CNMAC2012
XXXIV CONGRESSO NACIONAL DE
MATEMÁTICA APLICADA E COMPUTACIONAL
17 a 21 de setembro de 2012 – Águas de Lindóia/SP
ISSN 1984-8218

Comparação de Abordagens de Classificação para um Problema de Mineração de Dados Meteorológicos

Glauston Roberto Teixeira de Lima, Stephan Stephany,
INPE – Laboratório Associado de Computação e Matemática Aplicada (LAC)
12.227-010 São José dos Campos, SP
E-mail: glau11@gmail.com, stephan@lac.inpe.br

***Resumo**: Este trabalho propõe três abordagens de classificação para a detecção de ocorrências de eventos convectivos severos por meio do monitoramento das saídas do modelo de previsão numérica de tempo Eta. Assume-se que esses eventos possam ser correlacionados com um grande número de descargas elétricas atmosféricas, expressas pelo correspondente campo de densidade de ocorrências. Neste contexto, foram desenvolvidas e testadas duas abordagens de classificação baseadas em similaridade de vetores e agrupamentos e também um classificador neural com o objetivo de identificar padrões indicativos de eventos severos. O desempenho de classificação dessas abordagens é analisado para esquemas de validação cruzada e de predição para testes realizados com três mini-regiões do território brasileiro. Os resultados mostram ser possível desenvolver um classificador para auxílio ao meteorologista na previsão do tempo.*

Palavras-chave: mineração de dados, previsão meteorológica, eventos convectivos, similaridade de vetores, agrupamento, redes neurais artificiais.

1. Introdução

A detecção antecipada e semi-automática de eventos convectivos severos é desejável pela possibilidade de evitar, ou pelo menos minimizar, os impactos negativos causados por tais eventos e tornou-se um tema atual de pesquisa em Meteorologia. No entanto, a tendência atual é que a crescente quantidade de dados e imagens meteorológicos torne difícil, senão impossível, sua pronta análise por meteorologistas. Além disso, eventos convectivos severos podem ocorrer em escalas temporais e espaciais que dificultam ainda mais uma predição antecipada e precisa. Assim, seria interessante, que além de contar com o conhecimento do especialista, essa análise pudesse ter o suporte de ferramentas mais avançadas capazes de interpretar estes dados e auxiliar na tomada de decisão do próprio meteorologista, no sentido de agilizar a previsão do tempo sem perda de confiabilidade.

Nesse escopo, o presente trabalho faz uma comparação entre 3 abordagens de classificação que foram desenvolvidas com a finalidade de detectar ocorrências de eventos convectivos severos a partir das saídas do modelo meteorológico regional Eta. As abordagens de classificação propostas, denominadas abordagem I, II e III, atribuem um vetor de variáveis meteorológicas gerado pelo modelo Eta a uma das seguintes classes: atividade convectiva de intensidade desprezível/fraca, atividade convectiva de intensidade média ou atividade convectiva de intensidade forte. Os classificadores propostos incorporam conceitos de aprendizagem de máquina que possibilitam o treinamento a partir de um conjunto de instâncias conhecidas. A abordagem I, a primeira a ser desenvolvida, baseada nas freqüências de ocorrência associadas às variáveis de um vetor de acordo com suas faixas de valores, já apresentara desempenho promissor na identificação de eventos convectivos severos [1, 2]. Na

ISSN 1984-8218

seqüência foi desenvolvida a abordagem II que é baseada em agrupamentos gerados a partir de uma métrica de similaridade robusta e a abordagem III que utiliza uma rede neural artificial e atualmente encontra-se em fase de aperfeiçoamento.

As abordagens I e II tiveram bons desempenhos em esquemas de classificação por validação cruzada utilizando-se 10 partições. No entanto, foram também realizados testes para predição de eventos severos, separando-se um conjunto de dados para treinamento e outro conjunto, com instâncias temporalmente posteriores, para teste. Neste esquema de predição, as abordagens I e II tiveram um desempenho bem inferior àquele alcançado na classificação por validação cruzada. Estes resultados levaram à exploração de novas abordagens, surgindo assim, a rede neural proposta dado a difusão crescente desta técnica em mineração e modelagem de dados [3, 4, 5]. Os resultados obtidos até o momento com esta nova abordagem (III) são encorajadores.

A análise de desempenho destas três abordagens feita neste artigo é parte de uma pesquisa relativa à mineração de dados meteorológicos com o objetivo de detectar eventos convectivos severos iniciada em 2007.

2. A Base de Dados Meteorológicos

Os dados binários do modelo meteorológico Eta foram fornecidos pelo CPTEC/INPE, sendo referentes aos meses de janeiro e fevereiro de 2007 para uma área de 20° de latitude por 20° de longitude (ou 101 pixels por 101 pixels, considerando a resolução de 20 km dos dados) que abrange boa parte da América do Sul. Estes dados são gerados pelo Eta a cada 6 horas. Uma análise efetuada com um meteorologista, selecionou dentre as centenas de variáveis de saída do modelo Eta, 26 variáveis como sendo mais relevantes para a detecção de eventos severos. 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 [6] de forma a gerar os campos de densidade de ocorrência de descargas elétricas atmosféricas para uma extensão geográfica e intervalo de tempo selecionados. A densidade de ocorrência de descargas, que constitui o atributo de decisão, em termos de indicar o nível de atividade convectiva, foi então discretizada em 3 faixas com base numa avaliação feita para casos de atividade convectiva severa conhecidos: atividade convectiva desprezível/fraca, moderada ou forte, utilizando-se os limiares de densidades de descarga 0.0025 e 0.01, os quais foram determinados empiricamente. Essa divisão em 3 classes foi adotada para as abordagens I e II. Entretanto, para viabilizar a abordagem III, optou-se por usar apenas 2 classes: a fraca/moderada (obtida pela fusão dos elementos da desprezível/fraca e da moderada) e a forte (igual à original). Para testar as 3 abordagens de classificação, foram selecionados os dados de 3 mini-regiões de 1° de latitude por 1° de longitude no território brasileiro: a mini-região do Pantanal Sul Matogrossense (A), a mini-região da Alta Sorocabana paulista entre Bauru e Presidente (B) e, finalmente, a mini-região do Vale do Paraíba e Litoral Norte (C).

3. Abordagens Propostas de Classificação

3.1 Abordagem I - classificação por freqüência de ocorrência

Na abordagem I, primeiramente são formados 3 agrupamentos (*clusters*) representando as 3 classes adotando-se como critério de decisão apenas a densidade de ocorrência de descargas elétricas atmosféricas associada a cada vetor e de acordo com os limiares de divisão pré-estabelecidos. Os vetores de variáveis meteorológicas da classe atividade convectiva fraca/desprezível (C_1) são alocados no agrupamento que aqui também é denominado de C_1. Os vetores das classes atividade convectiva média (C_2) e atividade convectiva forte (C_3), são

ISSN 1984-8218

alocados nos agrupamentos C_2 e C_3, respectivamente. N_1, N_2 e N_3 são as respectivas quantidades de vetores alocadas em cada agrupamento. A seguir, aplica-se um esquema de discretização, baseado em subdivisões sucessivas, e as faixas de valores das variáveis em cada agrupamento são discretizadas em 16 sub-faixas.

Um vetor $\mathbf{V} = [v^1, v^2, \ldots, v^{25}, v^{26}]$ a ser classificado como pertencente a um dos 3 agrupamentos é considerado, na abordagem I, como sendo um conjunto de eventos, no sentido probabilístico do termo. Assim, sob o esquema de discretização adotado, o valor assumido por cada variável em **V** e os valores de todas as suas duplas de variáveis, constituem eventos (probabilísticos) para os quais, portanto, podem ser calculadas freqüências de ocorrência. Para tanto, os agrupamentos C_1, C_2 e C_3 funcionam como aproximações dos espaços amostrais da classe que representam e a partir deles são construídas para **V**, 3 matrizes de freqüência de ocorrência $\mathbf{M_1}$, $\mathbf{M_2}$ e $\mathbf{M_3}$, de dimensão 26x26, com as quais a pertinência de **V** a cada classe é estimada. O processo de construção destas matrizes é explicado a seguir.

Para cada variável v^j de **V** são executados os seguintes 4 passos de cálculo:

Passo 1: em cada agrupamento contam-se os vetores cujo valor da j-ésima variável está na mesma sub-faixa que o valor de v^j, resultando nas quantidades Q_1^j, Q_2^j e Q_3^j.

Passo 2: o j-ésimo elemento diagonal das matrizes $\mathbf{M_1}$, $\mathbf{M_2}$ e $\mathbf{M_3}$ é preenchidos com os valores: Q_1^j / N_1, Q_2^j / N_2, Q_3^j / N_3, respectivamente.

Passo3: em cada agrupamento, os vetores identificados no passo 1 são inspecionados com relação às demais 25 variáveis. Para cada remanescente variável, contam-se, dentre os vetores previamente identificados no passo 1, aqueles cujo valor da i-ésima variável encontra-se na mesma sub-faixa que o valor da variável v^i ($i \neq j$) de **V**. Essa inspeção resulta 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 $\mathbf{M_1}$, $\mathbf{M_2}$ e $\mathbf{M_3}$ com os valores: $[q_1^{j1}, \ldots, q_1^{j26}]/Q_1^j$, $[q_2^{j1}, \ldots, q_2^{j26}]/Q_2^j$ e $[q_3^{j1}, \ldots, q_3^{j26}]/Q_3^j$, respectivamente.

A Figura 1 mostra as matrizes $\mathbf{M_1}$, $\mathbf{M_2}$ e $\mathbf{M_3}$ com a j-ésima coluna preenchida de acordo com os passos acima descritos.

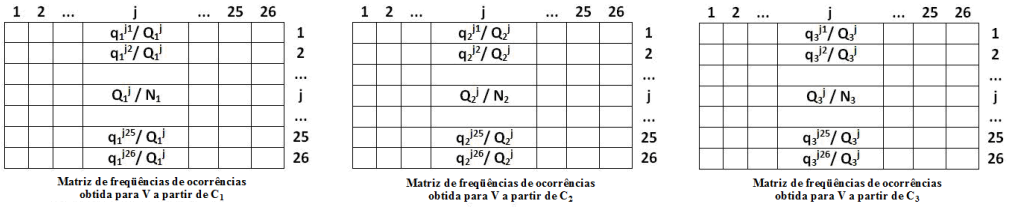

Figura 1: Construção das matrizes de freqüência de ocorrência M_1, M_2 e M_3 utilizadas na classificação do vetor V.

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

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

Esta abordagem, baseada na freqüência de ocorrência, tem a vantagem de prescindir do cálculo (ou estimação) de quaisquer funções de densidade de probabilidade. Além disso, na classificação, a comparação de vetores privilegiando as informações individuais de variáveis ou de duplas de variáveis, em vez da comparação

ISSN 1984-8218

simultânea das 26 variáveis, parece ser conveniente uma vez que em sistemas de decisão multivariável as correlações entre variáveis de informação e variável de decisão nem sempre são iguais. A complexidade algorítmica desta abordagem é O(N), onde N é o número de vetores disponíveis para formar os agrupamentos.

3.2. Abordagem II - classificação por agrupamentos

A abordagem II é, na maioria dos aspectos, muito similar às abordagens de agrupamento mais tradicionais, mas utiliza uma métrica de similaridade pouco comum para avaliar a pertinência de um vetor a um dado agrupamento. Na fase de formação dos agrupamentos, os vetores são agrupados segundo um esquema aglomerativo e iterativo sem se considerar a variável de decisão (que voltará a ser utilizada na fase de classificação). Inicialmente, cada vetor da base de dados é considerado um agrupamento. A cada iteração, os vetores são apresentados aleatoriamente e sua pertinência aos agrupamentos já existentes é medida utilizando-se a métrica de similaridade considerada. Conseqüentemente, quando as iterações prosseguem, os vetores podem migrar de um agrupamento para outro. Uma variável de controle interrompe as iterações quando essas migrações cessam. Uma vez que os agrupamentos estejam definidos, cada um é rotulado de acordo com a classe (definida pela variável de decisão) à qual pertencem a maioria de seus vetores. Quando um novo vetor (não utilizado na fase de formação dos agrupamentos) é apresentado para classificação, a mesma métrica de similaridade é usada para avaliar sua pertinência a cada agrupamento e o novo vetor é atribuído à classe do agrupamento "vencedor". A complexidade algorítmica desta abordagem é $O(N^2)$, onde N é o número de vetores disponíveis para formar os agrupamentos.

A métrica de similaridade proposta é apresentada aqui. Assumindo um dado conjunto de agrupamentos, um vetor $\mathbf{V} = [v^1, v^2, \ldots, v^{25}, v^{26}]$ é avaliado como pertencente a um agrupamento K_i , comparando-o ao centróide de cada agrupamento usando a média harmônica $H(V, K_i)$ de duas quantidades: a distância Euclidiana e (1/26) da distância de Manhattan conforme equação (1). O vetor que corresponde ao centróide de K_i é denotado por c_i. O vetor **V** é classificado como pertencente ao agrupamento que apresente a menor média harmônica. Uma vez que as faixas de valores originais das 26 variáveis diferem muito, foi necessário fazer uma normalização das mesmas para o intervalo [-1 1]. A métrica de similaridade em (1) foi proposta tendo em vista que ela combina a "medida geral" (no sentido de mostrar a influência conjunta de todas as variáveis) dada pela distância Euclidiana com a distância de Manhattan que enfatiza a influência particular das variáveis.

$$H(\mathbf{V}, \mathbf{K_i}) = \frac{2}{\frac{1}{\sqrt[2]{\sum_{j=1}^{26}\left(v^j - c_i^j\right)^2}} + \frac{26}{\sum_{j=1}^{26} abs\left(v^j - c_i^j\right)}} \qquad (1)$$

A abordagem II gerou mais de 1500 agrupamentos para cada mini-região contendo entre 3 e 12 vetores. A reprodutibilidade da abordagem foi testada tendo em vista sua estocasticidade. Para tanto, 10 execuções do algoritmo foram realizadas para cada mini-região. Os resultados, com relação à "pureza dos agrupamentos, mostraram que um percentual maior do que 97% foi obtido em termos de "agrupamentos puros", ou seja, agrupamentos contendo somente vetores pertencentes a uma determinada classe.

3.3. Abordagem III - classificação com rede neural artificial

Na abordagem de classificação III uma nova rede neural está sendo testada. O número de nós de entrada da rede é igual ao número de variáveis dos vetores meteorológicos. As 26 entradas são completamente conectadas a dois conjuntos de neurônios ocultos distintos: o grupo de neurônios $[O_{1p}, \ldots, O_{jp}, \ldots, O_{26p}]$ cujas funções de ativação são sigmóides bipolares

ISSN 1984-8218

com inclinação positiva (por isso o subscrito p) e o grupo de neurônios [O_{1n} , ... , O_{jn}, ... , $O2_{6n}$] cujas funções de ativação são sigmóides bipolares com inclinação negativas (por isso o subscrito n). Utilizam-se esses dois tipos de neurônios ocultos para que a rede tenha maior flexibilidade na construção da fronteira de classificação. Esta hipótese, no entanto, está sob avaliação. A rede tem apenas um neurônio de saída porque, por hora, está sendo testada para fazer o reconhecimento de duas classes: atividade convectiva não forte (que reúne os casos de atividade convectiva desprezível/fraca e média) e atividade convectiva forte. O projeto de uma rede neural (arquiteturas e algoritmos de treinamento) para reconhecimento de duas classes é menos complexo e nesta fase inicial do desenvolvimento, em que se busca "sentir" o potencial da abordagem para a tarefa pretendida, essa menor complexidade é conveniente. Com base na análise dos resultados que forem sendo colhidos, a rede poderá, posteriormente, ser projetada para o reconhecimento de 3 classes como feito nas abordagens I e II. A Figura 2 mostra a arquitetura da rede neural.

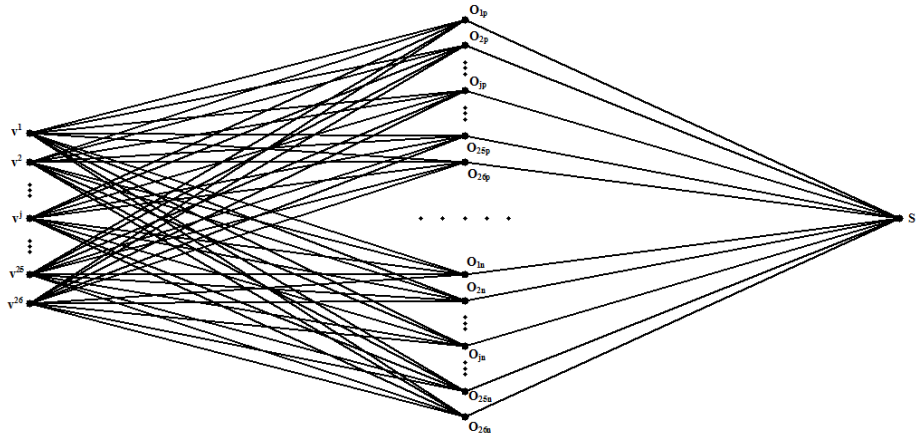

Figura 2: Arquitetura da rede neural utilizada para classificar os dados meteorológicos

Para a rede neural acima são definidos:

$\mathbf{P_{EOp}}$ e $\mathbf{P_{EOn}}$: são as matrizes de pesos conectando as entradas aos neurônios ocultos O_{jp}, e O_{jn}.
$\mathbf{P_{OSp}}$ e $\mathbf{P_{OSn}}$: são os vetores de pesos conectando os neurônios ocultos O_{jp} e O_{jn} à saída.
φ_p, φ_n e φ_s : são as funções de ativação de O_{jp} e O_{jn} e do neurônio de saída.

Se $\mathbf{V} = [v^1, v^2, \ldots, v^{25}, v^{26}]$ é apresentado à rede, a saída S é dada por:

$$S = \varphi_s(\varphi_p(\mathbf{V} * \mathbf{P_{EOp}}) * \mathbf{P_{OSp}} + \varphi_n(-\mathbf{V} * \mathbf{P_{EOn}}) * \mathbf{P_{OSn}}) \quad (2)$$

Se **V** pertence à classe atividade convectiva não forte, a saída S deve maior do que um limiar L e se **V** pertence à classe atividade convectiva forte, a saída deve ser menor do que -L. Caso contrário haverá um erro E dado por:

$$E = L - S \quad \text{ou} \quad E = -L - S \quad (3)$$

O ajuste dos pesos é feito ao término de cada época (*batch learning*) tomando-se as médias dos ajustes relativos aos erros E_i gerados por cada vetor V_i. A taxa de aprendizado aplicada nestes ajustes decresce exponencialmente com o número de épocas.

A função custo minimizada é :

$$F_c = EQM + \delta \sum_{i=1}^{m} P_i^2 \quad (4)$$

ISSN 1984-8218

Em (4), EQM é o erro quadrático médio, Pi é um dos pesos da rede, m é o número total de pesos da rede e δ é um parâmetro de regularização. O método utilizado para minimizar F_c é o gradiente descendente.

4. Resultados

Os resultados a seguir referem-se às 3 mini-regiões consideradas (A, B e C) para as 3 abordagens, sendo que aparecem os resultados de classificação usando validação cruzada para as abordagens I e II, com 3 classes. No caso da abordagem III (rede neural), mostram-se os resultados da classificação no esquema de predição usando apenas 2 classes. O desempenho de classificação pode ser avaliado por meio das matrizes de confusão resultantes e também por métricas conhecidas de avaliação de classificação: a acurácia e o índice Kappa. Estes resultados estão apresentados nas Tabelas 1, 2 e 3 abaixo.

	Mini-região A				Mini-região B				Mini-região C			
Abordagem I		C_1	C_2	C_3		C_1	C_2	C_3		C_1	C_2	C_3
	C_1	790	5	0	C_1	737	11	0	C_1	766	5	0
	C_2	0	39	2	C_2	2	59	8	C_2	1	54	6
	C_3	0	0	13	C_3	0	3	30	C_3	0	1	17
Abordagem II		C_1	C_2	C_3		C_1	C_2	C_3		C_1	C_2	C_3
	C_1	789	6	0	C_1	737	10	1	C_1	765	6	0
	C_2	5	35	1	C_2	11	53	5	C_2	9	48	4
	C_3	0	1	12	C_3	0	5	28	C_3	1	3	14

Tabela 1: Matrizes de confusão para as abordagens I e II.

	Mini-região A	Mini-região B	Mini-região C
Abordagem I	Kappa	Kappa	Kappa
	0.9345	0.8752	0.9130
	Acurácia	Acurácia	Acurácia
	0.9918	0.9718	0.9847
Abordagem II	Kappa	Kappa	Kappa
	0.8741	0.8270	0.8387
	Acurácia	Acurácia	Acurácia
	0.9847	0.9623	0.9729

Tabela 2: Métricas de avaliação para as abordagens I e II.

Mini-região A			Mini-região B			Mini-região C		
	(C_1, C_2)	C_3		(C_1, C_2)	C_3		(C_1, C_2)	C_3
(C_1, C_2)	140	4	(C_1, C_2)	137	5	(C_1, C_2)	126	9
C_3	1	22	C_3	1	22	C_3	6	92
Kappa : 0.8731			Kappa : 0.8587			Kappa : 0.8684		
Acurácia : 0.9685			Acurácia : 0.9636			Acurácia : 0.9356		

Tabela 3: Matrizes de confusão e métricas de avaliação para a abordagem III.

ISSN 1984-8218

5. Comentários Finais

Neste trabalho foram apresentadas e comparadas três abordagens de classificação para a detecção de ocorrências de eventos convectivos severos por meio do monitoramento das saídas do modelo de previsão numérica de tempo Eta. Assumiu-se como premissa que esses eventos possam ser correlacionados com um grande número de descargas elétricas atmosféricas, sendo empregado o correspondente campo de densidade de ocorrências. Assim, as variáveis selecionadas do modelo Eta constituem os atributos de informação, enquanto o atributo de decisão é dado pela densidade de descargas. Neste contexto, foram desenvolvidas e testadas duas abordagens de classificação baseadas em similaridade de vetores (abordagem I) e agrupamentos (abordagem II) e também um classificador neural (abordagem III) com o objetivo de identificar padrões indicativos de eventos severos nas saídas do modelo.

As abordagens I e II obtiveram bom desempenho de classificação apenas no esquema de validação cruzada, em que os dados são embaralhados aleatoriamente e separados em conjuntos de treinamento e de teste. Isso levou ao desenvolvimento da abordagem III, que tornou possível um esquema de predição em que uma data delimita o conjunto de treinamento (com dados anteriores a essa data) e de teste (com dados posteriores a essa data). Embora os testes sejam relativos a três mini-regiões do território brasileiro, os resultados da abordagem III são promissores no sentido de desenvolver uma ferramenta semi-automática de auxílio à previsão para o meteorologista. Assim, durante a execução do modelo de previsão, no caso o modelo Eta, suas saídas seriam monitoradas pelo classificador, indicando ao meteorologista a possibilidade de ocorrência de eventos convectivos severos.

Referências

[1] G. R. T. de Lima, J. D. S. Silva, S. Stephany, C. Strauss, M. Caetano, N. J. FERREIRA, Mineração de dados meteorológicos para previsão de eventos severos pela abordagem de similaridade de vetores. In: XXXIII Congresso Nacional de Matemática Aplicada e Computacional, Águas de Lindóia, pp.729-735, 2010.

[2] Pessoa, A. S. A., Lima, G. R. T., Silva, J. D. S., Stephany, S., Strauss, C., Caetano, M., Ferreira, N. J., Meteorological data mining for the prediction of severe convective events, Revista Brasileira de Meteorologia, (artigo aceito para publicação).

[3] S. Haykin, "Neural Networks: a comprehensive foundation" Prentice-Hall, Upper Saddle River, 1999.

[4] L. Fausett, "Fundamentals of Neural Networks: architectures, algorithms and applications", Prentice-Hall, Upper Saddle River, 1994.

[5] G. R. T. Lima, J. D. S. Silva, O. Saotome, Vehicle inductive signatures recognition using a Madaline neural network, Neural Computing and Applications**,** v.19, n.3, pp 421-436, 2009.

[6] C. Strauss, S. Stephany, M. Caetano, 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, Águas de Lindóia, pp. 269-275, 2010.