Revista Brasileira de Meteorologia, **v.27**, **n.1**, 61 - 74, 2012

MINERAÇÃO DE DADOS METEOROLÓGICOS PARA PREVISÃO DE EVENTOS SEVEROS

ALEX SANDRO AGUIAR PESSOA[1], GLAUSTON ROBERTO TEIXEIRA DE LIMA[1], JOSÉ DEMÍSIO SIMÕES DA SILVA[1], STEPHAN STEPHANY[1], CESAR STRAUSS[1], MIRIAN CAETANO[2] E NELSON JESUS FERREIRA[2]

[1]Instituto Nacional de Pesquisas Espaciais, Laboratório Associado de Computação e Matemática Aplicada (INPE/ LAC)
[2]Centro de Previsão de Tempo e Estudos Climáticos (INPE/CPTEC), São José dos Campos, SP, Brasil

asapessoa@gmail.com, glau11@gmail.com, demisio@lac.inpe.br, stephan@lac.inpe.br, cstrauss@cea.inpe.br, mirian.caetano@cptec.inpe.br, nelson.ferreira@cptec.inpe.br

Recebido Agosto de 2010 – Aceito Julho de 2011

RESUMO

O objetivo do trabalho proposto é detectar antecipadamente possíveis ocorrências de eventos convectivos severos, por meio do monitoramento das saídas do modelo de previsão numérica de tempo Eta, para cada intervalo de previsão e para um conjunto de variáveis selecionadas. O período de estudo estende-se de janeiro a fevereiro de 2007. Classificadores foram desenvolvidos pela abordagem de similaridade de vetores e de conjuntos aproximativos, de forma a identificar saídas do modelo Eta que possam ser associados a esses eventos. Assumiu-se como premissa que os eventos convectivos severos possam ser correlacionados com grande número de ocorrências de descargas elétricas atmosféricas. Os classificadores agruparam as saídas do modelo Eta, compostas por essas variáveis, com base na densidade de ocorrência de descargas elétricas atmosféricas nuvem-solo. Ambos os classificadores apresentaram bom desempenho para os testes realizados para um período de dois meses escolhido para três mini-regiões selecionadas do território brasileiro.
Palavras-Chave: mineração de dados, previsão meteorológica, eventos convectivos.

ABSTRACT: METEOROLOGICAL DATA MINING FOR THE PREDICTION OF SEVERE CONVECTIVE EVENTS

This work aims the early detection of possible occurrences of severe convective events in Central and Southeast Brazil by means of monitoring the output of the Eta numerical weather prediction model for each forecasted time interval and for a selected set of variables. The studied period ranges from January to February 2007. Classifiers were developed by two approaches, vector similarity and rough sets, in order to identify Eta outputs that can be associated to such events. It was assumed that severe convective events can be correlated to a large number of atmospheric electric discharges. The classifiers grouped the Eta meteorological model outputs for these selected variables based on the density of occurrences of cloud-to-ground atmospheric electrical discharges. Both classifiers show good performance for the chosen 2-month period at the three selected mini-regions of the Brazilian territory.
Keywords: data mining, weather forecast, convective events.

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 meteorológicos e imagens, gerados por sensores ou por modelos de previsão numérica de tempo, 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 (Fayyad et al., 1996).

No caso do presente trabalho, tenta-se inferir a possibilidade de ocorrência de eventos convectivos severos a partir das saídas do modelo de previsão numérica de tempo Eta,

as quais fornecem o valor de inúmeras variáveis meteorológicas para cada rodada de previsão de tempo.

Deve-se destacar que o modelo Eta tem um desempenho relativamente bom na América do Sul, essencialmente em situações associadas à presença de sistemas de escala sinótica (Bustamante et al., 2005). Entretanto, para o caso de atividade convectiva em mesoescala, a previsibilidade do modelo Eta (e dos demais modelos meteorológicos em geral) deve ser menor, em função da inexistência de redes meteorológicas de mesoescala e também devido às limitações dos próprios modelos de alta resolução. Nesse contexto, o uso combinado de dados de modelagem e de sensores remotos, aliado a metodologias que possibilitem assimilar inteligentemente esses volumes de dados e realizar previsões alternativas de tempo, como no caso de redes neurais, é de grande interesse.

Um classificador é o programa que atribui uma classe para o conjunto de valores das variáveis meteorológicas geradas pelo modelo meteorológico utilizado, sendo estes valores denominados de atributos. Embora as saídas do modelo refiram-se a previsões de 6 horas, seu passo de tempo de integração é de 15 minutos, tipicamente. 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 as 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 Caetano et al. (2009).

O agrupamento espaço-temporal de ocorrências de descargas elétricas atmosféricas do tipo nuvem-solo foi realizado por meio de uma técnica de análise espacial, a estimação de núcleo, *kernel estimation* (Scott, 1992; Silverman, 1990). 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 1º de latitude por 1º de longitude no território brasileiro (Figura 1), de forma a explorar a localidade espacial dos dados (em contraposição e considerar uma região mais extensa), para eventualmente poder reproduzir padrões específicos de cada mini-região.

A primeira mini-região (A) abrange o Pantanal Sul Matogrossense, a oeste da cidade de Corumbá, a segunda mini-região (B) é delimitada pelas cidades de Bauru e Presidente Prudente, na Alta Sorocabana paulista, enquanto que a terceira (C) fica no Vale do Paraíba, abrangendo São José dos Campos, Taubaté e parte do litoral norte paulista. Essas mini-regiões foram escolhidas por apresentar significativa atividade convectiva durante o período de estudo. Outra razão foi a escolha de mini-regiões que apresentassem um microclima homogêneo (A e B), em confronto com uma mini-região heterogênea (C), que engloba serra, planalto e litoral.

2. DADOS E METODOLOGIA

Os mesmos dados meteorológicos foram empregados nas duas abordagens de mineração de dados. Estes dados referem-se ao modelo de previsão numérica de tempo Eta (Seção 2.1) e à densidade de ocorrências de descargas elétricas atmosféricas (Seção 2.2). Os classificadores, que foram desenvolvidos neste trabalho, seguem a abordagem de similaridade de vetores (Seção 2.3) e a de conjuntos aproximativos (Seção 2.4), além da definição das métricas de desempenho usadas (Seção 2.5).

2.1 Dados do modelo de previsão numérica de tempo Eta

Utilizaram-se neste estudo variáveis meteorológicas geradas pelo modelo regional de previsão de tempo Eta, originalmente desenvolvido pela universidade de Belgrado em conjunto com o Instituto de Hidrometeorologia da Iugoslávia (Messinger et al, 1988; Black, 1994). O modelo Eta vem sendo utilizado operacionalmente pelo Centro de Previsão de Tempo e

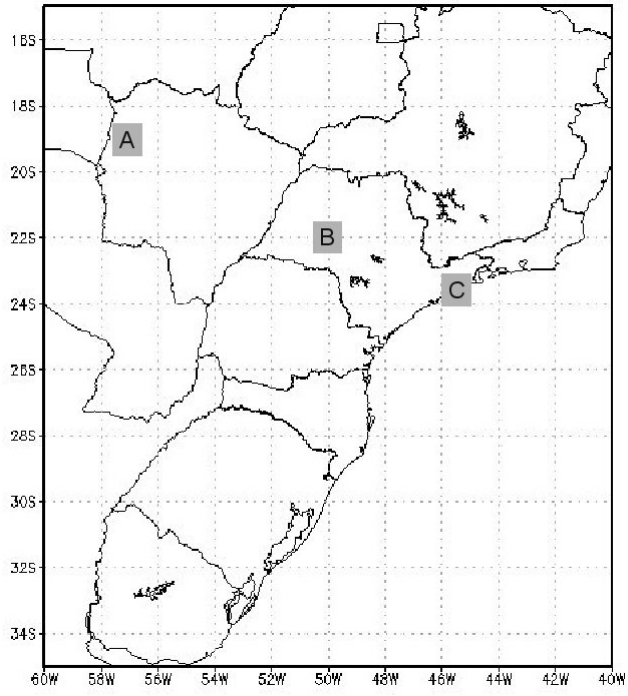

Figura 1- Mini-regiões A, B e C selecionadas para este trabalho.

Estudos Climáticos (CPTEC) do Instituto Nacional de Pesquisas Espaciais (INPE) desde 1996. Nessa versão operacional são geradas duas vezes ao dia previsões que se estendem a até 7 dias sobre grande parte da America do Sul, com resoluções espaciais de 20 a 40 Km. As variáveis prognósticas do modelo são: temperatura do ar, componente zonal e meridional do vento, umidade especifica, água líquida/gelo da nuvem, pressão à superfície e energia cinética turbulenta.

Em função de sua resolução espacial e das características dos processos físicos associados, o modelo Eta é indicado para prever com um maior detalhamento tanto fenômenos meteorológicos de mesoescala (sistemas convectivos, por exemplo) como os de escala sinótica. Este modelo utiliza a grade horizontal E de Arakawa (Arakawa e Lamb, 1977) e a coordenada vertical do modelo é a coordenada η (Mesinger,1984), desenvolvida para minimizar erros significativos de gradientes horizontais de pressão, advecção e difusão horizontal ao longo de uma região com topografia abrupta. O modelo utiliza como condições iniciais e condições de contorno laterais as análises do National Centers for Environmental Prediction (NCEP). Maiores detalhes do modelo Eta utilizado no CPTEC encontram-se em Chou, (1994), Bustamante et al. (2005), Rozante e Cavalcanti. (2006) e Torres e Ferreira (2011).

Os dados binários do modelo de previsão numérica de tempo Eta foram fornecidos pelo CPTEC/INPE, sendo referentes aos meses de janeiro e fevereiro de 2007. Uma análise inicial dos dados por parte de meteorologistas, visando a mineração de dados associada à detecção de eventos convectivos severos, levou a selecionar 17 variáveis deste modelo. A mineração de dados tem um caráter iterativo e qualquer uma de suas etapas pode ser revista em função da qualidade dos resultados obtidos. Assim, testes mais extensos, abrangendo outros meses, serão efetuados e poderão levar a uma revisão deste conjunto de variáveis. Foi realizado um pré-processamento de forma a se obter tabelas em formato texto (ASCII), para 6 meses de dados da primavera de 2006 ao verão de 2007 e 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). Cada registro contém, além da data e horário correspondentes ao intervalo de previsão de 6 hs do modelo Eta, a latitude e longitude correspondentes à grade do modelo, e as 17 variáveis selecionadas. Uma nova análise, efetuada posteriormente, reduziu o número de variáveis a 12, correspondentes a um total de 26 atributos, conforme abaixo. Este conjunto reduzido de variáveis, também poderá ser eventualmente revisto no futuro.

1. pslm: Pressão média ao nível do mar [hPa]
2. pslc: Pressão da superfície [hPa]
3. dp2m: Temperatura do ponto de orvalho a dois metros [K]
4. cape: Energia Potencial convectiva disponível [m^2/s^2]
5. bli: *Best Lifted Index* (a 500 hPa) [K]
6. agpl: Água precipitável instantânea [kg/m^2]
7. v: Vento meridional [m/s] nos níveis 925, 500 e 300 [hPa]
8. u: Vento zonal [m/s] nos níveis 925, 500 e 300 [hPa]
9. z: Altura geopotencial [gpm] nos níveis 1000, 500 e 300 [hPa]
10. tabs: Temperatura absoluta [K] nos níveis 925, 700, 500 e 300 [hPa]
11. omega: Omega [Pa/s] nos níveis 925, 500 e 300 [hPa]
12. umes: Umidade específica [kg/kg] nos níveis 925, 700, 500 e 300 [hPa]

Essas variáveis selecionadas são tipicamente utilizadas de forma direta ou indireta em centros operacionais de previsão de tempo para diagnosticar forçantes térmicas (advecção de temperatura, por exemplo) e forçantes dinâmicas presentes em situações de instabilidade atmosférica. Conforme conveniências da implementação computacional de cada metodologia de mineração adotada, os dados foram portados para o MySQL e também para o MATLAB, para permitir uma seleção em termos de intervalo de tempo e abrangência geográfica.

Os dados usados no presente trabalho são correspondentes aos meses de janeiro e fevereiro de 2007 e às 3 mini-regiões de interesse detalhadas abaixo, cada uma de 1º x 1º :

A. Pantanal Sul Matogrossense: latitudes 18ºS30' a 19ºS30' e longitudes 56ºW30' a 57ºW30'.

B. Alta Sorocabana paulista entre Bauru e Presidente Prudente: latitudes 21ºS30' a 22ºS30' e longitudes 49ºW30' a 50ºW30'.

C. Parte do Vale do Paraíba e Litoral Norte: latitudes 23ºS00' a 24ºS00' e longitudes 45ºW00' a 46ºW00'.

2.2 Densidade de ocorrência de descargas elétricas atmosféricas

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 (Strauss et al., 2010) de forma a gerar os campos de densidade de ocorrência de descargas elétricas atmosféricas para a 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 ferramenta de visualização meteorológica GRADS. Parâmetros específicos podem ser ajustados de forma a correlacionar a densidade com outros dados, objetivando seu uso na mineração de dados meteorológicos. Estes dados de densidade também foram portados para o MySQL e para o MATLAB.

A ferramenta EDDA permite também visualizar animações varrendo os registros de descargas por meio de uma janela deslizante, de forma a acompanhar a evolução temporal de estruturas convectivas. A pronta disponibilidade de dados de descargas permite animações que possibilitam monitorar os eventos meteorológicos com atividade elétrica em tempo quasi-real, contribuindo para que o meteorologista possa ter uma visão instantânea de estruturas convectivas e de sua evolução recente. Assim, a ferramenta proposta tem potencial operacional em meteorologia.

A densidade de ocorrência de descargas, que constitui o atributo de decisão, foi então discretizada em 3 ou 4 faixas com base numa avaliação feita para casos de atividade convectiva severa conhecidos. Na mineração de dados pela abordagem de similaridade de vetores optou-se por 4 faixas, sendo que testes posteriores demonstraram a conveniência de se adotar apenas 3 faixas, correspondentes a atividade convectiva desprezível/ fraca, média e forte. Por outro lado, na abordagem de teoria de conjuntos aproximativos, optou-se por adotar apenas duas faixas, num esquema de binarização (abrangendo o caso positivo apenas densidades altas).

O campo de densidade é calculado por uma técnica de estimação de densidade. A estimação de núcleo (*kernel estimation*), no caso adotando-se uma função gaussiana, permite gerar um campo de densidade de ocorrências, que é suave, e delimitar mais claramente a região de atividade convectiva a partir das descargas, as quais são muito esparsas no espaço e no tempo. A aplicação dessa abordagem foi proposta originalmente em (Politi, 2005; Politi et al., 2006), com o intuito de rastrear a atividade convectiva eletricamente ativa por meio das descargas nuvem-solo, tendo sido originalmente implementada na linguagem de programação MATLAB. Posteriormente, foi desenvolvido o software CAC (Menconi et al., 2010), escrito na linguagem C++ para compilação, com o compilador GNU/gcc, e que gera arquivos de saída referentes à densidade de descargas (ou outros parâmetros) em formato ASCII. Este software faz também a visualização dos campos gerados por meio do pacote GNU/Gnuplot. Este software foi empregado num estudo de correlação entre atividade convectiva e ocorrência de descargas elétricas atmosféricas nuvem-solo (Caetano et al., 2009), e foi inicialmente empregado no escopo dos projetos de mineração de dados meteorológicos citados adiante.

O software CAC mostrou-se útil como ambiente de experimentação, entretanto, um novo software foi desenvolvido visando a operacionalização da estimação da densidade de descargas em tempo quasi-real. Assim, foi desenvolvida a ferramenta EDDA - Estimação de Densidade de Descargas Atmosféricas (Strauss et al., 2010). EDDA é também o nome de um compêndio de contos e poesias da literatura pré-nórdica, que inclui Thor, o deus do trovão.

A estimativa de densidade $f(x)$ no ponto de grade x é obtida por:

$$F(h,(\phi,\lambda)) = \frac{1}{nh^2}\sum_{k=1}^{n} K(y_k) \quad (1)$$

Nessa expressão, n é o total de descargas consideradas, $y_k = d((\phi_i,\lambda_j),X_k)/h$ é uma distância normalizada baseada na distância Euclidiana $d((\phi_i, \lambda_j), X_k)$ do ponto de ocorrência X_k ao ponto de grade (ϕ_i, λ_j). O parâmetro h é usado no cálculo das distâncias normalizadas e define a vizinhança circular das ocorrências que mais influem o estimador para cada ponto de grade. Uma vez que h pondera as distâncias das ocorrências no cálculo da superfície expressa por $F(h,(\phi, \lambda))$, ele atua como um *parâmetro suavizador*. Quanto maior o h, mais suave será a superfície gerada e vice-versa. Além do mais, ele define o tamanho médio dos campos gerados. Finalmente, a função de núcleo $K(y_k)$ adotada é a gaussiana bidimensional. A distância normalizada foi corrigida a partir das diferenças de latitude e longitude do ponto de ocorrência ao ponto de grade considerado, para levar em conta a curvatura da Terra. Finalmente, é possível calcular um valor ótimo para o parâmetro h que minimize o *mean integrated standard error*, para cada ponto da grade, com base no desvio padrão σ das distâncias das ocorrências ao ponto considerado (Scott, 1992; Strauss et al., 2010), dado por:

$$h(\phi_i,\lambda_j) = n^{-1/6}\sigma(\phi_i,\lambda_j) \quad (2)$$

2.3 Abordagem de similaridade de vetores

O presente classificador, baseado na abordagem de similaridade de vetores, foi desenvolvido admitindo-se a hipótese de que "seja viável formar os *clusters* de vetores de atributos correspondentes a variáveis meteorológicas, que servirão como bases de representação para os diversos níveis (ou classes) de atividade convectiva, adotando-se como critério de decisão apenas o valor da densidade de ocorrência de descargas elétricas atmosféricas, atribuído ao vetor."

A primeira etapa do projeto consistiu em fazer a escolha dos subconjuntos de dados que seriam utilizados para treinar o classificador e dos subconjuntos de dados para posteriormente testá-lo. Conforme mencionado na Seção 2, foram selecionadas 3 mini-regiões de 1° x 1°, obtendo-se assim 3 conjuntos de dados correspondentes, cada um com 8496 vetores, sendo cada um dos quais divididos em dois subconjuntos: o primeiro servindo como base de treinamento e o outro como base de teste para o classificador. Os vetores das bases de dados destinadas ao treinamento do classificador foram agrupados em *clusters* de acordo com a discretização dos valores de densidades de descargas inicialmente adotada, que considerava os seguintes 4 níveis: densidade desprezível (ou nula), fraca, média e forte.

Considerando os vetores das 3 mini-regiões selecionadas, os correspondentes valores das densidades de descarga variaram entre 0 e $\approx$ 0,1900. Por sugestão dos meteorologistas, o limiar inferior de densidade de descarga, delimitando os casos de atividade convectiva desprezível, foi tomado como 0,0001. Esse valor foi determinado com base em comparações entre as imagens de densidade de descargas e as correspondentes imagens do satélite GOES-10, pseudocoloridas em função das temperaturas de topo de nuvens para alguns casos significativos de eventos convectivos que foram analisados pelo CPTEC. Embora esta discretização das densidades de descargas em 4 níveis tenha sido adotada inicialmente, no decorrer dos trabalhos, analisando-se 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 um dos 26 atributos era calculada sua variação máxima na base de treinamento e, para cada atributo 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 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:

$$\frac{abs\,(v_A^j - v_B^j)}{a_j} \leq d_j \quad \text{para todo } j = 1, \ldots, 26 \qquad (3)$$

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 do j-ésimo atributo 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 da Equação 3. 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 atributo do vetor a ser classificado, e considerando-se a distribuição destes valores nos *clusters* representativos de cada classe. Denominam-se estas matrizes de 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.

Seja $V = [v^1, v^2, \ldots, v^{25}, v^{26}]$ um vetor a ser classificado como pertencente a um dos 3 *clusters*. Para cada atributo 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 seu j-ésimo atributo na vizinhança $v^j \pm d(j)$, onde d(j) corresponde ao desvio padrão desse atributo, 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.

Passo 3: consideram-se os demais atributos 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 passo 1) têm também o valor de seu i-ésimo atributo na vizinhança $v^i \pm d(i)$. Resultando nas quantidades $[q_1^{j1},\ldots, q_1^{j26}]$, $[q_2^{j1}, \ldots, q_2^{j26}]$ e $[q_3^{j1}, \ldots, q_3^{j26}]$.

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

A Figura 2 abaixo mostra as matrizes M_1, M_2 e M_3 com as j-ésimas colunas preenchidas, relativamente ao Passo 4.

Passo 5: 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 atributos correspondentes a valores numéricos de variáveis meteorológicas considerando-o como um conjunto de eventos discretos (o valor de cada atributo 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, de modo geral, 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 atributo do vetor a ser classificado). São estimadas probabilidades para cada atributo individualmente, depois para todas as combinações de atributos tomados dois a dois 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 (Duda et al., 2000), 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 os atributos individualmente e em grupos de dois) é preferível a uma comparação simultânea dos 26 atributos, porque em sistemas de decisão multi-atributos, as correlações entre cada atributo

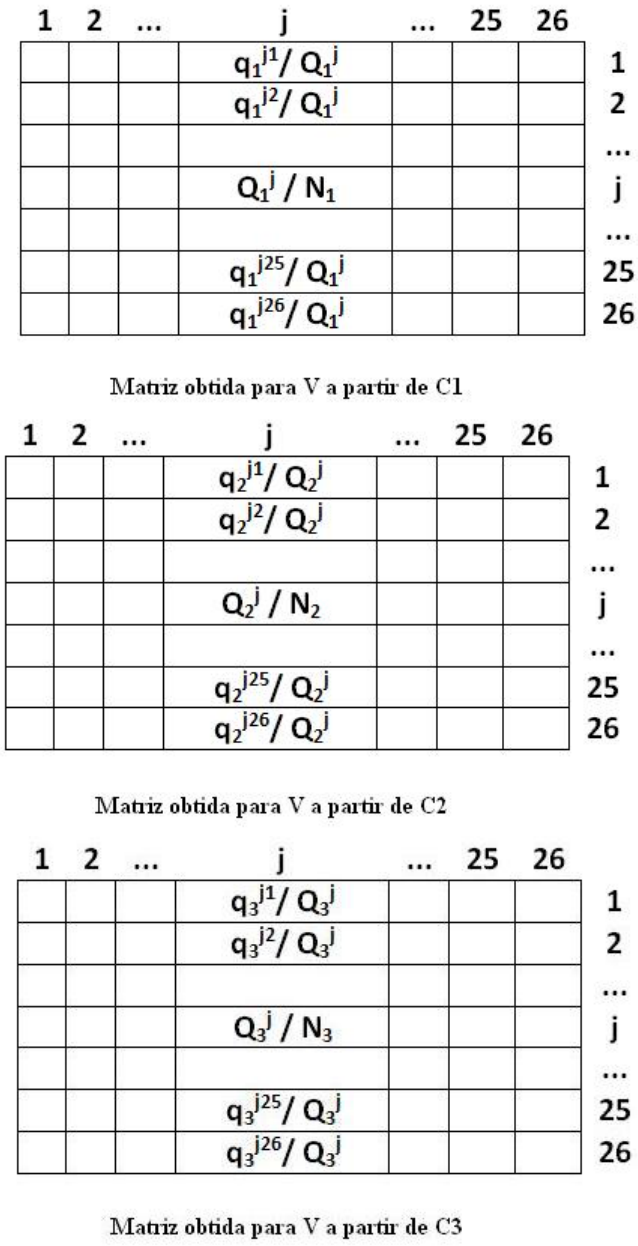

Figura 2 - Construção de matrizes M_1, M_2 e M_3 de probabilidades estimadas com base em freqüências de ocorrência dos valores de um vetor V a ser classificado.

de informação e o atributo de decisão nem sempre são iguais. Esse esquema inédito foi assim proposto visando alcançar um melhor desempenho de classificação, conforme acabou sendo demonstrado pelos resultados apresentados.

2.4 Abordagem de conjuntos aproximativos

No mundo real as informações são freqüentemente incertas, imprecisas ou incompletas, devido a restrições de observação, medição ou mesmo devido a limitações relativas à resolução espacial e temporal. Diversas teorias foram desenvolvidas para tratar tais imperfeições, tais como a teoria dos conjuntos nebulosos, teoria de Dempster-Shafer e a teoria das possibilidades. No início da década de 80 surgiu a Teoria dos Conjuntos Aproximativos (TCA), do inglês *Rough Set Theory* (Pawlak, 1982), a qual se caracterizada pela simplicidade e bom formalismo matemático, o que facilita a manipulação de informações, e em especial, as informações incertas. A TCA é uma extensão da teoria dos conjuntos, que enfoca o tratamento de incerteza dos dados por meio de uma relação de indiscernibilidade e que avalia se os elementos de um conjunto são indiscerníveis, ou seja, se possuem as mesmas propriedades, segundo Leibniz (Scuderi, 2003). 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.

Na TCA usa-se o conceito de sistema de informação, que constitui um conjunto particular de dados, representado por uma tabela onde cada linha representa um objeto, correspondente a um caso, evento ou paciente, e cada coluna representa um atributo, correspondente a uma variável, observação ou propriedade. Formalmente, um sistema de informação é definido por um par $S = (U; A)$, onde U é um conjunto finito não-vazio de objetos chamado de universo e A é um conjunto finito não-vazio de atributos, tal que $a : U \rightarrow V_a$ para todo $a \in A$. O conjunto V_a é chamado de conjuntos de valores do atributo a. Em muitas aplicações, para fim classificatório, certo atributo é distinguido dos demais, sendo denominado atributo de decisão. Os sistemas de informação deste tipo são chamados sistemas de decisão (SD), apresentando a forma $S = (U; A \cup \{d\})$, onde $d \notin A$ é o atributo de decisão. Os elementos do conjunto A são chamados atributos condicionais ou simplesmente condições (Komorowski et al., 1999).

A tabela correspondente a um SD pode ser desnecessariamente grande devido à repetição de elementos "iguais" ou devido a atributos supérfluos. Essa repetição pode ser eliminada por meio da noção de indiscernibilidade, inerente à TCA, implementada por uma relação de equivalência que faz com que apenas um objeto represente toda uma classe, ou seja, apenas um registro é mantido na tabela correspondente ao SD. Complementarmente, pode-se reduzir os atributos supérfluos por meio de reduções, conforme descrito adiante.

Dado $S = (U; A)$ como sistema de informação, então para qualquer subconjunto de atributos pode-se definir uma relação de equivalência $IND_A(B)$

$$IND_A(B) = \{(x, x' \in U) \mid \forall a \in B, a(x) = a(x')\} \quad (4)$$

Essa relação é chamada de relação de *B-indiscernibilidade*, que particiona A em classes de equivalência, sendo cada classe

denotada por $[x]_B$ (Komorowski et al., 1999). Assim, uma relação de equivalência agrupa elementos indiscerníveis em classes, particionando o universo de objetos para um determinado conjunto de atributos de decisão.

As partições, originalmente obtidas para o sistema de informação considerado, permitem uma classificação aproximada dos objetos, uma vez que, para o conjunto de atributos selecionados para a relação de equivalência, essa classificação não é exata. Ou seja, um dado objeto pode ser classificado como pertencente a uma dada classe, mas possuir um ou mais atributos que não são característicos dessa classe. Essa imprecisão levou à TCA, que visa estabelecer uma relação de pertinência não determinística de cada objeto em relação a uma dada classe. Dessa forma, na TCA, em função das classes definidas por uma relação de equivalência qualquer para um dado sistema de informação (por exemplo, classes X_1, X_2 e X_3), um novo objeto pode ser classificado, em particular em relação à classe X_1, como: (i) pertencente à classe X_1 sem que haja ambiguidade no sentido de pertencer a X_2 ou X_3, (ii) pertencente a mais de uma classe (por exemplo, X_1 e X_2) ou (iii) não pertencente à classe X_1. Se existem elementos que não podem ser definidos como pertencentes a uma só classe, o conjunto é dito *aproximativo* (Pawlak, 1982).

Assim, a partir das classes definidas por uma relação de equivalência qualquer aplicada a um dado sistema de informação $S = (U; A)$, e dado um subconjunto de atributos e um subconjunto de objetos, pode-se construir uma aproximação de X usando somente as informações contidas no conjunto de atributos B construindo as aproximações B-inferiores e B-superiores de X, denotados respectivamente $\underline{B}X$ e $\overline{B}X$, onde:

$$\underline{B}X = \{x \mid [x]_B \subseteq X\} \quad (5)$$

$$\overline{B}X = \{x \mid [x]_B \cap X \neq \varnothing\} \quad (6)$$

Os objetos em $\underline{B}X$ podem ser certamente classificados como membros de X na base de conhecimento (conjunto de atributos) B, enquanto os objetos em $\overline{B}X$ podem somente serem classificados como possíveis membros de X na base de conhecimento B. O conjunto $F_B(X) = \overline{B}X - \underline{B}X$ é chamado de *região de fronteira* de X, sendo que estes objetos não podem ser classificados pertencentes a X na base de conhecimento B com absoluta certeza. O conjunto $E_B(X) = U - \overline{B}X$ é então chamado de *região externa* de X, e estes objetos podem ser classificados como não pertencentes a X. Assim, um conjunto é dito *aproximativo* se a região da fronteira não é vazia, ou caso contrário, *conjunto preciso*.

Além da compactação da base de dados baseada na indiscernibilidade, que permite representar elementos iguais com um só registro, outro artifício é o uso de subconjuntos de atributos chamados *reduções*, de forma a manter somente os atributos que preservam a relação de indiscernibilidade, eliminando os supérfluos. Normalmente existem vários subconjuntos de atributos que possuem esta característica e os que são mínimos são chamados de *reduções*. A determinação das *reduções* é um problema "*NP-hard*" (Komorowski et al., 1999).

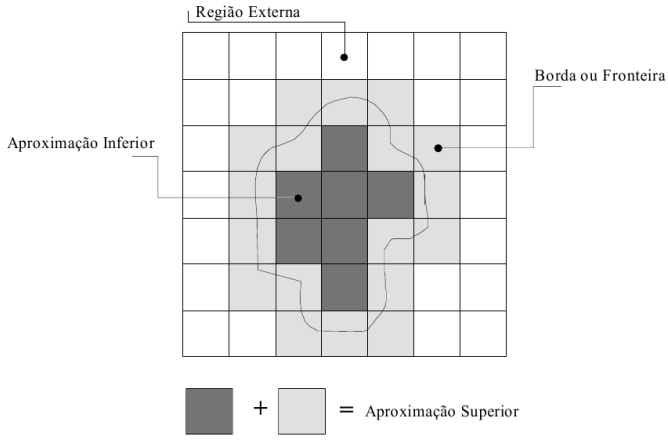

Figura 3 - Aproximações de um conjunto na TCA

Dado o sistema de informação $S = (U; A)$ uma redução de S é um conjunto mínimo de atributos, tal que $IND_S(B) = IND_S(A)$. Em outras palavras, uma redução, $RED(B)$, é o conjunto mínimo de atributos de A, que preserva o particionamento do universo realizado pela relação de indiscernibilidade, e conseqüentemente permite executar classificações equivalentes àquelas permitidas pelo conjunto de atributos original (completo).

Dado um sistema de informação S com n objetos, define-se a correspondente matriz de discernibilidade como sendo uma matriz simétrica de dimensão n x n para a qual, cada entrada c_{ij} para (i ≠ j) consiste em um subconjunto de atributos que difere os objetos x_i e x_j, sendo os elementos da diagonal nulos, conforme mostra a equação abaixo:

$$c_{ij} = \{a \in A \mid a(x_i) \neq a(x_j)\} \quad \text{para } i,j = 1, ..., n \text{ e } i \neq j \quad (7)$$

A função de discernibilidade f_A para S é uma função de m variáveis booleanas correspondentes aos atributos $a_1,..., a_m$ e é dada pela seguinte equação:

$$f_A(a_1^*,...a_n^*) = \wedge\{\vee \ c_{ij}^* \mid 1 \leq j \leq i \leq n, \ \ c_{ij} \neq \varnothing\} \quad (8)$$

onde $c_{ij}^* = \{a^* \mid a \in c_{ij}\}$.

O conjunto de atributos determinados pela simplificação booleana de f_A, produz o conjunto de reduções de A. Cada linha da função de discernibilidade corresponde a uma coluna da matriz de discernibilidade.

Quando são calculadas as reduções para toda a matriz de discernibilidade e, conseqüentemente, simplificada toda a função de discernibilidade, tem-se as chamadas *reduções*

completas. Por outro lado, pode-se calcular as reduções de um elemento em relação ao resto, de uma classe em relação às outras, etc. A estas reduções dá-se o nome de *reduções k-relativas,* e em geral, constituem a simplificação de uma parte da função de discernibilidade. Por exemplo, como cada linha de f_A corresponde a uma coluna matriz de discernibilidade e cada coluna é relativa a um elemento, então simplificando apenas uma linha da função, relativa a um elemento *x,* tem-se as reduções deste elemento em relação a todos os outros de *U*.

A mineração de dados pela abordagem da TCA proposta levou à implementação de um classificador que empregou os dados meteorológicos, no caso, do modelo de previsão numérica de tempo Eta, como atributos de informação, e dados de descargas elétricas, cuja densidade foi associada à ocorrência de atividade convectiva severa, constituindo o atributo de decisão. Os softwares usados nessa última abordagem estão listados a seguir, não havendo nenhum software proprietário:

A. MySQL (http://www.mysql.com/) – MySQL *Structured Query Language, da Sun Mycrosystems (EUA)*.

B. Weka (www.cs.waikato.ac.nz/ml/weka/) – *Data Mining Software in Java*, da Universidade de Waikato (Nova Zelândia);

C. ROSETTA (http://www.lcb.uu.se/tools/rosetta/) - *Rough Set Toolkit for Analysis of Data*, da Universidade de Uppsala (Suécia);

D. RSES (http://www.univ.rzeszow.pl/eng/inst_math.php) – *Rough Set Exploration System*, da Universidade de Rzeszów (Polônia).

E. 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 ferramenta 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.

Inicialmente, haviam sido feitos testes usando uma árvore de decisão J48, o qual é uma evolução do algoritmo C4.5 (Quinlan, 1993) , mas o desempenho baixo desse classificador levou ao classificador baseado na TCA, que demonstrou maior robustez.

2.5 Avaliação de desempenho de um classificador

Na avaliação de desempenho de um sistema de classificação, o uso da probabilidade de erro é tipicamente considerada insuficiente. Uma importante métrica para avaliar o desempenho do classificador é verificar quantitativamente o número de instâncias classificadas corretamente ou não para cada classe. Esses números permitem montar a matriz de confusão $\boldsymbol{MC} = [MC(i,j)]$, definida como sendo uma matriz n x n (sendo n o número de classes), cujos elementos $MC(i,j)$, com i, $j = 1,...n$, expressam a quantidade de instâncias que são da classe i e foram classificados como sendo da classe j (Theodoridis e Koutroumbas, 2006), de tal forma que os elementos da diagonal expressam as instâncias classificadas corretamente e a soma de cada coluna indica o total de instâncias de cada classe. Assim, a matriz de confusão para duas classes ($n=2$), exemplificadas como de instâncias positivas e negativas seria esquematizada conforme abaixo:

$$\mathbf{MC} = \begin{bmatrix} VN & FP \\ FN & VP \end{bmatrix} \quad (9)$$

Sendo:

- VN = verdadeiros negativos (classificados corretamente como negativos);
- VP = verdadeiros positivos (classificados corretamente como positivos);
- FN = falsos negativos (classificados incorretamente como sendo negativos);
- FP = falsos positivos (classificados incorretamente como sendo positivos).

Uma métrica que pode ser extraída facilmente desta matriz é a acurácia (*acc*) ou precisão global, que é a soma dos acertos do classificador dividida pelo total de elementos envolvidos na avaliação e é dado pela expressão abaixo, ou seja, a soma dos elementos da diagonal (acertos) dividida pela soma de todos os elementos da matriz:

$$acc = \frac{\sum_{i=1}^{n} MC(i,i)}{\sum_{i=1}^{n}\sum_{j=1}^{n} MC(i,j)} \quad (10)$$

Outra métrica bastante empregada em aprendizado supervisionado é o índice de concordância *kappa*, ou simplesmente κ, que mensura a concordância entre o real ou verdade e o estimado por um classificador com base na própria matriz de confusão. O índice *kappa* (κ) é dado por:

$$\kappa = \frac{acc - pae}{1 - pae} \quad (11)$$

onde, *pae* é a proporção de acertos esperados, definida a seguir (a acurácia constitui a proporção de certos observados):

$$pae = \frac{\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{n} MC(j,i)\right)\left(\sum_{j=1}^{n} MC(i,j)\right)\right]}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} MC(i,j)\right)^{2}} \qquad (12)$$

3. RESULTADOS

Apresentam-se a seguir os resultados obtidos pelos classificadores para dados dos meses de janeiro de fevereiro de 2007 para as 3 mini-regiões consideradas.

3.1 Resultados do classificador baseado em similaridade de vetores

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 0,0001 e 0,1900. Para os vetores das 3 mini-regiões selecionadas, os melhores limiares encontrados foram 0,0025 e 0,0100. Estes valores foram então adotados para montar os *clusters* representativos das 3 classes correspondentes à atividade convectiva desprezível/fraca, média e forte. Abaixo, a título de exemplo, mostra-se como ficou a distribuição dos 8496 vetores de uma das mini-regiões de 1º x 1º grau nos 3 *clusters*:

C_1: 4568 (desprezível) + 2891 (fraca) = 7459 vetores (88,17% do total)

C_2: 802 vetores (9,48% do total)

C_3: 199 vetores (2,35% do total)

O forte desbalanceamento entre C_1 e as outras duas classes visto no caso exemplo dado acima, foi verificado também nas demais mini-regiões. Para reduzir eventuais efeitos nocivos, devidos a este desbalanceamento, no desempenho do classificador, optou-se por se dividir o *cluster* C_1 em 5 *sub-clusters*, cada um deles contendo 20% dos exemplares de C_1. Na amostragem feita para gerar esses 5 subconjuntos, foram mantidas as mesmas proporções entre número de vetores por sub-faixa de densidade de descargas e o número total de vetores que havia no *cluster* C_1 original. Assim, considerando novamente o caso apresentado como exemplo, cada um dos 5 *sub-clusters* representativos da classe C_1 ficou com a seguinte quantidade de vetores:

C_1: 914 (desprezível) + 578 (fraca) = 1492 vetores (17,64% do total)

Os testes foram realizados segundo um esquema de *cross-validation* com 10 *folds,* no qual 9 *folds* (90% dos dados) foram utilizados para treinamento e 1 *fold* (10% dos dados) foi utilizado para teste. Considerando que os dados de cada mini-região foram divididos em 5 diferentes conjuntos (cada um deles formado por um dos 5 *sub-clusters* de C_1 mais os *clusters* C_2 e C_3), os resultados finais da classificação para cada mini-região foram calculados como a média dos resultados parciais obtidos nos 50 testes realizados.

Os resultados para as 3 mini-regiões são apresentados nas Tabelas 1, 2, e 3 e correspondem às matrizes de confusão obtidas com o esquema de classificação proposto. As quantidades totais de novos vetores (os vetores de teste), de cada classe, que foram submetidos à classificação pelo esquema proposto, são iguais as somas das respectivas linhas.

A Tabela 1 mostra a matriz de confusão resultante quando o classificador proposto, depois de treinado com os vetores de atributos correspondentes às variáveis meteorológicas da mini-região do Pantanal Sul Matogrossense (A), foi testado com novos vetores desta mesma mini-região.

A Tabela 2 mostra a matriz de confusão resultante quando o classificador proposto, depois de treinado com os vetores de atributos correspondentes às variáveis meteorológicas da mini-região da Alta Sorocabana paulista (B), foi testado com novos vetores desta mesma mini-região.

A Tabela 3 mostra a matriz de confusão resultante quando o classificador proposto, depois de treinado com os vetores de atributos correspondentes às variáveis meteorológicas da mini-região do Vale do Paraíba paulista (C), foi testado com novos vetores desta mesma mini-região.

Tabela 1 - Matriz de confusão para o teste da mini-região A, do Pantanal Sul Matogrossense (abordagem de similaridade de vetores).

	Desprezível/Fraco	Médio	Forte
Desprezível/Fraco	**142**	**6**	**1**
Médio	**1**	**75**	**4**
Forte	**0**	**1**	**19**

Tabela 2 - Matriz de confusão para o teste da mini-região B, da Alta Sorocabana paulista (abordagem de similaridade de vetores).

	Desprezível/Fraco	Médio	Forte
Desprezível/Fraco	**130**	**6**	**2**
Médio	**3**	**76**	**11**
Forte	**0**	**7**	**61**

Tabela 3 - Matriz de confusão para o teste da mini-região C, do Vale do Paraíba Paulista (abordagem de similaridade de vetores).

	Desprezível/Fraco	Médio	Forte
Desprezível/Fraco	**138**	**5**	**3**
Médio	**1**	**54**	**9**
Forte	**0**	**3**	**45**

Tabela 4 - Índices de avaliação de desempenho da classificação para as 3 mini-regiões (abordagem de similaridade de vetores).

Mini-região	Índice Kappa	Acurácia
Pantanal Sul Matogrossense (A)	**0,90**	**0,95**
Alta Sorocabana paulista (B)	**0,85**	**0,90**
Vale do Paraíba e Litoral Norte paulista (C)	**0,86**	**0,92**

Finalmente, na Tabela 4 são apresentados os índices de desempenho (índice Kappa e acurácia) do sistema de classificação proposto referentes aos resultados das Tabelas 1, 2 e 3.

Os índices de desempenho apresentados na Tabela 4, todos acima de 0.84, demonstram que o esquema de classificação proposto constitui-se numa interessante alternativa para a tarefa de identificar, a partir de vetores de atributos correspondentes às variáveis meteorológicas, condições atmosféricas favoráveis à ocorrência de um determinado nível de atividade convectiva. É preciso registrar, no entanto, que o esquema de classificação proposto foi ainda submetido a outro tipo de teste, no qual os dados de treinamento e os dados novos para validação foram selecionados em diferentes mini-regiões de 1° x 1° e que, nestes testes, o classificador não proporcionou bons resultados.

3.2 Resultados do classificador baseado na teoria de conjuntos aproximativos

As análises para as três mini-regiões supramencionadas, denominadas A, B e C (Pantanal Sul Matogrossense, Alta Sorocabana Paulista e parte do Vale do Paraíba e Litoral Norte, respectivamente), foram efetuadas com o software ROSETTA, no esquema de amostragem de validação cruzada (*cross-validation*) com 10 partições, e de *holdout* com divisão de 80% das instâncias para treinamento e as demais 20% para validação do classificador. Cada instância representa o estado da atmosfera, dado por um vetor de atributos correspondentes às variáveis meteorológicas do modelo Eta, para um dado ponto de grade e para uma dada saída do próprio modelo. O tamanho da amostra é de 5900 instâncias para A e B, e 8496 instâncias C devido ao ajuste da região selecionada à grade do modelo. Para as 3 mini-regiões consideradas, foram adotados os mesmos limiares da abordagem descritos na seção 3.a, ou seja, 0,0025 e 0,0100 de forma a se obter 3 classes.

A matriz de confusão mostrada na Tabela 5 é referente ao teste do classificador produzido pelo esquema de validação cruzada para a mini-região A (Pantanal Sul Matogrossense). Vale lembrar que os resultados apresentados para este tipo de amostragem são resultados médios, uma vez que a classificação e validação (treino e teste) são computadas 10 vezes. A acurácia média é de 0,95, que é a melhor acurácia dentre as três mini-regiões e índice Kappa 0,81.

Os resultados obtidos para a validação cruzada da mini-região B (Alta Sorocabana paulista) podem ser conferidos na Tabela 6. Nesta mini-região foi obtido o melhor índice Kappa, embora seja ligeiramente superior às outras, com valor de 0,82 e acurácia de 0,92.

Tabela 5 - Matriz de confusão para o teste da mini-região A, do Pantanal Sul Matogrossense (abordagem da Teoria de Conjuntos Aproximativos com validação cruzada).

	Desprezível/Fraco	Médio	Forte
Desprezível/Fraco	**501**	**3**	**0**
Médio	**13**	**41**	**1**
Forte	**4**	**1**	**17**

Tabela 6 - Matriz de confusão para o teste da mini-região B, da Alta Sorocabana paulista (abordagem da Teoria de Conjuntos Aproximativos com validação cruzada).

	Desprezível/Fraco	Médio	Forte
Desprezível/Fraco	**465**	**2**	**0**
Médio	**14**	**42**	**5**
Forte	**4**	**5**	**42**

Tabela 7 - Matriz de confusão para o teste da mini-região C, do Vale do Paraíba e Litoral Norte Paulista (abordagem da Teoria de Conjuntos Aproximativos com validação cruzada).

	Desprezível/Fraco	Médio	Forte
Desprezível/Fraco	**702**	**3**	**0**
Médio	**21**	**45**	**4**
Forte	**7**	**3**	**47**

Para o esquema de validação cruzada, a matriz de confusão da mini-região C (Vale do Paraíba e Litoral Norte paulista) é exibida na Tabela 7. Para esta região a acurácia média é de 0,94 e índice Kappa de 0,80.

Levando em conta os índices que mensuram o desempenho do classificador, que podem ser avaliados na Tabela 8, os resultados são equivalentes para as mini-regiões analisadas. A acurácia média está acima de 90% para todas as mini-regiões e o índice Kappa igual ou superior a 0,80. Estes resultados indicam excelente habilidade discriminatória do classificador, expressado pela acurácia, e a ótima concordância do classificador com o real (ou desejado), dado pelo índice Kappa.

Os resultados da validação cruzada serviram para validar o emprego da metodologia da TCA na modelagem de eventos severos, utilizando a densidade de ocorrência de descargas elétricas atmosféricas como atributo de decisão, e dados do modelo Eta como atributos condicionais, ou seja, mediante aos resultados conseguidos se torna possível a construção de um classificador para a criação de um modelo baseado em regras para os dados.

Então, frente aos resultados otimistas da fase de validação cruzada, o próximo passo deste estudo foi à constituição dos classificadores para as três mini-regiões utilizando o esquema de amostragem *holdout*, cujas matrizes de confusão resultantes são mostradas a seguir, sendo o total de elementos de teste de 1180 para as mini-regiões A e B e 1699, para a mini-região C.

Neste primeiro caso, a matriz de confusão mostrada na Tabela 9, é referente às reduções da mini-região do Pantanal Sul Matogrossense (A). O conjunto de regras induzido a partir das reduções encontradas possui cerca de 16000 regras. A acurácia calculada para a matriz de confusão da região A é de 0,99 e o índice Kappa de 0,85.

O segundo caso refere-se à mini-região da Alta Sorocabana Paulista (B). Diferentemente da mini-região anterior, os resultados para esta mini-região são ligeiramente inferiores em termos de acurácia e Kappa, porém esses índices ainda estão em patamares considerados excelentes para classificação. Na matriz de confusão (Tabela 10) produzida pelas reduções, tem acurácia 0,94 e Kappa 0,84. O conjunto de regras produzido tem aproximadamente 18800 elementos.

A seguir, apresentam-se os resultados referentes à mini-região do Vale do Paraíba e Litoral Norte paulista (C) na Tabela 11. Semelhantemente aos resultados exibidos pela mini-região B, essa região possui índices parecidos, sendo 0,95 e 0,85 de acurácia e Kappa, respectivamente, com um número elevado de regras, aproximadamente 27500.

Tabela 8 - Índices médios de avaliação de desempenho da classificação para as 3 mini-regiões (abordagem da Teoria de Conjuntos Aproximativos com validação cruzada).

Mini-região	Índice Kappa	Acurácia
Pantanal Sul Matogrossense (A)	0,81	0,95
Alta Sorocabana Paulista (B)	0,82	0,93
Vale do Paraíba e Litoral Norte paulista (C)	0,80	0,94

Tabela 9 - Matriz de confusão para o teste da mini-região A, do Pantanal Sul Matogrossense (abordagem da Teoria de Conjuntos Aproximativos com *holdout*).

	Desprezível/Fraco	Médio	Forte
Desprezível/Fraco	996	7	0
Médio	20	95	7
Forte	4	5	41

Tabela 10 - Matriz de confusão para o teste da mini-região B, da Alta Sorocabana paulista. (abordagem da Teoria de Conjuntos Aproximativos com *holdout*).

	Desprezível/Fraco	Médio	Forte
Desprezível/Fraco	933	5	3
Médio	16	96	15
Forte	7	14	77

Claramente o volume de regras criadas, nesta mini-região C, indica uma maior complexidade no micro clima em questão, pois mais regras são necessárias para descrever o comportamento da atmosfera, evidentemente quando comparado com as outras mini-regiões analisadas.

Finalmente, na Tabela 12 é mostrado um sumário dos resultados apresentados anteriormente. É possível notar que a região A possui o melhor resultado. Entretanto, todos os resultados são considerados excelentes para construção de classificadores, pois tanto a acurácia, quando o índice Kappa, possuem valores relevantes, mediante os dados utilizados para treinamento e validação do modelo criado.

4. CONCLUSÕES

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 de previsão numérica de tempo 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. Dois classificadores, o primeiro baseado na abordagem de Similaridade de Vetores e o segundo, na Teoria de Conjuntos Aproximativos foram desenvolvidos e testados. Os resultados foram expressivos, mostrando que a abordagem proposta pode ser viável para a previsão de ocorrências de eventos convectivos severos a partir das previsões de tempo do modelo meteorológico Eta. Isso pode demonstrar que existe um uso potencial deste modelo para tal tipo de previsão, mesmo considerando-se eventuais discrepâncias com os dados observados (Farias e Chan, 2006), ou seja, entre a previsão e a realidade.

A pesquisa apresentada continuará dividindo partes do território brasileiro numa grade de mini-regiões de tamanho ótimo ainda sendo avaliado, possivelmente de 2º x 2º, uma vez que tamanhos maiores podem levar a imprecisões devidas às diferenças de relevo, vegetação, etc. entre áreas consideradas na mesma mini-região, enquanto que tamanhos menores implicariam num volume de processamento muito alto. Novos dados meteorológicos referentes a novos períodos serão adquiridos do CPTEC/INPE, de forma a incorporar efeitos das estações do ano e de sazonalidade. O objetivo é obter um conjunto de classificadores adequados para diferentes regiões do Brasil e para diferentes contextos temporais. Os

Tabela 11 - Matriz de confusão para o teste da mini-região C, do Vale do Paraíba e Litoral Norte paulista (abordagem da Teoria de Conjuntos Aproximativos com *holdout*).

	Desprezível/Fraco	Médio	Forte
Desprezível/Fraco	**1400**	**15**	**3**
Médio	**23**	**103**	**13**
Forte	**8**	**10**	**103**

Tabela 12 - Índices de avaliação de desempenho da classificação para as 3 mini-regiões (abordagem da Teoria de Conjuntos Aproximativos com *holdout*).

Mini-região	Índice Kappa	Acurácia
Pantanal Sul Matogrossense (A)	**0,85**	**0,99**
Alta Sorocabana paulista (B)	**0,84**	**0,94**
Vale do Paraíba e Litoral Norte paulista (C)	**0,85**	**0,95**

resultados obtidos com estes novos conjuntos de dados poderão eventualmente levar a rever o conjunto utilizado de variáveis meteorológicas do modelo Eta, eliminado algumas variáveis ou acrescentando outras. A própria metodologia de mineração empregada nas duas abordagens poderá ser revista de forma a se obter classificadores mais robustos, para no futuro testá-los e validá-los em ambiente operacional, com o objetivo de se obter uma ferramenta de auxílio à previsão meteorológica.

5. 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". Agradecem também ao CPTEC/INPE pelos dados de descargas elétricas atmosféricas do sistema RINDAT. Finalmente, os demais autores agradecem, *in memoriam,* ao autor José Demísio Simões da Silva pela sua contribuição fundamental ao presente trabalho, pela liderança e companheirismo.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ARAKAWA, A.; LAMB, V. R. Computational design of the basic dynamical processes of the UCLA general circulation model. **Methods in Computational Physics**, v.17, p.173-265, 1977.

BLACK, T. L. The new NMC mesoscale Eta model: description and forecast examples. **Weather and Forecasting**, v.9, n.2, p.265-278, 1994.

BUSTAMANTE, J. F.; CHOU, S. C.; ROZANTE, J. R.; GOMES, J. L. Uma avaliacao da previsibilidade de tempo do Modelo ETA para a America do Sul. **Revista Brasileira de Meteorologia**, v. 20, n. 1, p. 59-70, Abr. 2005. (INPE-12500-PRE/7802). Disponível em: <http://urlib.net/sid.inpe.br/iris@1915/2005/05.12.17.41>. Acesso em: 07 jun. 2011.

CAETANO, M.; ESCOBAR, G.C.J.; STEPHANY, S.; MENCONI, V.E.; FERREIRA, N.J.; DOMINGUES, M. O.; MENDES JUNIOR, O. Visualização de campo de densidade de ocorrências de descargas elétricas atmosféricas como ferramenta auxiliar no *nowcasting*. In: Xiii Latin American and Iberian Congress on Meteorology (CLIMET XIII) and X Argentine Congress on Meteorology (CONGREMET X). **Proceedings**. Buenos Aires, 2009.

CHOU, S.C. Modelo Regional Eta. **Climanálise – Boletim de Monitoramento e Análise Climática**, v.1, n.especial 10 anos, p. on line, 1996.(INPE-12512-PRE/7814).

DUDA, R.O.; HART, P.E.; STORK, D.G. **Pattern Classification**. New York: John Wiley & Sons, 2000. 654p.

FARIAS, S. E. M.; CHAN, C. S. Simulação das características microclimatológicas para o Pantanal Sul-mato-grossense. In: PRIMEIRO SIMPÓSIO DE GEOTECNOLOGIAS NO PANTANAL, EMBRAPA INFORMÁTICA AGROPECUÁRIA/INPE. **Anais**. Campo Grande, 2006.

FAYYAD, U. M.; PIATETSKY-S, G.; SMYTH, P. From data mining to knowledge discovery: an overview. In: FAYYAD, U. M.; PIATETSKY-S, G.; SMYTH, P; UTHURUSAMY, R.

(ed.). **Advances in knowledge discovery and data mining.** Menlo Park-CA: AAAI Press, 1996. Cap. 1, p. 1-34.

KOMOROWSKI, J.; POLKOWISKI, L.; SKOWRON, A. Rough sets: a tutorial. In: PAL, S.K.; SKOWRON, A. (ed.). **Rough fuzzy hybridization: A new trend in decision-making**. Singapore : Springer-Verlag, 1999. Cap. 1, p. 3-98.

MENCONI, V. E.; MENDES JUNIOR, O.; DOMINGUES, M. O.; CAETANO, M.; STEPHANY, S. **Software para visualização de centros de atividade convectiva (CAC)**. São José dos Campos: INPE, 2010.

MESINGER, F. A blocking for representation of mountains in atmospheric models. **Rivista di Meteorologia Aeronautica**, v.44, p.195-202, 1984.

MESINGER, F., Z. I. JANJIC, S. NICKOVI , D. GAVRILOV, e D. G. DEAVEN The step-mountain coordinate: Model description and performance for cases of Alpine lee cyclogenesis and for a case of Appalachian redevelopment. **Monthly Weather Review**, v.116, n.7, p.1493-1518, 1988.

PAWLAK, Z. Rough sets. **International Journal of Computer and Information Sciences**, v.11, n. 5, p. 341-356, 1982.

POLITI, J. **Implementação de um ambiente para mineração de dados aplicada ao estudo de núcleos convectivos**. 2005. 137 p. (INPE-14165-TDI/1082). Dissertação (Mestrado em Computação Aplicada) - INPE, São José dos Campos. 2005.

POLITI, J.; STEPHANY, S.; MENDES JUNIOR, O.; DOMINGUES, M. O. Mineração de dados meteorológicos associados à atividade convectiva empregando dados de descargas elétricas atmosféricas. **Revista Brasileira de Meteorologia**, v. 21, n. 2, p. 232-244, 2006.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Francisco-CA: Morgan Kaufmann Publishers Inc, 1993. 302p.

ROZANTE, J. R.; CAVALCANTI, I. F. A. Eta model experiments during the SALLJEX period. In: **International Conference on Southern Hemisphere Meteorology and Oceanography (ICSHMO)**, 2006, Foz do Iguaçu. **Proceedings...** 2006. p. 1963-1968. CD-ROM, On-line. Disponível em: <http://urlib.net/cptec.inpe.br/adm_conf/2005/10.20.16.39>. Acesso em: 07 jun. 2011.

SCOTT, D. W. **Multivariate Density Estimation: Theory, Practice, and Visualization**. New York: John Wiley and Sons, 1992. 376p.

SILVERMAN, B. W. **Density Estimation for Statistics and Data Analysis (Monographs on Statistics and Applied Probability 26)**. London: Chapman and Hall, 1990. 176p.

SCUDERI, S. Conjuntos rough, Leopoldianum. **Revista de Estudos e Comunicação da Universidade Católica de Santos**, v. 27, n. 75, p. 185-197, 2003.

STRAUSS, C.; STEPHANY, S.; CAETANO, M. A ferramenta EDDA de geração de campos de densidade de descargas atmosféricas para mineração de dados meteorológicos. In: XXXIII Congresso Nacional de Matemática Aplicada e Computacional (XXXIII CNMAC). **Anais**. Águas de Lindóia, 2010.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. San Diego-CA: Academic Press, 2006. 837p.

TORRES R. R.; FERREIRA, N. J. Case Studies of Easterly Wave Disturbances over Northeast Brazil suing the Eta Model. **Weather and Forecasting**, v.26, n.2, p.225-235, 2011.