ISSN 1984-820X
Anais do CNMAC
v.3
PUBLICAÇÃO DA SBMAC
CNMAC2010
XXXIII CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL

Mineração de Dados Meteorológicos para Previsão de Eventos Severos pela Abordagem de Similaridade de Vetores

Glauston Roberto Teixeira de Lima,
INPE - Programa de Pós-graduação em Computação Aplicada
12.227-010 São José dos Campos, SP
E-mail: asapessoa@gmail.com

José Demísio Simões da Silva, **Stephan Stephany,**
INPE – Laboratório Associado de Computação e Matemática Aplicada (LAC)
12.227-010 São José dos Campos, SP
E-mail: demisio@lac.inpe.br, stephan@lac.inpe.br

César Strauss,
INPE – Coordenação de Ciências Espaciais e Atmosféricas (CEA)
12.227-010 São José dos Campos, SP
E-mail: cstrauss@cea.inpe.br

Mirian Caetano, **Nelson Jesus Ferreira**
INPE – Centro de Previsão de Tempo e Estudos Climáticos (CPTEC)
12.630-000 Cachoeira Paulista, SP
E-mail: mirian.caetano@cptec.inpe.br, nelson.ferreira@cptec.inpe.br

***Resumo**: O objetivo do trabalho proposto é detetar possíveis ocorrências de eventos convectivos severos por meio do monitoramento das saídas do modelo meteorológico Eta para cada timestep simulado e para um conjunto de variáveis selecionadas. Um classificador foi desenvolvido pela abordagem de similaridade de vetores de forma a identificar padrões dessas variáveis que possam ser associados a esses eventos. Assumiu-se como premissa que os mesmos possam ser correlacionados com grande número de ocorrências de descargas elétricas atmosféricas. O classificador foi treinado agrupando-se saídas do modelo Eta compostas por essas variáveis com base na densidade de ocorrência de descargas elétricas atmosféricas nuvem-solo. O classificador apresentou bom desempenho para os testes realizados para um período mensal escolhido para 3 mini-regiões selecionadas.*

Palavras-chave: mineração de dados, previsão meteorológica, eventos convectivos, similaridade de vetores

1. Introdução

A previsão de eventos convectivos severos de forma semi-automática e com antecedência desejável é um tema atual de pesquisa em Meteorologia. A necessidade de análise da crescente quantidade de dados e imagens meteorológicos, gerados por sensores ou por simulações, demanda técnicas computacionais avançadas. Nesse escopo, um dos objetivos da mineração de dados é descobrir correlações potencialmente úteis entre os diversos dados ou encontrar regras quantitativas associadas aos mesmos.

No caso do presente trabalho, tenta-se inferir a possibilidade de ocorrência de eventos convectivos severos a partir das saídas do modelo meteorológico regional Eta, as quais fornecem o valor simulado de muitas dezenas de variáveis meteorológicas para um tempo de simulação futuro. Um classificador é o programa que atribui uma classe para o conjunto de valores das variáveis meteorológicas de cada timestep gerado pelo modelo meteorológico. As classes compreendem, por exemplo, evento convectivo severo, ou de média ou fraca

ISSN 1984-820X
Anais do CNMAC
v.3
PUBLICAÇÃO DA SBMAC
CNMAC2010
XXXIII CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL

intensidade, ou ainda ausência de atividade convectiva. O classificador incorpora conceitos de aprendizagem de máquina, os quais possibilitam que o mesmo seja "treinado" a partir de um conjunto de instâncias conhecidas. No caso, as instâncias são o conjunto de saídas do modelo Eta para os quais a intensidade da atividade convectiva é conhecida de forma indireta por meio da densidade de ocorrências de descargas elétricas atmosféricas nuvem-solo. Assume-se aqui que esta densidade possa ser associada à severidade dos eventos convectivos, tal como proposto em [1].

O agrupamento espaço-temporal de ocorrências de descargas elétricas atmosféricas do tipo nuvem-solo foi realizado por meio de uma técnica de análise espacial [5][6] aplicada de maneira inovadora [3][4]. Esse agrupamento gera um campo de densidade de ocorrências de descargas que permite identificar regiões mais densas como sendo centros de atividade elétrica (CAEs). O próprio processo de mineração de dados permite estabelecer de maneira conveniente os limites das faixas de densidade associadas a cada classe.

Foram selecionadas 3 mini-regiões de 1 grau de latitude por 1 grau de longitude no território brasileiro de forma a explorar a localidade espacial dos dados (em contraposição a considerar uma região mais extensa), de forma a eventualmente poder reproduzir padrões específicos de cada mini-região. A primeira mini-região abrange o Pantanal Sul Matogrossense, a oeste da cidade de Corumbá, a segunda mini-região é delimitada pelas cidades de Bauru e Presidente Prudente, na Alta Sorocabana paulista e a terceira fica no Vale do Paraíba, abrangendo São José dos Campos, Taubaté e parte do litoral norte paulista.

2. Metodologia

Os dados binários do modelo meteorológico Eta foram fornecidos pelo CPTEC/INPE, sendo referentes aos meses de janeiro e fevereiro de 2007. Uma análise inicial dos dados por parte de meterologistas visando a mineração de dados associada à deteção de eventos convectivos severos levou à seleção de 65 variáveis deste modelo. Foi realizado um pré-processamento de forma a se obter tabelas em formato texto (ASCII) para 6 mêses de dados da primavera de 2006 ao verão de 2007 para uma extensão geográfica correspondente a uma faixa de 20 graus de latitude por 20 graus de longitude (ou 101 pixels por 101 pixels, considerando a resolução de 20 km dos dados). Uma nova análise, efetuada posteriormente com um meteorologista, reduziu o número de variáveis a 26, sendo os dados portados para o ambiente MATLAB para seleção em termos de intervalo de tempo e abrangência geográfica. Os dados brutos de descargas, contendo os registros individuais em formato ASCII foram gerados pela Rede Integrada Nacional de Detecção de Descargas Atmosféricas (RINDAT), fornecidos pelo CPTEC/INPE, e processados pela ferramenta EDDA [7] de forma a gerar os campos de densidade de ocorrência de descargas elétricas atmosféricas para uma extensão geográfica e intervalo de tempo selecionados. A ferramenta implementa o estimador de núcleo gaussiano com janela adaptativa, sendo gerados arquivos em formato ASCII adequados a algoritmos de mineração e em formato de grade binário para as ferramentas de visualização meteorológica GRADS. Parâmetros específicos podem ser ajustados de forma a se poder correlacionar a densidade com outros dados, objetivando seu uso na mineração de dados meteorológicos. A densidade de ocorrência de descargas, que constitui no caso o atributo de decisão, em termos de indicar ou não ocorrência de atividade convectiva, foi então discretizada em 4 faixas com base numa avaliação feita para casos de atividade convectiva severa conhecidos. Posteriormente, os testes demonstraram a conveniência de se adotar apenas 3 faixas, correspondentes a atividade convectiva desprezível/fraca, média e forte. As 3 mini-regiões foram definidas como:

A) Pantanal Sul Matogrossense: latitudes 18:30 a 19:30 graus sul e longitudes 56:30 a 57:30 graus oeste.
B) Alta Sorocabana paulista entre Bauru e Presidente Prudente: latitudes 21:30 a 22:30 graus sul e longitudes 49:30 a 50:30 graus oeste.
C) Parte do Vale do Paraíba e Litoral Norte : latitudes 23:00 a 24:00 graus sul e longitudes 45:00 a 46:00 graus oeste.

ISSN 1984-820X
Anais do CNMAC v.3 PUBLICAÇÃO DA SBMAC
CNMAC2010 XXXIII CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL

2.1 Projeto do classificador

O presente classificador, baseado na abordagem de similaridade de vetores, foi desenvolvido admitindo-se duas hipóteses:

- **Hipótese 1**: a informação contida em um vetor de variáveis meteorológicas é suficiente para identificar o nível de atividade convectiva expresso pela densidade de ocorrência de descargas elétricas atmosféricas.
- **Hipótese 2**: é viável agrupar os vetores de variáveis meteorológicas nos *clusters* que representarão os diversos níveis (ou classes) de atividade convectiva adotando como critério de decisão apenas a densidade de ocorrência de descargas elétricas atmosféricas.

O primeiro passo de projeto foi a formação de um subconjunto de dados para o treinamento do classificador. A escolha recaiu sobre os vetores de 5 datas nas quais se verificou bom equilíbrio entre casos com atividade convectiva significativa e casos sem atividade convectiva significativa. Obteve-se assim uma base de treinamento com 204.020 vetores que foram agrupados em *clusters* de acordo com uma discretização dos valores de densidades de descargas inicialmente adotada que considerava os seguintes 4 níveis: densidade desprezível (ou nula), fraca, média e forte.

Para os vetores da base de treinamento, os valores das densidades de descarga variaram entre 0 e 0.2517. Por sugestão dos meteorologistas, o limiar inferior de densidade de descarga, delimitando os casos de atividade convectiva desprezível, foi tomado como $1.0E10^{-4}$. Esta discretização das densidades de descargas em 4 níveis foi adotada inicialmente, mas, no decorrer dos trabalhos, analisando os resultados de vários testes, optou-se por se fundir todos os vetores representantes de atividade desprezível e fraca numa só classe resultando nos *clusters* C_1, C_2 e C_3, (classes desprezível/fraca, média e forte) contendo N_1, N_2 e N_3 vetores, respectivamente.

A primeira abordagem de classificação testada comparava o vetor a ser classificado com todos os vetores alocados nos 3 *clusters* utilizando a métrica de similaridade descrita a seguir. Para cada uma das 26 variáveis era calculada sua variação máxima na base de treinamento e, para cada variável separadamente, eram calculadas as diferenças entre todos os seus valores normalizando-se essas diferenças pelas respectivas variações máximas. Os desvios padrão de cada um desses conjuntos de diferenças normalizadas foram adotados como limiares para se decidir se dois vetores eram semelhantes ou não. Assim, sendo v_A^j e v_B^j, para j = 1, ..., 26, os vetores em comparação, os mesmos eram considerados semelhantes se atendessem à condição:

$$\frac{\text{abs}\,(v_A^j - v_B^j)}{a_j} \leq d_j \quad \text{para todo } j = 1, \ldots, 26 \qquad (1)$$

Onde, abs corresponde ao valor absoluto e a_j e d_j são, respectivamente, os valores da variação máxima e do desvio padrão previamente calculados a partir de todos os valores da j-ésima variável na base de treinamento. O vetor em classificação era atribuído ao *cluster* (ou classe) com o maior número de vetores semelhantes ao próprio de acordo com a métrica em (1).

Entretanto, essa primeira abordagem não produziu bons resultados. Foi desenvolvido então um novo esquema de classificação baseado em matrizes de probabilidades cruzadas estimadas a partir dos valores de cada variável do vetor a ser classificado e considerando-se a distribuição de valores destas variáveis nos *clusters* de cada classe. Denominam-se estas matrizes de probabilidade como M_1, M_2 e M_3, cada uma delas com dimensão 26 e correspondendo a cada classe. Este esquema de classificação é explicado mais detalhadamente a seguir.

ISSN 1984-820X
Anais do CNMAC v.3
PUBLICAÇÃO DA SBMAC
CNMAC2010 XXXIII CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL

Seja V = [v^1, v^2, ... , v^{25}, v^{26}] um vetor a ser classificado como pertencente a um dos 3 *clusters*.

Para cada variável v^j de V (ou seja, para cada j) são realizados os seguintes 4 passos de cálculo:

Passo 1: considerando-se as j-ésimas colunas de C_1, C_2 e C_3 são contados quantos vetores de cada *cluster* têm o valor de sua j-ésima variável na vizinhança $v^j \pm d(j)$, onde d(j) corresponde ao desvio padrão dessa variável, resultando nas quantidades Q_1^j, Q_2^j e Q_3^J.

Passo 2: constroem-se então as matrizes M_1, M_2 e M_3, sendo atribuídos aos j-ésimos elementos da diagonal os valores: Q_1^j / N_1, Q_2^j / N_2, Q_3^j / N_3.

Passo3: consideram-se as demais variáveis v^i ($i \neq j$) do vetor V a ser classificado e contam-se nas i-ésimas colunas de C_1, C_2 e C_3 quantos dentre os Q_1^j, Q_2^j e Q_3^j vetores (identificados no passo1) têm também o valor de sua i-ésima variável na vizinhança $v^i \pm d(i)$. Resultando nas quantidades [q_1^{j1}, ..., q_1^{j26}], [q_2^{j1} , ..., q_2^{j26}] e [q_3^{j1} , ..., q_3^{j26}].

Passo 4: preenchem-se então as demais posições da j-ésima coluna das matrizes M_1, M_2 e M_3 com os valores: [q_1^{j1}, ..., q_1^{j26}]/Q_1^j, [q_2^{j1} , ..., q_2^{j26}]/Q_2^j e [q_3^{j1} , ..., q_3^{j26}] /Q_3^j.

A Figura 1 mostra as matrizes M_1, M_2 e M_3 com as j-ésimas colunas preenchidas.

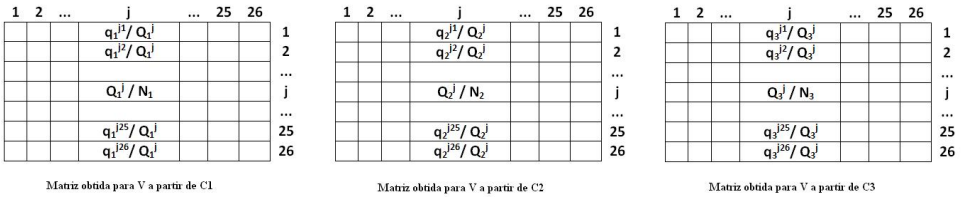

Figura 1: Construção de matrizes M_1, M_2 e M_3 de probabilidades estimadas com base em freqüências de ocorrência dos valores de um vetor V a ser classificado.

Passo 5: a repetem-se os passos de 1 a 4 para os demais *j*'s até que as 3 matrizes estejam completamente preenchidas.

Passo 6: finalmente, somam-se os valores de todos os elementos de cada uma das 3 matrizes e o vetor V é classificado como pertencente à classe (ao *cluster*) correspondente à maior soma.

O esquema acima descrito classifica um vetor de variáveis meteorológicas considerando-o como um conjunto de eventos discretos (o valor de cada variável sendo o evento) e, para cada um desses eventos, estima probabilidades de pertinência a uma dada classe com base em freqüências de ocorrência no *cluster* representativo da classe que, grosso modo, é o espaço amostral da mesma. Deve-se notar que a idéia de similaridade entre vetores é explorada neste esquema de classificação uma vez que ela é a ferramenta utilizada para se apurar as freqüências de ocorrências de cada evento (através da delimitação nos *clusters* das "regiões de semelhança" em torno do valor de cada variável do vetor a ser classificado). São estimadas probabilidades para cada variável individualmente, depois para todas as combinações de variáveis tomadas duas as duas e a probabilidade geral de pertinência a uma classe é obtida como uma soma simples desses valores. Pode-se dizer então que o esquema proposto se inspira no método de máxima verossimilhança [2], mas seu mérito está no fato de que a classificação, ao contrário deste método, é feita sem a necessidade do cálculo (ou estimação) de quaisquer funções de densidade de probabilidades. A forma de comparar os vetores de dados (tomando-se as variáveis individualmente e em grupos de duas) é preferível a uma comparação simultânea das 26 variáveis porque, em sistemas de decisão multi-variáveis, as correlações entre cada variável de informação e a variável de decisão nem sempre são iguais. Esse esquema foi assim proposto visando alcançar um melhor desempenho de classificação, conforme acabou sendo demonstrado pelos resultados apresentados.

3. Resultados

Para realizar os testes de validação do classificador proposto era preciso definir antes os dois limiares para a divisão das densidades de descarga em três faixas, correspondentes às 3 classes propostas. Para este fim, o critério adotado foi o desempenho da classificação com base no índice Kappa. Assim, foi realizada uma varredura para 2.561 pares de limiares selecionados para teste na faixa entre 1.0E-4 e 0.2517. Considerando-se os 204.020 vetores da base de treinamento, os dois melhores limiares encontrados foram $2.5E10^{-3}$ e $1.0E10^{-2}$. Estes valores foram então adotados para montar os *clusters* representativos das 3 classes (correspondentes a atividade convectiva desprezível/fraca, média e forte), que apresentaram as seguintes quantidades de vetores:

C_1: 140.870 (desprezível) + 51.304 (fraca) = 192.174 vetores (94.19% do total)

C_2: 8.168 vetores (4% do total)

C_3: 3.678 vetores (1.8% do total)

Devido ao forte desbalanceamento entre C_1 e as outras duas classes, optou-se por se fazer uma amostragem na qual 10% dos exemplares de C_1 foram selecionados, de forma que fossem mantidas as mesmas proporções entre número de vetores por sub-faixa de densidade de descargas e o número total de vetores que estavam presentes no *cluster* com 192.174 vetores. Assim, nessa amostragem, o novo *cluster* para C_1 ficou com a seguinte quantidade de vetores:

C_1: 14.088 (desprezível) + 5.130 (fraca) = 19.218 vetores (9.42% do total)

Utilizando esses três *clusters* como base de treinamento, foram realizados novos testes e obtidos os resultados (matrizes de confusão) mostrados nas Tabelas 1, 2, 3 e 4, que correspondem ao conjunto de treinamento e a cada uma das 3 mini-regiões escolhidas. Nessas tabelas, as primeiras 3 linhas contêm os percentuais de acertos e de erros, sendo a soma de cada linha igual a 100%. Os números nas quartas linhas dessas tabelas correspondem às quantidades de vetores que foram submetidos à classificação pelo esquema proposto.

A Tabela 1 mostra a matriz de confusão resultante quando o classificador proposto foi testado com os próprios vetores da base de treinamento tendo sido obtido índice Kappa de 0.8189.

	Desprezível/Fraco	**Médio**	**Forte**
Desprezível/Fraco	92,03	6,08	1,89
Médio	4,43	82,97	12,60
Forte	0,44	4,68	94,89
Quantidade de vetores nos *clusters*	19218	8168	3678

Tabela 1: Matriz de confusão para o teste com os vetores da base de treinamento.

A Tabela 2 mostra a matriz de confusão resultante quando o classificador proposto foi testado com os vetores de variáveis meteorológicas da mini-região do Pantanal Sul Matogrossense, tendo sido obtido índice Kappa de 0.9159.

	Desprezível/Fraco	**Médio**	**Forte**
Desprezível/Fraco	98,27	1,54	0,19
Médio	0,26	97,10	2,64
Forte	0	0	100
Quantidade de vetores nos *clusters*	3761	379	91

Tabela 2: Matriz de confusão para o teste com os vetores da mini-região do Pantanal Sul Matogrossense.

ISSN 1984-820X
Anais do CNMAC
v.3
PUBLICAÇÃO DA SBMAC
CNMAC2010
XXXIII CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL

A Tabela 3 mostra a matriz de confusão resultante quando o classificador proposto foi testado com os vetores de variáveis meteorológicas da mini-região da Alta Sorocabana paulista, tendo sido obtido índice Kappa de 0.9085.

	Desprezível/Fraco	**Médio**	**Forte**
Desprezível/Fraco	97,24	2,27	0,49
Médio	1,61	95,87	2,52
Forte	0	3,48	96,52
Quantidade de vetores nos *clusters*	3479	436	316

Tabela 3: Matriz de confusão para o teste com os vetores da mini-região da Alta Sorocabana paulista.

Finalmente, a Tabela 4 mostra a matriz de confusão obtida quando o classificador proposto foi testado com os vetores de variáveis meteorológicas da mini-região do Vale do Paraíba paulista, tendo sido obtido índice Kappa de 0.8950.

	Desprezível/Fraco	**Médio**	**Forte**
Desprezível/Fraco	97,29	1,54	1,17
Médio	0,29	97,35	2,35
Forte	0	0,99	99,01
Quantidade de vetores nos *clusters*	3690	340	202

Tabela 4: Matriz de confusão para o teste com os vetores da mini-região do Vale do Paraíba paulista.

4. Comentários Finais

Este trabalho apresentou resultados da mineração de dados meteorológicos aplicada a eventos para 3 mini-regiões selecionadas do território brasileiro. Foram empregados dados selecionados do modelo meteorológico Eta como atributos de informação e dados de densidade de descargas atmosféricas como atributo de decisão, assumindo que alta densidade de descargas seja indicativa de atividade convectiva severa. Um classificador baseado na abordagem de Similaridade de Vetores foi treinado para uma área mais extensa e testado para cada uma das 3 mini-regiões escolhidas. Os resultados foram expressivos, mostrando que a abordagem proposta pode ser viável para a previsão de ocorrências de eventos convectivos severos a partir das saídas correspondentes aos *timesteps* de tempo simulado futuro de um modelo meteorológico.

Agradecimentos: Os autores agradecem o suporte recebido do CNPq por meio do projeto do Edital Universal denominado “Mineração de Dados Associados a Sistemas Convectivos” ("Cb-Mining", processo 479510/2006-7), bem como o suporte recebido da FINEP por meio do projeto “ADAPT – Tempestades: desenvolvimento de um sistema dinamicamente adaptativo para produção de alertas para a região Sul/Sudeste”, mais especificamente sua Meta 2 – “Mineração de dados para identificação de condições favoráveis à gênese e evolução de tempestades”.

Referências

[1] Caetano, Escobar, Stephany, Menconi, Ferreira, Domingues, Mendes Junior, Visualização de campo de densidade de ocorrências de descargas elétricas atmosféricas como ferramenta auxiliar no nowcasting, em "Proceedings of XIII Latin American and Iberian Congress on Meteorology (CLIMET XIII) and X Argentine Congress on Meteorology (CONGREMET X)", Buenos Aires, 2009.

[2] R.O. Duda, P.E. Hart, D.G. Stork, "Pattern Classification", 2. ed., John Wiley & Sons, New York, 2000.

[3] J. Politi, "Implementação de um Ambiente para Mineração de Dados Aplicada ao Estudo de Núcleos Convectivos", Dissertação de Mestrado em Computação Aplicada, INPE, 2005, INPE-14165-TDI/1082.

[4] J. Politi, S. Stephany, M.O. Domingues, O. Mendes Jr., "Mineração de dados meteorológicos associados à atividade convectiva empregando dados de descargas elétricas atmosféricas", Revista Brasileira de Meteorologia, v.21, n.2, pp. 232-244, (2006).

[5] D. W. Scott, "Multivariate Density Estimation: Theory, Practice, and Visualization", JohnWiley, 1992.

[6] B. W. Silverman, "Density Estimation for Statistics and Data Analysis (Monographs on Statitics and Applied Probability 26)", Chapman and Hall, London, 1990.

[7] C. Strauss, S. Stephany, M. Caetano, "A Ferramenta EDDA de Geração de Campos de Densidade de Descargas Atmosféricas para Mineração de Dados Meteorológicos", submetido ao CNMAC-2010.