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

Mineração de Dados Meteorológicos Associada a Eventos Severos no Pantanal Sul Matogrossense

Alex Sandro Aguiar Pessoa,
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 conjuntos aproximativos de forma a identificar saídas referentes a timesteps simulados do modelo 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 ótimo desempenho para os testes realizados para o Pantanal Sul Matogrossense.*

Palavras-chave: mineração de dados, previsão meteorológica, eventos convectivos, Teoria dos Conjuntos Aproximativos

1. Introdução

A previsão de eventos convectivos severos de forma semi-automática e com antecedência desejável é um tem 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 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 e considerar uma região mais extensa), de forma a eventualmente poder reproduzir padrões específicos de cada mini-região. Embora o artigo enfoque o Pantanal Sul Matogrossense, delimitando a região considerada pela cidade de Corumbá a Oeste, outras duas mini-regiões também foram exploradas, ambas no estado de São Paulo, uma delimitada pelas cidades de Bauru e Presidente Prudente, na Alta Sorocabana, e outra no Vale do Paraíba, abrangendo São José dos Campos, Taubaté e parte do litoral norte paulista.

2. Metodologia

2.1 Softwares empregados

A mineração de dados proposta requer dados meteorológicos, no caso, do modelo meteorológico Eta, usados como atributos de informação, dados de descargas elétricas, cuja densidade foi associada à ocorrência de atividade convectiva severa, constituindo o atributo de decisão e um classificador. Inicialmente foram efetuados testes usando uma árvore de decisão J48 (evolução do algoritmo C4.5 [8]), mas posteriormente passou-se a usar um classificador baseado na teoria de Conjuntos Aproximativos (*Rough Sets*) [5], que demonstrou maior robustez. Os softwares usados no processo estão listados a seguir, não havendo nenhum software proprietário:

- MySQL (http://www.mysql.com/) – MySQL *Structured Query Language*, da Sun Mycrosystems (EUA).
- Weka (www.cs.waikato.ac.nz/ml/**weka**/) – *Data Mining Software in Java*, da Universidade de Waikato (Nova Zelândia);
- ROSETTA (http://www.lcb.uu.se/tools/**rosetta**/) - *Rough Set Toolkit for Analysis of Data*, da Universidade de Uppsala (Suécia);
- RSES (*logic.mimuw.edu.pl/~**rses**/*) – *Rough Set Exploration System*, da Universidade de Rzeszów (Polônia).
- EDDA - *Estimação de Densidade de Descargas Elétricas Atmosféricas*, desenvolvido pelo LAC/INPE.

O MySQL é um sistema de gerenciamento de base de dados e o Weka é um ambiente para mineração de dados que incorpora vários algoritmos, inclusive a árvore de decisão J48, testada preliminarmente no escopo deste trabalho. O ROSETTA é também um ambiente para mineração, mas voltado para a teoria de conjuntos aproximativos e utiliza a biblioteca RSES, a qual pode também ser utilizada independentemente do ROSETTA. Finalmente, a ferramenta EDDA, estima a densidade de ocorrência 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 a 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.

2.2 Dados empregados

Os dados binários do modelo meteorológico Eta foram fornecidos pelo CPTEC/INPE, sendo referentes aos mêses de janeiro e fevereiro de 2007. Uma análise inicial dos dados por parte de meteorologistas visando a mineração de dados associada à deteção de eventos convectivos severos levou a selecionar 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 e sendo os dados portados para o MySQL para seleção em termos de intervalo de tempo e abrangência geográfica, sendo os dados usados no presente trabalho, correspondentes aos mêses de janeiro e fevereiro de 2007 e às 3 mini-regiões de interesse. 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 de forma a gerar os campos de densidade de ocorrência de descargas elétricas atmosféricas, também portados para o MySQL. A densidade de ocorrência de descargas, que constitui o atributo de decisão foi então discretizada em 4 faixas com base numa avaliação feita para casos de atividade convectiva severa conhecidos. Posteriormente, optou-se por adotar apenas duas faixas, num esquema de binarização (abrangendo o caso positivo apenas densidades altas). 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.

2.3 Teoria dos Conjuntos Aproximativos (TCA)

No mundo real as informações são frequentemente incertas, imprecisas ou incompletas, talvez devido à dificuldade em relatar os fenômenos naturais observáveis, expressar acontecimentos ou fatos, etc. Diversas teorias foram desenvolvidas para "tratar" tais imperfeições, dentre elas a teoria dos conjuntos nebulosos, teoria de Dempster-Shafer, teoria das possibilidades. No início da década de 80, surgiu uma teoria, caracterizada pela simplicidade e bom formalismo matemático, o que facilita a manipulação de informações, em especial, incertas, conhecida como Teoria dos Conjuntos Aproximativos (TCA), ou do inglês Rough Set Theory [5]. A TCA é uma extensão da teoria dos conjuntos, que enfoca o tratamento de incerteza dos dados através de uma relação de indiscernibilidade que diz que dois elementos são ditos indiscerníveis, se possuem as mesmas propriedades, segundo Leibniz [10]. Alguns autores apontam como a principal vantagem da teoria dos conjuntos aproximativos a não necessidade de utilização de informações adicionais, tais como distribuição de probabilidade, grau de pertinência, possibilidade ou atribuição de crença.

A teoria dos conjuntos aproximativos tem como alicerce três conceitos importantes para a análise de dados: a relação de indiscernibilidade, as aproximações dos conjuntos e as reduções de variáveis ou atributos. Em TCA o conjunto de dados é representado por meio de um sistema de decisão (SD), que é S = (U; A ∪ $\{d\}$), onde U é o universo de discurso, A são os atributos condicionais e $d \notin A$ é o atributo de decisão.

Uma característica importante da TCA é a compactação dos dados durante as análise em ao menos dois modos: através da relação de indiscernibilidade (IND) e das reduções de atributos condicionais (RED). Na primeira, elementos iguais são representados somente por um elemento pertencente a uma classe de equivalência ou em outras palavras, elementos iguais pertencem a mesma classe de indiscernibilidade. Na segunda o conjunto reduzido (B) deve possuir a mesma partição do universo (U) do que o conjunto completo de atributos (A), ou seja, IND(B) = IND(A).

Em termos práticos, para um classificador TCA, a saída (i.e. o resultado da classificação) é determinada pelo limiar *Thr* (0<=*Thr* <=1), conforme mostra a equação abaixo:

$$\theta(\phi(x)) = 1 \; se \; \phi(x) \geq Thr \quad \text{ou} \quad \theta(\phi(x)) = 0 \; se \; \phi(x) < Thr \qquad (1)$$

onde $\theta(\Phi(x))$ é a saída dada em função do grau de certeza $\Phi(x)$ da instância x, que é comparado ao limiar *Thr*. O grau de certeza para a classe alvo é calculado por um esquema de votação, com base na cobertura relativa das instâncias dessa classe que foram usadas no treinamento.

2.4 Métricas de Avaliação

Foram utilizadas algumas métricas de avaliação comuns de classificadores, tais como matriz de confusão e a curva ROC (*Receiver Operating Characteristic*).

A matriz de confusão é uma matriz cuja dimensão é dado pelo número de classes que sumariza os resultados de um dado classificador. A diagonal principal da matriz exibe o número de acertos para as classes analisadas, enquanto que os elementos fora da diagonal, o número de erros. No caso de duas classes (sim ou não, d(x) = 0 ou 1) tem-se uma matriz 2x2, como mostrado abaixo:

		Predição			
Real	d(x)	d(x)			
			0	1	
		0	NV	FP	Especificidade
		1	FN	VP	Sensibilidade
			VNP	VPP	Acurácia

Figura 1: Matriz de confusão de ordem 2

Onde NV são negativos verdadeiros, VP são verdadeiros positivos, FN são falsos negativos e FP são falsos positivos. Os dois primeiros referem-se aos acertos do classificador e os dois últimos aos erros. Ainda existem outras métricas importantes sendo estas:

- Especificidade = NV/(NV+FP);
- Sensibilidade = VP/(VP+FN);
- VNP (Valor Negativo Preditivo) = NV/(NV+FN);
- VPP (Valor Positivo Preditivo) = VP/(VP+FP);
- Acurácia = VP+VN/(VP+VN+FP+FN).

A especificidade e sensibilidade estão ligados as taxas de acertos do classificador com relação ao real ou verdade. Os valores de VPP e VNP dizem respeito às taxas de acerto do classificador com relação às classes estudadas. E o parâmetro mais importante da matriz de confusão é a acurácia que expressa a taxa de acerto global.

Outra forma de avaliação de resultados é a análise da curva ROC, que é a representação gráfica da discriminação. Define-se discriminação como sendo a habilidade de um classificador separar objetos de classes de decisão diferentes. A Figura 2, ilustra como a curva ROC é

sempre delimitada pelo caso ideal, em que um classificador tem perfeita habilidade discriminatória (segmento de reta vermelho) e o pior caso, em que o classificador não tem habilidade discriminatória (segmento de reta azul). Assim, como a área sob o primeiro segmento é unitária, quanto melhor o classificador, mais próxima da unidade será a área sob a curva ROC.

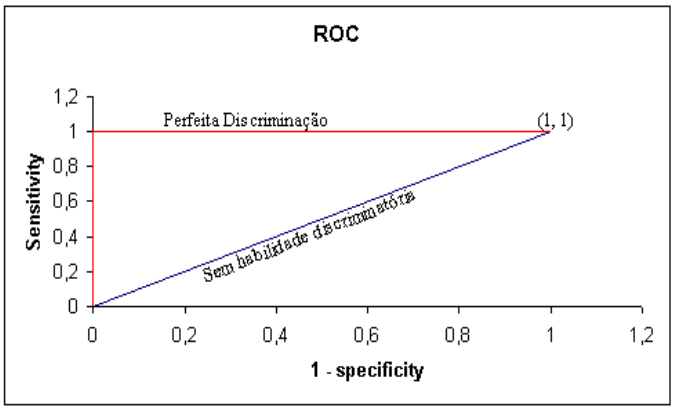

Figura 2: Segmentos de reta que delimitam a curva ROC.

3. Resultados

As análises para as três regiões supramencionadas, denominadas A, B e C (Pantanal Sul Matogrossense, Alta Sorocabana Paulista e parte do Vale do Paraíba e Litoral Norte, respectivamente), foram processadas no software ROSETTA, no esquema de amostragem *holdout*, que usa 80% das instâncias para treinamento e as demais 20% para validação do classificador. O tamanho da amostra é de 5900 elementos para A e B e 8496 elementos C. Na Figura 3 são mostradas as saídas do software ROSETTA para as regiões A, B e C. Aparecem as correspondentes matrizes de confusão e outros parâmetros tais como os valores da área sob a curva ROC para a classe alvo (classe 1, correspondente á ocorrência de atividade convectiva severa). Para a mesma classe, aparece o valor do limiar *Thr* (0 ,1) usado na classificação, bem como outros parâmetros definidos na Figura 1.

Todos os resultados em termos de acurácia possuem valores acima de 95%, porém a sensibilidade (medida relacionada com os valores de acerto da classe 1) difere entre as regiões, sendo a mini-região A a de melhor desempenho, com aproximadamente 83% de acertos para a classe 1. Para as regiões B e C este índice cai para pouco mais de 60%. Isto significa que a classificação de instâncias correspondentes a atividade convectiva severa foi melhor para a mini-região A, ficando um pouco comprometida para as demais regiões (B e C). Pode-se especular que esse resultado deva-se à capacidade do modelo meteorológico Eta reproduzir em maior ou menor grau as características do micro-clima de região.

		Predicted		
Actual		0.000	1.000	
	0.000	1148	3	0.997394
	1.000	5	24	0.827586
		0.995663	0.888889	0.99322
ROC	Class	1.000		
	Area	0.986758		
	Std. error	0.014971		
	Thr. (0,1)	0.172		
	Thr. acc.	0.504		

(a) Mini-região A

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

		Predicted		
Actual		0.000	1.000	
	0.000	1104	8	0.992806
	1.000	26	42	0.617647
		0.976991	0.84	0.971186
ROC	Class	1.000		
	Area	0.922754		
	Std. error	0.02254		
	Thr. (0,1)	0.108		
	Thr. acc.	0.508		

(b) Mini-região B

		Predicted		
Actual		0.000	1.000	
	0.000	1565	31	0.980576
	1.000	37	66	0.640777
		0.976904	0.680412	0.959976
ROC	Class	1.000		
	Area	0.885268		
	Std. error	0.021676		
	Thr. (0,1)	0.08		
	Thr. acc.	0.692		

(c) Mini-região C

Figura 3: Saída gráfica do software ROSETTA mostrando a matriz de confusão e parâmetros da curva ROC para as 3 mini-regiões.

Analisando esses resultados, pode-se ver que o esquema de classificação proposto teve melhor desempenho na mini-região do Pantanal Sul Matogrossense, apresentando alto índice de acertos para a classe alvo, relativa a atividade convectiva severa. Assim, de um total de total de 1180 saídas do modelo das quais 29 eram de atividade convectiva, o classificador identificou 24 saídas deixando de identificar apenas 5, enquanto que outras 3 foram assim classificadas embora constituíssem falsos positivos. Mas vale ressaltar que os resultados para as demais áreas de estudo, também foram bem aceitáveis.

4. Comentários Finais

Este trabalho apresentou resultados da mineração de dados meteorológicos aplicada a eventos convectivos severos no Pantanal Sul Matogrossense. 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 Teoria de Conjuntos Aproximativos foi treinado e testado para cada uma das 3 mini-regiões escolhidas. Os resultados para o Pantanal Sul Matogrossense 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. Isso mostra que existe um uso potencial do modelo Eta para tal previsão, mesmo considerando-se eventuais discrepâncias com os dados observados [2], ou seja, entre simulação e realidade.

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] M. Caetano, G.C.J. Escobar, S. Stephany, V.E. Menconi, N.J. Ferreira, M.O. Domingues, O. 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] S. E. M. Farias, C. S. Chan, Simulação das características microclimatológicas para o Pantanal Sul-mato-grossense, em "Anais do Primeiro Simpósio de Geotecnologias no Pantanal", Embrapa Informática Agropecuária/INPE, Campo Grande, 2006.

[3] U. Fayyad et al. From data mining to knowledge discovery: an overview, AAAI Press, (1996).

[4] J. Komorowski et al, Rough sets: a tutorial, em: "Rough fuzzy hybridization: A new trend in decision-making" (S.K. Pal e A. Skowron, eds.), Springer-Verlag, Singapore, 1999.

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

[6] 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.

[7] J. Politi, S. Stephany, M.O. Domingues, O. Mendes Junior, "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).

[8] J. R. Quinlan. C4.5: programs for machine learning, em: "Machine Learning", Morgan Kaufmann Pub, 1993.

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

[10] S. Scuderi, Conjuntos rough, Leopoldianum, Revista de Estudos e Comunicação da Universidade Católica de Santos. Ano 27, n. 75, pp. 185-197, (2003).

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

[12] 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).