	**Estação**	**RNA**	**TCA**
1998	Outono	0,0417	0,0138
1998	Inverno	0,0052	0,0701
1998	Primavera	0,0036	0,0489
1998	Verão	0,3930	0,6970
1999	Outono	0,0004	0,0554
1999	Inverno	0,0567	0,0199
1999	Primavera	0,0360	0,2767
1999	Verão	0,1751	0,1347
2000	Outono	0,0056	0,0079
2000	Inverno	0,0451	0,0001
2000	Primavera	0,0658	0,0454

Os resultados são mostrados em um mapa gerado pela ferramenta GrADS [Doty 2009], para as quatro estações do ano de 1999. As Figuras 6.2(a), 6.2(b) e 6.3(c) representam respectivamente as situações observadas, a estimativa obtida pela rede com os dados processados pela TCA e a estimativa pela rede neural com todos os dados disponíveis, para a estação verão de 1999, observa-se que a estimativa usando a TCA apresentou um melhor resultado em relação a rede com os dados completos.

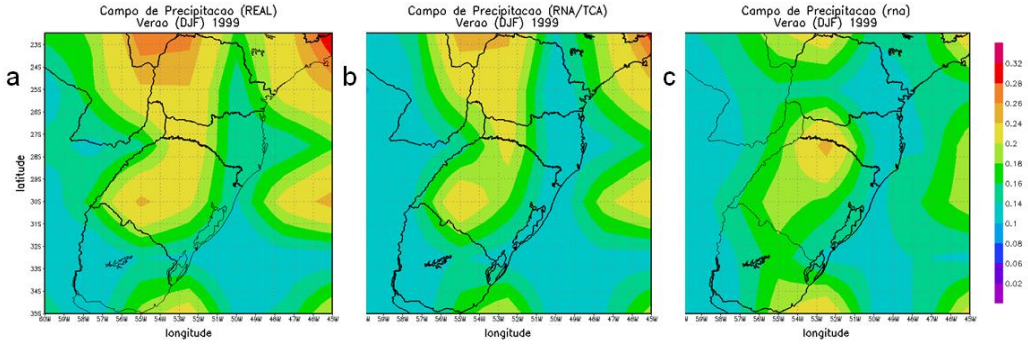

Figura 6.2 Estimativa de precipitação. Verão de 1999. (a) Precipitação Real; (b) Estimativa com dados reduzidos por TCA, (c) Estimativa com RNA com todos os dados.

Na Figura 6.3 são mostrados os resultados de precipitação obtidos para a estação outono de 1999. A estimativa realizada com os dados reduzidos pela TCA apresenta padrões visuais mais semelhantes àqueles mostrado na Figura 6.3(a).

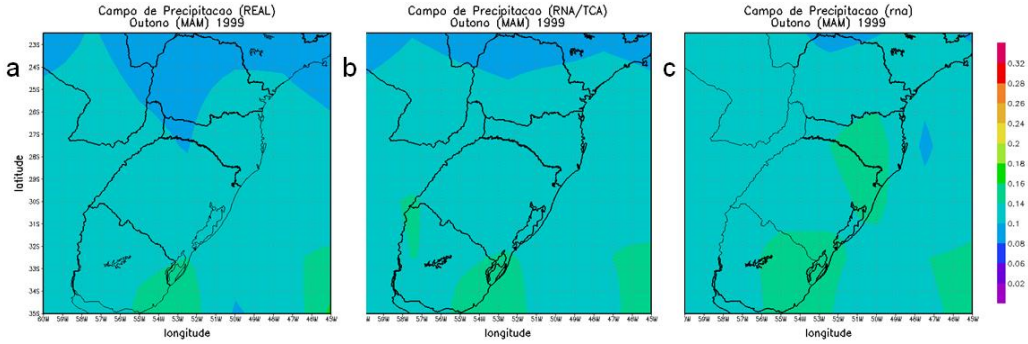

Figura 6.3 Estimativa de precipitação. Outono de 1999. (a) Precipitação Real; (b) Estimativa com dados reduzidos por TCA, (c) Estimativa com RNA com todos os dados.

Na Figura 6.4 são mostradas as estimativas de precipitação obtidas pelos modelos de redes neurais para a estação inverno de 1999. O resultado obtido de previsão utilizando dados processados pela TCA apresenta um melhor resultado em relação a previsão usando todos os dados.

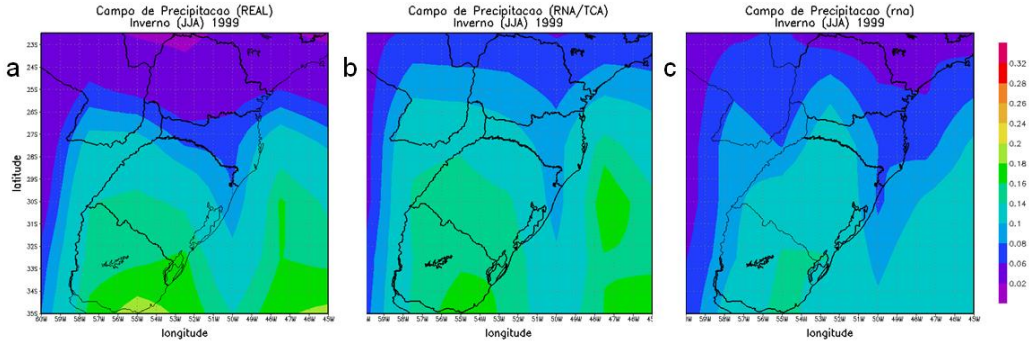

Figura 6.4 Estimativa de precipitação. Inverno de 1999. (a) Precipitação Real; (b) Estimativa com dados reduzidos por TCA, (c) Estimativa com RNA com todos os dados.

Na Figura 6.5 são apresentados os resultados de previsão de precipitação para a estação primavera de 1999. Neste caso a previsão realizada com todos os dados tem padrões mas semelhantes em relação ao observado.

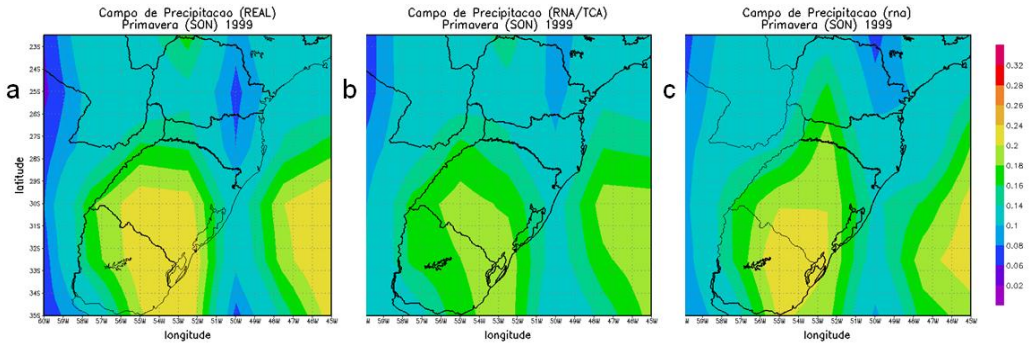

Figura 6.5 Estimativa de precipitação. Primavera de 1999. (a) Precipitação Real; (b) Estimativa com dados reduzidos por TCA, (c) Estimativa com RNA com todos os dados.

7. Considerações Finais

Neste trabalho foi apresentado o uso de técnicas de Inteligência Artificial para estimar o comportamento médio atmosférico sobre a região Sul do Brasil. Os dados disponíveis foram utilizados para treinar modelos de redes neurais artificiais para realizar estimativa de precipitação. Analisando qualitativamente os resultados de maneira gráfica, observou-se que as estimativas produzidas com o modelo de rede neural usando os dados pré-processados com a TCA, produziram estimativas muito semelhantes ao observado.

Com base nos experimentos realizados, a metodologia proposta neste trabalho mostra-se como uma boa alternativa para a obtenção de modelos localizados de previsão de clima, consistindo em informação de previsão com base nos dados históricos.

Referências Bibliográficas

Doty, B. (2009). Grid Analysis and Display System (GrADS). Disponível em: http://grads.iges.org/grads/head.html. Acesso em: 23-fev2009.

Fayyad U. and Shapiro G.P. and Smyth P. (1996). From Data Mining to Knowledge Discovery in databases. *AAAI Press*.

Haykin S. (2001). *Redes Neurais: Princípios e Práticas*. Bookman, Porto Alegre, 1 edition.

Komorowski, J. and Øhrn, A. (1999). Modelling prognostic power of cardiac tests using rough sets. *Artificial Intelligence in Medicine*. pp. 167-1991.

Øhrn A. (1999). Discernibility and Rough Sets in Medicine: Tools and Applications. Tese de 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.1. pp. 341-356.

Vianello, R. L. and Alvez A. R. (2000). “Meteorologia Básica e Aplicações”, Viçosa, UFV.