Implementação de um Ambiente para Mineração de Dados Aplicado ao Estudo de Núcleos Convectivos

Jacques Politi
Programa de Pós Graduação em Computação Aplicada (CAP-INPE)
jpoliti@lac.inpe.br

Stephan Stephany, Margarete Oliveira Domingues
Laboratório de Computação e Matemática Aplicada (LAC-INPE)
{stephan,margarete}@lac.inpe.br

Odim Mendes Junior
Divisão de Geofísica Espacial(DGE-INPE)
odim@dge.inpe.br

Resumo

Neste trabalho foi implementado um ambiente para mineração de dados utilizando a teoria dos conjuntos aproximativos (rough sets). A mineração de dados tem sido utilizada para analisar grandes volumes de dados tentando identificar correlações, padrões freqüentes, anomalias, nos mais variados domínios de aplicações. Utilizou-se o ambiente desenvolvido no estudo de núcleos convectivos, por meio de um de seus traçadores, no caso, descargas do tipo nuvem-solo. Devido à grande quantidade de dados disponíveis, necessitou-se de um método otimizado para a redução do volume de dados de descargas elétricas. Para isto, foram investigados diversos métodos de representação espacial, visando agrupar espacialmente as ocorrências de descargas elétricas em entidades denominadas centros de atividade elétrica (CAEs). Este estudo buscou correlações ocultas e potencialmente úteis em uma base de dados formada pelos CAEs e outros parâmetros metereológicos coletados em estações de radiosondagem.

Palavras-chave: *mineração de dados, rough sets, núcleos convectivos, estações de radiosondagem*

1. Introdução

Nas últimas duas décadas houve um crescimento significativo na quantidade de informação armazenada em formatos eletrônicos. Estima-se que a quantidade de informação no mundo dobra a cada 20 meses [8]. Isso foi proporcionado basicamente pela queda de preços dos equipamentos de processamento e armazenamento, e aos avanços nos mecanismos de captura e geração de dados, tais como, leitores de código de barras, sensores remotos e satélites espaciais.

Segundo Piatetsky-Shapiro [5], estes dados, produzidos e armazenados em larga escala, são inviáveis de serem lidos ou analisados por especialistas por meio de métodos manuais tradicionais, tais como planilhas de cálculos e relatórios operacionais informativos. Por outro lado, sabe-se que grandes quantidades de dados possuem um maior potencial de informação.

Analisar essa crescente quantidade de informação, não é uma tarefa trivial, e técnicas computacionais avançadas são necessárias para descobrir padrões ocultos e potencialmente úteis entre os dados. Esse é o objetivo da mineração de dados, também conhecida como extração de conhecimento, arqueologia de dados ou colheita de informações.

A aplicação dessas técnicas ao estudo de núcleos convectivos é de grande interesse aos metereologistas. Uma vez que o volume de informações coletadas é muito elevado, torna-se muito complexa a tarefa de relacionar os dados de descargas elétricas com outros parâmetros metereológicos, tais como, temperatura, pressão, umidade etc, utilizando técnicas tradicionais.

Em geral, os núcleos convectivos estão associados a um ou mais aglomerados de nuvens Cumulonimbus (Cb). Essas nuvens são caracterizadas pelo forte movimento vertical e grande extensão, cerca de 16km a 18km de altura nos trópicos. [3].

O ciclo de vida dessas nuvens Cb divide-se em três estágios: inicial (ou Cumulus), maduro e dissipativo. Essas nuvens, em seu estado maduro, geram descargas elétricas, que são conseqüência das cargas elétricas que se acumulam a partir da colisão entre diferentes tipos de partículas tais como os cristais de gelo e granizo, atingindo às vezes a carga elétrica total de até centenas de Coulombs. Essas descargas são classificadas em nuvem-solo (NS), solo-nuvem, intranuvens, entre-nuvens, horizontais e para ionosfera. Os relâmpagos são

constituídos por uma ou mais dessas descargas elétricas atmosféricas, de caráter transiente, portando uma alta corrente elétrica (em geral, superior a várias dezenas de quilo-ampéres). Eles são conseqüências das cargas elétricas (10-100 C) que se acumulam em nuvens Cumulonimbus e ocorrem quando o campo elétrico excede localmente a capacidade isolante do ar (acima de 400 kV/m).

2. Metodologia

O processo de mineração de dados, também conhecido por KDD (Knowledge Discovery in Databases – Descoberta de Conhecimento em Banco de Dados) consiste basicamente de seis fases e cada fase pode possuir uma intersecção com as demais. A figura 1 ilustra todo o processo, sendo possíveis realimentações de dados entre as fases.

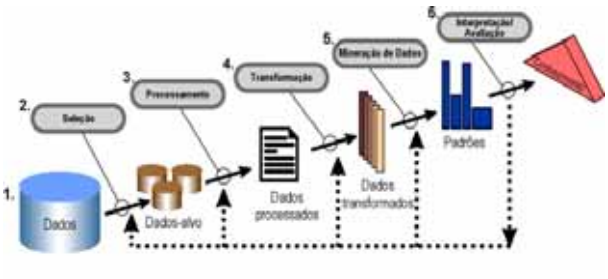

Figura 1. Etapas do ciclo de mineração de dados [2]

As etapas que tomam a maior parte do tempo no processo de descoberta são o pre-processamento e a transformação de dados. Essas etapas visam construir uma base de dados integrada, sem ruídos e erros, e principalmente reduzida, proporcionando um melhor desempenho aos algoritmos de mineração de dados da etapa posterior.

A metodologia proposta divide-se em duas partes distintas: a primeira diz respeito às etapas de pre-processamento e transformação dos dados de descargas elétricas e estações de radiosondagem, e foi implementada no ambiente MATLAB® (versão 6.5). A segunda refere-se à mineração de dados propriamente dita utilizando o sistema ROSETTA [9].

2.1 Pré-processamento no ambiente MATLAB®

Descrevem-se a seguir dois métodos implementados para o pré-processamento dos dados de descargas elétricas, um baseado nos CAEs e outro, nas estações de radiosondagem. Inicialmente, selecionam-se apenas os atributos temporais e espaciais dos dados de descargas elétricas visando gerar campos que expressam de forma reduzida o agrupamento espacial das descargas elétricas. Tais campos podem ser denominados, no contexto deste trabalho, como centros de atividade elétrica (CAEs). Para que esses CAEs sejam formados é necessário utilizar técnicas de representação espacial e também integração temporal dos dados separando as ocorrências de descargas em intervalos de tempo compatíveis com a escala de tempo do fenômeno observado, ou seja, o ciclo de vida das estruturas convectivas.

Dentre as técnicas de representação espacial investigadas, temos: *paintball* (plotagem de eventos), agrupamento em grade, *clustering* (aglomerado) e estimadores de densidade. Esta última técnica foi adotada aqui, pois possibilita gerar CAEs com aparência suavizada [6]. Nessa técnica, utiliza-se uma grade bidimensional, e para cada ponto desta faz-se a contagem das ocorrências de descargas elétricas dentro de um raio limite que atua como um parâmetro de suavização. Na Figura 2, pode-se observar um exemplo de CAE gerado, onde os pontos pretos representam as ocorrências de descargas elétricas.

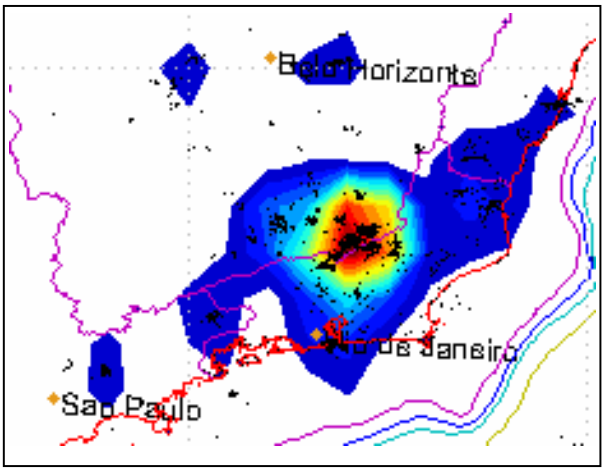

Figura 2. Centros de atividade elétrica

Em seguida, define-se um filtro para os pontos da grade com baixa contagem, eliminando com isso as descargas esparsas e delimitando melhor os CAEs gerados. Terminada a identificação dos CAEs, estes são agrupados com parâmetros representativos da estabilidade atmosférica, provenientes de dados coletados por radiosondagem atmosféricas. Esses parâmetros são os índices CAPE (Convective Available Potencial Energy), K, TT (Totals), e SLI (Lift Index)[1,3].

2.1.1 Pré-processamento baseado nos CAEs. Verifica-se se uma dada estação, está dentro da área de atuação de cada CAE, e em caso positivo, agrupam-se todas as informações da estação ao respectivo CAE. Existem dois testes de pertinência implementados para essa verificação. O primeiro teste, aplicável na maioria dos casos, verifica se a estação está dentro do *perímetro* do CAE. No caso de CAEs compostos por apenas um ponto de grade ou dispostos em uma única linha da grade bidimensional, nos quais não é possível determinar o perímetro, recorre-se ao segundo teste de pertinência. Neste, verifica-se se uma estação está dentro de um raio de influência pré-estabelecido, a partir do centro do CAE. Os CAEs são agrupados em

uma tabela contendo um resumo de suas características tais como a posição espacial de seu centro, sua área, densidade, número de ocorrências de descargas, parâmetros das estações de radiosondagem, entre outras.

2.1.2 Pré-processamento baseado nos dados de radiosondagem. Analisam-se os CAEs tomando por referência todas estas estações disponíveis dentro do raio definido nos horários das 12UTC e 00UTC. A tabela resultante dessa abordagem, contém todas as informações das estações e um parâmetro de decisão que indica a existência de CAEs nas proximidades. As tabelas geradas por ambos os métodos são então utilizadas como entrada do sistema de mineração de dados ROSETTA. Esse sistema é baseado na teoria dos conjuntos aproximativos (*rough sets*), e será detalhada na seção seguinte.

2.2. Mineração de dados utilizando o ROSETTA

O sistema ROSETTA (Rough Set Toolkit for Analisys Data) é um conjunto de componentes de software escrito em C++ utilizado para análise de dados, baseado na Teoria dos Conjuntos Aproximativos (*Rough Sets*). Foi desenvolvido em um esforço cooperativo entre o grupo Knowledge Discover Group da NTNU (Norwegian University of Science and Technology), na Noruega e o Logic Group da Universidade de Varsóvia, Polônia. A Teoria dos Conjuntos Aproximativos foi desenvolvida por Pawlak[4] no começo da década de 80 para lidar com dados incertos e vagos em aplicações de Inteligência Artificial, e tem constituído uma base teórica para a solução de muitos problemas de descoberta de conhecimento. O sistema ROSETTA é capaz de suportar todo o ciclo de mineração de dados, incluindo o pré-processamento e transformação dos dados. Entretanto, os algoritmos responsáveis por essas etapas, não apresentaram resultados satisfatórios aos dados deste trabalho, o que levou à utilização dos métodos descritos na seção anterior. Os dados contidos nas tabelas resultantes do pré-processamento no ambiente MATLAB foram ainda agrupados em classes discretas.

Estes dados foram então utilizados pelo algoritmo de mineração de dados propriamente dito, no caso um algoritmo genético. Este algoritmo efetua as reduções de atributos da base de dados utilizando o conceito de *rough sets*. Geram-se então regras do tipo IF-THEN, que estão sempre associadas a medidas estatísticas que auxiliam na determinação da importância de cada regra com base em sua cobertura estatística, como exemplificado na Tabela 1. Essas regras representam relações entre a ocorrência de descargas elétricas, e os parâmetros coletados pelas estações de radiosondagem. Na primeira coluna desta tabela, tem-se a regra gerada pelo algoritmo genético de redução do sistema ROSETTA. A regra é separada em duas partes pelo símbolo "=>". O lado esquerdo da regra (LHS, *Left Hand Side*), apresenta os atributos condicionais ligados por ***AND (E lógico)***, e lado direito (RHS, *Right Hand Side*), apresenta o atributo de decisão. Os números entre parênteses correspondem aos valores dos atributos e o "*" significa infinito. Na coluna **SUP**, tem-se o número de registros que satisfazem o RHS da regra. Na coluna **COB** (cobertura), tem-se o percentual do número de registros da base de dados que satisfazem uma determinada regra. Na primeira linha da tabela, por exemplo, pode-se interpretar a regra como:

"Em 71% dos casos em que ocorreram descargas próximas das estações de radiosondagem, o parâmetro K foi MAIOR que 32 "

Tabela 1. Exemplos de regras geradas

REGRA	SUP	COB(%)
K([32, *)) => descarga(1)	68	71
SLI([*, -1)) => descarga(1)	54	56
K([32,*))AND TT([44,*))=>descarga(1)	52	54

3. Testes realizados

Os dados de descargas NS analisados neste trabalho são referentes às regiões Sul, Sudeste e Centro-Oeste do Brasil, num período que engloba a terceira campanha do experimento interdisciplinar do Pantanal (IPE-3). Estes dados referem-se ao período de 01 de fevereiro a 30 de março de 2002 e foram cedidos pelo RINDAT (Rede Integrada Nacional de Descargas Atmosféricas). Incluem 25 parâmetros para cada ocorrência de descarga, incluindo tempo, localização, bem como outras características físicas, tais como, pico de corrente, polaridade, e multiplicidade. Os dados das radiosondagens são obtidos de estações de radiosondagem padrões e das radiosondagens do experimento IPE-3 para o mesmo período. E as comparações foram feitas quando ambos os dados estavam disponíveis. Os parâmetros configuráveis de cada teste, são descritos na Tabela 2.

Tabela 2. Descrição dos parâmetros dos testes

Parâmetro	Descrição
MATLAB®	
Definição	Comprimento da lateral de cada cédula da grade bidimensional utilizada para gerar os CAEs (em graus)
Raio	Raio de corte responsável pela suavização dos CAEs gerados (em graus)
Timestep	Tempo de integração das descagas elétricas (em horas)
Área de Influência Radiosonda	Distância máxima para verificar se uma estação de radiosondagem pertence ao domínio do CAE gerado. (em graus)
ROSETTA	
Intervalo	No caso de discretização automática, pode-se escolher entre diversos métodos. O método que apresentou melhores resultados foi *Equal Frequency Binning*, que consiste em dividir a base de dados em um número determinado de intervalos contendo o mesmo número de elementos.

Na Tabela 3 têm-se todos os testes realizados para os dados de descargas do período acima mencionado. O valor automático do raio é calculado a partir de Regra de Silverman [7].

Tabela 3. Descrição dos testes realizados

	IDENTIFICADOR DO TESTE					
PARAMETRO	**1**	**2**	**3**	**4**	**5**	**6**
Definição	0.5	0.5	0.5	0.5	0.5	0.5
Raio	auto	auto	auto	auto	auto	auto
Timestep(h)	1	3	6	1	3	6
Área de Influência Radiosonda	1	1	1	2	2	2
Intervalo	2	2	2	2	2	2

3.1. Análise da ocorrência de descargas elétricas baseado nas estações de radiosondagem

Dentre todas as regras geradas, considerou-se apenas aquelas cuja cobertura foi superior a 50% da base de dados para a classe em que o parâmetro de decisão relativo à proximidade de CAEs da estação era 1 (o valor 0 implica não-proximidade).

Tabela 4. Resultados da análise de ocorrência de descargas elétricas

REGRA	SUP	COB(%)
Teste 1		
K([32, *)) => descarga(1)	68	71
SLI([*, -1)) => descarga(1)	54	56
K([32, *)) AND TT([44, *)) => descarga(1)	52	54
CAPE([*, 703)) => descarga(0)	363	52
Teste 2		
K([32, *)) => descarga(1)	79	71
SLI([*, -1)) => descarga(1)	60	54
CAPE([*, 703)) => descarga(0)	355	52
TT([*, 44)) => descarga(0)	346	51
Teste 3		
K([32, *)) => descarga(1)	80	71
K([32, *)) AND TT([44, *)) =>descarga(1)	66	59
SLI([*, -1)) => descarga(1)	61	54
CAPE([*, 703)) => descarga(0)	351	52
Teste 4		
K([32, *)) => descarga(1)	121	68
K([32, *)) AND TT([44, *)) =>descarga(1)	97	54
TT([*, 44)) => descarga(0)	326	53
SLI([*, -1)) => descarga(1)	93	52
Teste 5		
K([32, *)) => descarga(1)	134	69
TT([*, 44)) => descarga(0)	318	54
CAPE([*, 703)) => descarga(0)	316	53
K([32, *)) AND TT([44, *)) => descarga(1)	103	53
Teste 6		
K([32, *)) => descarga(1)	137	69
K([32, *)) AND TT([44, *)) => descarga(1)	109	55
TT([*, 44)) => descarga(0)	317	54
CAPE([*, 703)) => descarga(0)	309	52

Podem-se observar dois grupos de resultados. No primeiro grupo, composto pelos testes nº 1,2 e 3, os parâmetros que mais se destacaram para a caracterização da ocorrência de descargas elétricas, foram os parâmetros K e SLI, apresentando valores superiores a 32 para o K, e inferiores a -1 para o SLI.

No segundo grupo, o parâmetro K continua sendo importante, entretanto o SLI deixou de ser determinante, sendo substituído pelo parâmetro TT. Para a ocorrência de descargas os valores de K e TT devem ser maiores que 32 e 44 respectivamente. O parâmetro que variou nesses dois grupos de resultados foi à Área de Influência Radiosonda, que no primeiro grupo considerava os CAEs que estivessem até 1º (cerca de 110km) de distância da estação de radiosondagem. Isso resulta em um número reduzido de registros considerados com atividade elétrica, em relação ao segundo grupo. Por outro lado, sabe-se que quanto menor a distância a ser considerada, mais certeza tem-se que os parâmetros influenciam na ocorrência de descargas. Portanto deve-se dar prioridade aos resultados cujo limite de distância seja 1º, e assim, deduz-se que os parâmetros K e SLI sejam os mais importantes nessa análise.

3.2. Análise da densidade de descargas elétricas baseado nos CAEs

Nesta análise, o cálculo da densidade é efetuado a partir da área total de um CAE e do número total de ocorrências de descargas elétricas pertinentes a essa área. Consideraram-se apenas as regras cuja cobertura foi superior a 50%.

Tabela 5. Resultados da Análise de densidade de descargas elétricas

REGRA	SUP	COB(%)
Teste 1		
SLI([*, -1)) => densidade([62, *))	31	69
CAPE([*, 1040)) => densidade([*, 62))	24	53
Teste 2		
SLI([*, -1)) => densidade([163, *))	33	63
CAPE([*, 969)) => densidade([*, 163))	31	58
Teste 3		
SLI([*, -1)) => densidade([330, *))	34	65
CAPE([*, 900)) => densidade([*, 330))	29	55
Teste 4		
SLI([-1, *)) => densidade([*, 72))	53	52
CAPE([754, *)) => densidade([72, *))	52	51
Teste 5		
TT([*, 46)) => densidade([*, 181))	60	55
SLI([*, -1)) => densidade([181, *))	57	52
Teste 6		
SLI([-1, *)) => densidade([*, 379))	64	55
TT([*, 46)) => densidade([*, 379))	63	54
CAPE([874, *)) => densidade([379, *))	59	51

Pode-se perceber que nestes testes, a densidade elevada de descargas (representada pelo * à direita do valor), está relacionada com SLI baixo (inferior a –1) e CAPE elevado. Por outro lado, as baixas densidades estão relacionadas com valores altos de SLI, e valores baixos de TT e CAPE.

3.3. Análise do número de ocorrência de descargas elétricas baseados nos CAEs

Nesta análise, tem-se como parâmetro de decisão, o número total de ocorrências de descargas elétricas. Esse número é calculado a partir da contagem das ocorrências dentro do perímetro de um determinado CAE.

Tabela 6. Resultados da Análise do número de ocorrências de descargas elétricas

REGRA	SUP	COB(%)
Teste 1		
SLI([*, -1)) => n_instancias([246, *))	30	67
CAPE([*, 1040)) => n_instancias([*,246))	24	53
Teste 2		
SLI([*, -1)) => n_instancias([347, *))	33	61
CAPE([*, 969)) => n_instancias([*, 347))	31	57
Teste 3		
SLI([*, -1)) => n_instancias([681, *))	35	65
CAPE([*, 900)) => n_instancias([*, 681))	31	57
Teste 4		
CAPE([754, *)) => n_instancias([94, *))	61	60
SLI([-1, *)) => n_instancias([*, 94))	59	58
TT([46, *)) => n_instancias([94, *))	53	52
Teste 5		
Sem regras		
Teste 6		
SLI([-1, *)) => n_instancias([*, 243))	69	58
CAPE([874, *)) => n_instancias([243, *))	65	55
K([*, 34)) => n_instancias([*, 243))	63	53
TT([46, *)) => n_instancias([243, *))	61	52

Nessa análise, os valores altos para os parâmetros CAPE e TT, e valores baixos para o parâmetro SLI (inferiores a –1) correspondem a uma alta taxa de ocorrência de descargas elétricas. Os valores baixos para o parâmetro CAPE e K, bem como valores altos para SLI, correspondem a uma baixa incidência ou a não existência de descargas elétricas.

4. Considerações finais

Este trabalho apresentou uma metodologia de mineração de dados aplicada à análise de núcleos convectivos e sua correspondente implementação. Na redução espaço-temporal dos dados de ocorrência de descargas foi empregada uma técnica baseada em estimadores de densidade. Na mineração de dados propriamente dita foi empregado o sistema ROSETTA, baseado em rough sets. Os testes realizados visaram encontrar padrões globais entre a ocorrência de descargas, densidade, e a quantidade de descargas em relação aos parâmetros K, TT, SLI e CAPE, processados em estações de radiosondagem.

Dentre os resultados que mais significativos, tem-se os testes referentes à ocorrência de descargas (Tabela 4), onde se pode perceber que os valores elevados de K e valores baixos de SLI correspondem à ocorrência de descargas elétricas.

Apesar de alguns resultados apresentarem significados estatísticos baixos, de maneira global são capazes de expressar de forma resumida alguns padrões de conhecimento geral dos metereologistas, validando, portanto a metodologia proposta. Outras técnicas estão sendo investigadas, para buscar melhorar o suporte estatístico das regras encontradas. Com a mesma finalidade, pretende-se realizar testes com bases de dados mais extensas. Isso possibilitaria encontrar padrões desconhecidos que possam ser úteis para os metereologistas.

5. Agradecimentos

Os autores agradecem ao Met. Cesar A. A. Beneti (SIMEPAR) e ao RINDAT pelos dados de descargas elétricas atmosféricas utilizados neste trabalho, ao CPTEC/INPE (pelos dados de radiosondagem), à FAPESP (projeto IPE, processo n° 1988/0105-5, pelos dados), e ao CNPq pelo apoio financeiro fornecido (processos n° 478707/2003-7, 477819/03-6 e 131384/2003-1) .

6. Referências

[1] Domingues, M. O. , Mendes O. Jr, Chan, C. S. , Sá, L. D. A. e Manzi, A. O., *Análise do Tempo durante o Experimento Interdisciplinar do Pantanal Fase 2*, Revista Brasileira de Meteorologia, 19(1), 73-88, 2004.

[2]Fayyad, U. M., Piatetsky Shapiro, G., Smyth - The KDD Process for Extracting Useful Knowledge from Volumes of Data - *Knowledge Discovery In Communications Of The Acm* November 1996/Vol. 39, No. 11

[3] MacGorman, D. R.; Rust, W. D. *The electrical nature of storms*. Oxford: Oxford University, 1998. 422 p.

[4] Pawlak, Z., Rough Sets. *International Journal of Computer and Information Sciences* 11, 1982, pp. 341-356

[5] Piatetsky-Shapiro, G. Knowledge discovery in real databases: *A report on the IJCAI-89 Workshop. AI Magazine*, Vol. 11, No. 5, Jan. 1991, Special issue, 68-70.

[6] Politi. J., Stephany. S., Domingues, M. O., Mendes Jr. O. Uma Metodologia para Representação Espaço-Temporal de Ocorrências de Descargas Elétricas Nuvem-Solo – *Revista Brasileira de Metereologia (submetido, julho de 2004),*

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

[8]Szalay,A.; Kunszt,P. Z.; Thakar,A.; Gray,J.; Slut, D. R. Designing and mining multi-terabyte astronomy archives: The sloan digital sky survey. *Proceedings of the ACM SIGMOD*, pages 451-462. ACM Press, 2000.

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