Mineração de Dados Meteorológicos pela Teoria dos Conjuntos Aproximativos para Aplicação na Previsão de Precipitação Sazonal

Juliana Aparecida Anochi
Instituto Nacional de Pesquisas Espaciais, INPE
12227-010, São José dos Campos, SP
E-mail: juliana.anochi@lac.inpe.br

José Demisio Simões da Silva
Instituto Nacional de Pesquisas Espaciais, INPE
12227-010, São José dos Campos, SP
E-mail: demisio@lac.inpe.br

***Resumo**: Este artigo visa mostrar um método de redução de atributos, baseado em técnicas de inteligência artificial para a realização de previsão de precipitação sazonal sobre dados de reanálise. A metodologia usa a Teoria dos Conjuntos Aproximativos para extrair informações relevantes dos dados, visando reduzir os esforços computacionais na realização dos estudos de previsão climática.*

1. Introdução

Com o desenvolvimento da tecnologia computacional e a possibilidade de armazenamento de informações em grandes bases de dados, a análise e extração de conhecimento necessitam de novas abordagens para se obter resultados em tempo hábil para uso deste conhecimento em processos de tomada de decisão, principalmente naqueles que envolvem situações críticas para o ser humano, como nos processos de análise de clima e de tempo em Meteorologia, por exemplo.

Em particular, na Meteorologia, a grande disponibilidade de dados, oriundos de diferentes tipos de sensores, implica no aumento de dificuldade no processo de previsão meteorológica. Apesar da disponibilidade de vários instrumentos, que medem diferentes grandezas associadas aos fenômenos meteorológicos, as decisões sobre previsões podem ser tomadas com base em algumas das variáveis medidas. Entretanto, os modelos computacionais para previsão meteorológica que tentam simular a física dos processos atmosféricos, necessitam de todas as informações disponíveis para reproduzir, o mais fiel possível, o comportamento atmosférico.

Neste trabalho, uma técnica de mineração de dados é usada em um estudo de análise de variáveis meteorológicas na previsão climática. O objetivo é obter subsídios para analisar e compreender o comportamento dos dados meteorológicos e identificar informações relevantes que possam ser usadas em processos de previsão.

Como técnica de mineração de dados, este trabalho propõe o uso da Teoria dos Conjuntos Aproximativos (TCA) introduzida por Zdzislaw Pawlak em 1982, cuja característica intrínseca é de extrair o volume de dados, tratando informações incertas e imprecisas, por meio de aproximações de um conjunto de dados. Na TCA existem dois conceitos que estão diretamente ligados à compactação da base de dados: a relação de indiscernibilidade, na qual um elemento representa toda a classe, e as reduções de atributos, que são constituídas dos atributos mais relevantes e indispensáveis.

Os dados minerados com a TCA são em seguida utilizados para treinar redes neurais artificiais para uma tarefa de previsão climática. Os resultados são comparados com previsões feitas por redes neurais artificiais treinadas com todos os dados disponíveis.

Neste trabalho são utilizados dados meteorológicos da região Nordeste do Brasil. As redes neurais utilizadas foram treinadas para compor modelos de previsão de precipitação sobre a região, considerando todo o conjunto de dados e os dados reduzidos pelo uso da TCA.

Os resultados mostram que o uso de técnicas de mineração de dados pode diminuir a complexidade de sistemas de previsão meteorológicos, com a redução de informações úteis para os processos de previsão, mantendo boa precisão nas previsões.

2. Previsão Climática

Previsão climática é definida como a estimativa do comportamento médio da atmosfera com alguns meses de antecedência. Por exemplo, em escala de tempo sazonal, pode-se prever se o próximo inverno será mais frio que a média, ou ainda, se haverá mais chuva que a estação anterior. Cabe ainda à previsão climática analisar a friagem no inverno e as ondas de calor, visando prever as propriedades estatísticas do estado climático [8].

Para as previsões climáticas existem os modelos numéricos. Dentre esses modelos existe o chamado Modelo de Circulação Geral Atmosférico (MCGA), que tem sido utilizado para a realização de previsão climática, de forma experimental, para estudar a variabilidade e as mudanças climáticas. Outro modelo numérico é o modelo regional, que é uma solução para aumentar a resolução do modelo sem aumentar o custo computacional. Este modelo prevê fenômenos de pequena escala como tempestades, brisa marítimas entre outras [1].

3.Teoria dos Conjuntos Aproximativos

A Teoria dos Conjuntos Aproximativos (TCA) foi proposta no início da década de 80 pelo matemático polonês, Zdzislaw Pawlak em 1982, esta teoria baseia-se nas relações de similaridades entre objetos através da relação de indiscernibilidade.

A forma de representar os dados na abordagem de TCA é através de um Sistema de Informação (SI), organizados em formato de tabela, em que cada linha representa um objeto e as colunas representam os atributos [5]. Um SI é definido como um par ordenado SI = (U, A) em que U é um conjunto finito de elementos não vazio, chamado de universo, e A é um conjunto finito não vazio de elementos chamados atributos. Um Sistema de Decisão (SD) é qualquer $SI = (U, A \cup \{d\})$, onde $d \notin A$ é o atributo de decisão.

O processo de redução dos dados é feito através dos chamados redutos, que são subconjuntos de atributos com capacidade de representar o conhecimento da base de dados [7]. Na Tabela 3.1, observa-se um SI, composto por: um conjunto de objetos U = {a_1, a_2, a_3, a_4, a_5} e os atributos condicionais A = {Estação do Ano, Temperatura, Vento}.

U	**Atributos Condicionais**		
	Estação do Ano	Temperatura	Vento
a_1	Outono	média	forte
a_2	Inverno	baixa	moderado
a_3	Primavera	alta	moderado
a_4	Verão	alta	forte
a_5	Outono	média	forte

Tabela 3.1: Sistema de Informação

Dado um SI = (U; A), então com qualquer B $\subseteq$ A existe uma relação de equivalência IND_A (B):

$$IND_A(B) = \{(x, x') \in U \mid \forall a \in B, a(x) = a(x')\} \quad (3.1)$$

A relação de indiscernibilidade é a similaridade entre dois ou mais objetos caracterizados pelos mesmos valores. Para o subconjunto IND(B) = {Estação do Ano, Temperatura, Vento}, os objetos a_1 e a_5 são indiscerníveis, dessa forma, é possível reduzi-los, formando assim a classe C_1. Na Tabela 3.2 são apresentadas as classes para o subconjunto {Estação do Ano, Temperatura, Vento}.

U	**Atributos Condicionais**		
	Estação do Ano	Temperatura	Vento
C_1	Outono	média	forte
C_2	Inverno	baixa	moderado
C_3	Primavera	alta	moderado
C_4	Verão	alta	forte

Tabela 3.2: Classe para IND(B) = {Estação do Ano, Temperatura, Vento}

Neste trabalho utilizou-se a ferramenta ROSETTA (Rough Set Toolkit for Analysis of Data) que é um software utilizado para análise de dados, baseado na teoria dos conjuntos aproximativos, para realização dos processos de [6].

4. Redes Neurais Artificiais

Redes Neurais Artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através de experiência, seu comportamento inteligente surge das interações entre as unidades de processamento da rede [4].

As redes neurais apresentam como principais vantagens às características de adaptabilidade, generalização e tolerância a ruídos [4]. Essas características parecem ser de importantes na aplicação de redes neurais em problema de previsão climática, devido à complexidade de tal problema.

O tipo de rede utilizado para desempenhar a previsão de precipitação foi o Perceptron de Múltiplas Camadas (MLP), utilizando 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.

5. Implementação e resultados

A metodologia adotada neste trabalho considera duas abordagens para previsão climática: na primeira uma rede neural é treinada com todos os dados disponíveis, selecionados para treinamento; na segunda abordagem, os dados disponíveis para treinamento são minerados para reduzir o volume de dados usado na previsão. Na fase de mineração, os dados são submetidos à TCA que busca identificar os redutos mais significativos para realizar previsão de clima. Os dados obtidos nos redutos são utilizados para treinar uma rede neural artificial do tipo Perceptron de Múltiplas Camadas (MLP) com o algoritmo de treinamento por retro-propagação do erro. Neste processo de descoberta de conhecimento, a TCA identifica os atributos relevantes para o processo de previsão climática, gerando como resultado os atributos com maior ocorrência, segundo a relação de indiscernibilidade, que são então escolhidos como reduções para o treinamento da rede neural. As redes neurais treinadas com os dados completos e com os dados reduzidos são comparadas.

Para a visualização dos resultados obtidos, utilizou-se a ferramenta meteorológica GrADS. Os dados de saída da RNA estão no formato de texto ASCII (txt) e requerem um pré-

processamento para integrá-lo ao ambiente GrADS, esses dados são convertidos para o formato binário, utilizando scripts desenvolvidos pela ferramenta FORTRAN [2].

5.1 Base de dados

Os dados utilizados nos experimentos conduzidos neste trabalho foram coletados da base de dados de reanálise do NCEP/NCAR (National Centers for Environmental Prediction / The National Center for Atmospheric Research) [http://www.ncep.noaa.gov/], para o período de janeiro de 1980 a dezembro de 2000, para uma área contida entre as latitudes [10° N, 35° S] e longitudes [80° W, 30° W], referente à América do Sul, com resolução espacial, em ambas as dimensões (x, y), de 2.5° e resolução temporal (t) de 1 mês. As variáveis contidas na base de dados são: temperatura do ar, divergência, precipitação, umidade específica, pressão da superfície, componentes vento zonal -300, 500 e 850 hPa e meridional -300, 500 e 850 hPa.

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 3 anos (janeiro de 1998 a dezembro de 2000) para a validação do treinamento. A métrica para quantificar o desempenho da previsão foi o erro quadrático médio (EQM) dado por:

$$EQM = \frac{1}{N}\sum_{k=1}^{N}(d(k) - \mathrm{y}(k))^2 \tag{5.1}$$

Em que N é o número de padrões apresentados para a rede, $d(k)$ é o vetor esperado e $y(k)$ é a saída obtida pela rede.

5.2 Resultados

Nesta seção são mostrados os resultados obtidos através da estimativa de redes neurais artificiais com o conjunto de dados reduzidos pelo uso da TCA e o conjunto de dados completos, para previsões climáticas de precipitação.

A arquitetura da rede neural utilizada neste trabalho foi escolhida durante testes preliminares variando o número de neurônios na camada escondida e o número de épocas. Após vários testes, verificou-se que o número de neurônios entre 18 e 22, com 1000 épocas de treinamento, era suficiente para a maioria dos casos testados, então se escolheu configurações condizentes com estes números encontrados empiricamente. A função de transferência é do tipo logística sigmoidal.

Os experimentos foram realizados sobre a região Nordeste do Brasil, entre as longitudes: 312.5°, 315°, 317.5°, 320°, 322.5° e 325° e entre as latitudes: -17.5°, -15°, -12.5°, -10°, -7.5°, -5°, 2.5° e 0°, abrangendo todos os estados da região Nordeste, como pode ser observado na Figura 5.1.

Na fase da redução dos atributos, utilizou-se o ROSETTA para calcular os redutos mínimos. Inicialmente é feita uma discretização por meio de um algoritmo do próprio sistema ROSETTA o *Equal frequency binning*, para fazer a discretização de forma automática. Os dados discretizados são então submetidos ao algoritmo de redução *RSESGeneticReducer* para calcular os redutos mínimos e selecionar os atributos relevantes, com base na TCA.

A redução dos atributos pode ser observada na Tabela 5.1, em que as variáveis com ocorrência igual ou superior a 70% de presença na função de indiscernimento são escolhidas como entradas para o treinamento das redes neurais. Pela Tabela 5.1, observa-se que a dimensão do problema foi reduzida de 11 atributos para 5 atributos.

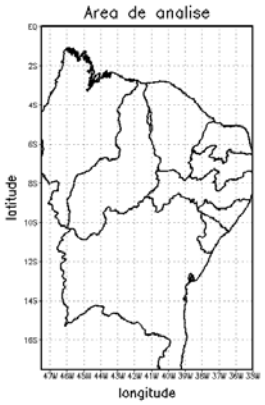

Figura 5.1: Região de análise

Variável	%
airt	75%
u850	82%
u500	71%
v500	73%
v300	85%

Tabela 5.1: Variáveis reduzidas

Nas Figuras 5.2 a 5.5 são apresentados os resultados obtidos no processo de previsão climática, usando todos os dados disponíveis e os dados processados através da TCA. Os resultados são mostrados em um mapa criado pela ferramenta GrADS, para as quatro estações do ano de 1999, em 48 pontos de grade. As Figuras 5.2(a), 5.3(a), 5.4(a) e 5.5(a) representam as situações observadas (denominadas REAL), às quais serão comparadas os resultados das estimativas feitas pelas redes neurais.

Na Figura 5.2 são mostrados os resultados de precipitação obtidos no processo de previsão climática pelas redes neurais, para a estação outono de 1999. Observa-se que a previsão realizada com os dados processados pela TCA, mostra padrões visuais mais semelhantes àqueles mostrado na Figura 5.2(a).

Na Figura 5.3 são apresentados os resultados de precipitação obtidos pelas redes neurais para a estação inverno de 1999. Observa-se que a previsão realizada com os dados pré-processados pela TCA tem padrões mas semelhantes àqueles observados na Figura 5.3(a).

Na Figura 5.4 são mostrados os resultados de previsão de precipitação para a estação primavera de 1999. Observa-se que o resultado de previsão utilizando dados processados pela TCA apresenta uma previsão compatível com o que foi observado e é apresentado na Figura 5.4(a).

Na Figura 5.5 são apresentados os resultados de precipitação obtidos para a estação verão de 1999. Observa-se que ambas as estimativas por redes neurais apresentaram padrões muito semelhantes àqueles presentes na Figura 5.5(a).

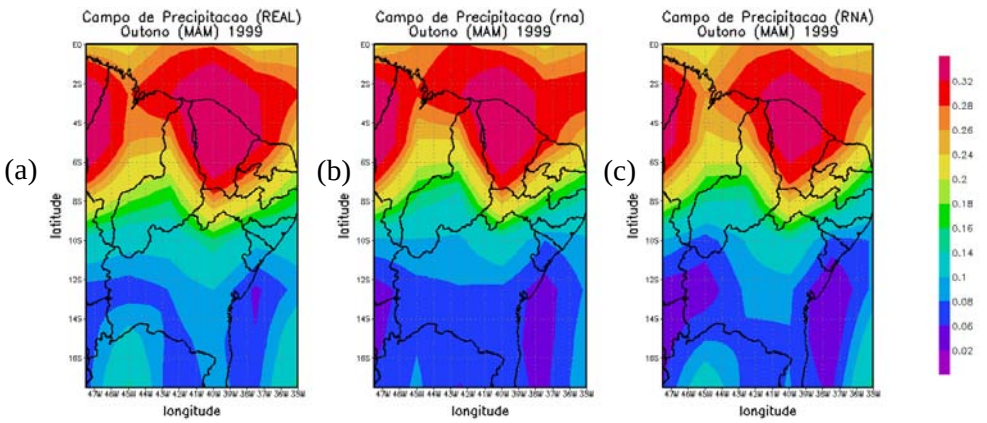

Figura 5.2: Resultado de precipitação. Estação outono de 1999. (a) Precipitação Real; (b) Estimativa com RNA com todos os dados; (c) Estimativa com dados processados por TCA.

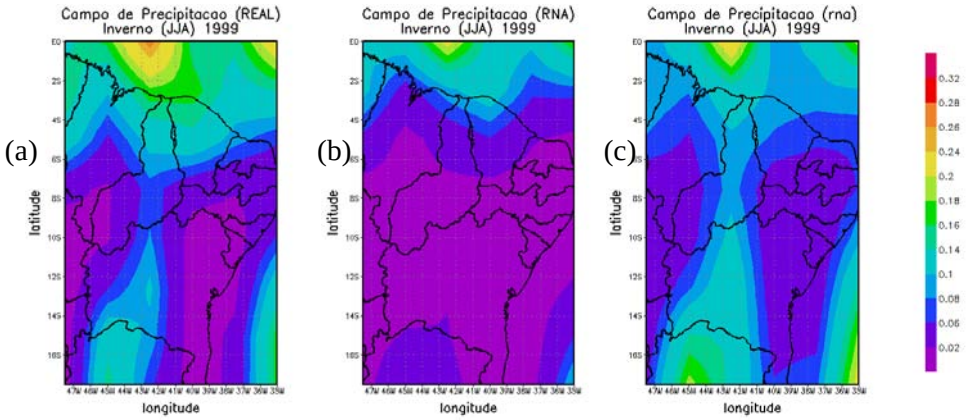

Figura 5.3: Resultado de precipitação. Estação inverno de 1999. (a) Precipitação Real; (b) Estimativa com RNA com todos os dados; (c) Estimativa com dados processados por TCA.

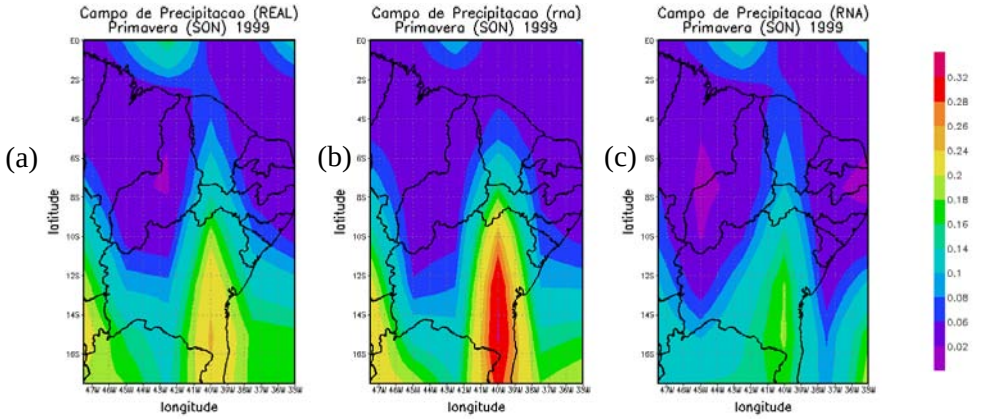

Figura 5.4: Resultado de precipitação. Primavera de 1999. (a) Precipitação Real; (b) Estimativa com RNA com todos os dados; (c) Estimativa com RNA usando dados processados por TCA.

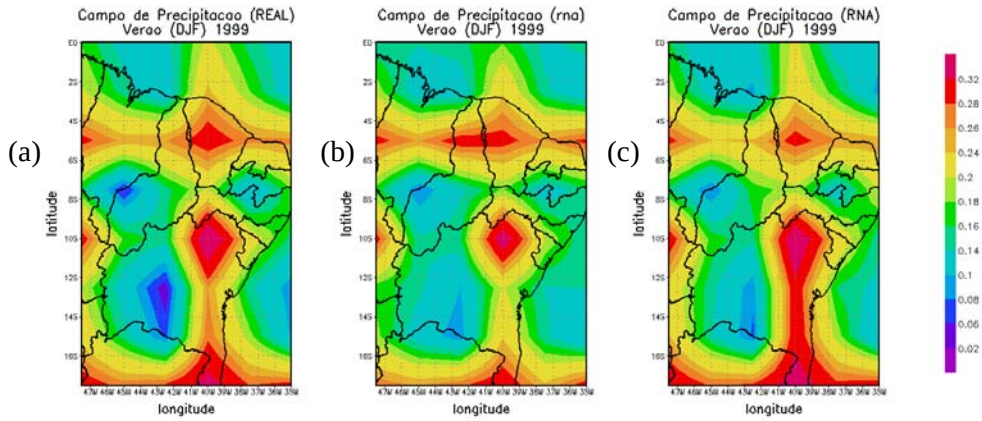

Figura 5.5: Resultado de precipitação. Estação verão de 1999. (a) Precipitação Real; (b) Estimativa com RNA com todos os dados; (c) Estimativa com dados processados por TCA.

A Tabela 5.2 exibe os erros quadráticos médios obtidos durante o processo de previsão climática.

Estação do ano	Dados completos	Dados processados por TCA
Outono	$1{,}34 \times 10^{-5}$	$5{,}8 \times 10^{-5}$
Inverno	$3{,}92 \times 10^{-5}$	$9{,}48 \times 10^{-5}$
Primavera	$8{,}67 \times 10^{-5}$	$5{,}87 \times 10^{-5}$
Verão	$2{,}11 \times 10^{-5}$	$8{,}77 \times 10^{-5}$

Tabela 5.2: Erro quadrático médio

Considerações finais

Neste trabalho foi apresentado o uso de técnicas de Inteligência Artificial para estimar o comportamento médio atmosférico sobre a região Nordeste do Brasil. Os dados disponíveis foram utilizados para treinar redes neurais artificiais para fazer estimativa de precipitação na região, sob duas abordagens: uma utilizando todos os dados selecionados para treinamento das redes neurais; e uma segunda, em que os dados foram primeiro pré-processados pela técnica de TCA com o objetivo de identificar as variáveis que mais contribuem para o processo de previsão.

Nos experimentos iniciais a base de dados, correspondente aos 21 anos da série história, continha 11 atributos, com descrito na Seção 5.1.

O uso da TCA reduziu o número de atributos para 5 (Tabela 5.1), implicando na redução do custo computacional no treinamento das redes neurais. Em um computador PC, Pentium Core 2 Duo, 2.2 Ghz, 1GB de memória, a redução de tempo de processamento para o treinamento da rede foi de 34% (72s para 47s). O tempo de processamento para obtenção das reduções depende da dimensão da base de dados disponível. Para a base considerada o tempo foi inferior a 60s, utilizando o mesmo hardware citado.

As estimativas produzidas pelas redes neurais foram comparadas com os dados reais existentes. Em todos os experimentos, as redes fizeram estimativas muito próximas dos dados reais, mostrando-se adequadas para esta tarefa. Os resultados das redes neurais treinadas com as reduções foram comparados com os resultados das redes neurais treinadas com todos os atributos sendo ambos da mesma ordem de grandeza.

Apesar dos resultados satisfatórios obtidos neste trabalho, a eficiência da metodologia usando TCA na redução dos dados, deverá ainda ser analisada, utilizando outros modelos de redes neurais como modelo de previsão.

Referências bibliográficas

[1] Cavalcanti, I. F. A. Previsão climática no CPTEC-INPE. Disponível em: <http://tucupi.cptec.inpe.br/products/climanalise/cliesp10a/precli.html>. Acesso em: abr 2009.

[2] Doty, B. Grid Analysis and Display System (GrADS). Maryland: Center for Ocean- Land-Atmosphere Studies (COLA). Disponível em: <http://grads.iges.org/grads/head.html>, Acesso em: 23-fev2008.

[3] U. Fayyad, G. P. Shapiro, P. Smyth, “From Data Mining to Knowledge Discovery in databases”, AAAI Press, 1996.

[4] S. Haykin, “Redes Neurais: Princípios e Práticas”, Bookman, Porto Alegre, 2001.

[5] Komorowski, J. e Øhrn, A. “Modelling prognostic power of cardiac tests using rough sets. Artificial Intelligence in Medicine”, pp. 167-1991, (1999).

[6] A. Øhrn, “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, 1999.

[7] Pawlak Z. Rough sets. “International Journal of Computer and Information Sciences”, vol.1. pp. 341-356, (1982).

[8] R. L. Vianello, A. R. Alves, “Meteorologia Básica e Aplicações”, Viçosa, UFV, 2000.