Utilização do Algoritmo de Johnson e o Particionamento Aleatório do Conjunto de Treinamento na pela Teoria dos Conjuntos Aproximativos para a Mineração de Dados Meteorológicos

Alex S. A. Pessoa[1], Stephan Stephany[2]

[1]Programa de Mestrado ou Doutorado em Computação Aplicada – CAP
Instituto Nacional de Pesquisas Espaciais – INPE

[2]Laboratório Associado de Computação e Matemática Aplicada – LAC
Instituto Nacional de Pesquisas Espaciais – INPE

asapessoa@gmail.com, stephan@lac.inpe.br

Abstract. *This work aims to detect patterns associated with the occurrence of severe convective events in meteorological data. It is used the Rough Sets Theory in conjunction with Johnson's algorithm, which is a heuristic that simplifies the calculation of reductions. It also adopts a partitioning scheme of the training data in order to enable the mining of large volumes of data.*

Resumo. *Este trabalho busca a detecção de padrões associados à ocorrência de eventos convectivos severos em dados meteorológicos. Utiliza-se a Teoria dos Conjuntos Aproximativos conjuntamente com o algoritmo de Johnson, que é uma heurística que simplifica o cálculo das reduções. Adota-se também um esquema de particionamento dos dados de treinamento de forma a viabilizar a mineração de grandes volumes de dados.*

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

1. Introdução

A grande quantidade e multiplicidade de dados e imagens meteorológicas gerados por modelos numéricos e sensores, embarcados em satélites ou não, torna mais complexo o trabalho do meteorologista na previsão do tempo. Assim, ferramentas auxiliares tornam-se desejáveis e ultimamente a aplicação de técnicas de mineração de dados vem se expandindo. Um dos objetivos é a detecção automática nos dados de padrões associados a determinados fenômenos meteorológicos. Em particular, a detecção de eventos convectivos severos tem importância devido a seu impacto sócio-econômico, sendo objeto do presente trabalho.

Eventos convectivos severos meteorológicos são fenômenos associados à chuvas e ventos fortes de curta duração ou então de intensidade média com duração prolongada, que em geral, causam sérios danos, tornando sua previsão altamente desejável. Eventos

deste tipo, devido à sua baixa freqüência são de difícil previsão por meteorologistas, embora existam esquemas de alerta antecipado. Neste contexto, este trabalho aborda a mineração de dados meteorológicos com o objetivo de detectar padrões associados à atividade convectiva severa nos dados do modelo numérico de previsão de tempo Eta [Mesinger et al. 1988] por meio de um classificador baseado na Teoria dos Conjuntos Aproximativos (TCA).

Se o modelo numérico fosse suficientemente preciso, poderia-se detectar atividade convectiva severa analisando-se certas variáveis nas saídas do próprio modelo. Obviamente isso não se verifica, embora o modelo Eta, em particular, possa simular com razoável precisão variáveis tais como pressão, temperatura, conteúdo de umidade ou ventos em vários níveis de pressão. Optou-se então pelo uso de dados de descargas elétricas atmosféricas, assumindo-se que uma grande quantidade destas possa ser indicativa de atividade convectiva severa.

Os dados de descargas são tratados e agrupados espaço-temporalmente por meio de uma técnica de análise espacial aplicada de maneira inovadora a esse tipo de dados [Politi et al. 2005][Caetano et al. 2009]. Esse agrupamento gera um campo de densidade de ocorrências de descargas que permite identificar regiões mais densas como sendo núcleos de atividade elétrica (NAEs). O campo de densidade gerado é dividido em três classes de densidade: D, M e F, associados respectivamente a eventos convectivos fracos/desprezíveis, moderados ou fortes, tal como proposto em [Pessoa et al. 2010].

A TCA é amplamente empregada em mineração de dados e construção de classificadores, devido ao seu tratamento eficaz e eficiente de incertezas e imprecisões em bases de dados diversas. Neste trabalho, cada elemento ou objeto da base de dados refere-se a um pixel do modelo Eta e ao instante de tempo considerado. Assim, as variáveis do modelo constituem os atributos condicionais, enquanto que a densidade de descargas, para o mesmo pixel e instante, o atributo de decisão, referente ao nível de atividade convectiva.

Em Pessoa et al. (2010), foi aplicada a TCA para dados do modelo Eta e de densidade de descargas com o objetivo de detectar padrões associados à atividade convectiva severa para 3 mini-regiões do território brasileiro, empregando apenas dados dos meses de janeiro e fevereiro de 2007. A aplicação da TCA consiste na geração de conjuntos reduzidos de atributos, as chamadas reduções, e nas regras decorrentes que mapeiam os atributos de informação aos atributos de decisão. Os resultados obtidos foram promissores, mas a extensão da abordagem então empregada para um volume maior de dados era limitada pelo volume de processamento, pela grande quantidade de regras geradas e pela falta de escalabilidade, em termos do acréscimo de mais meses de dados.

O presente trabalho objetiva contornar essas dificuldades com o emprego do algoritmo de Johnson [Johnson 1974] para o cálculo de reduções, o qual é rápido e retorna apenas a melhor redução. No trabalho anterior, empregou-se um algoritmo genético, que além de ser mais lento, retorna várias reduções, induzindo um conjunto de regras muito grande. Entretanto, o algoritmo de Johnson poderia levar a um conjunto de regras muito reduzido que não fosse representativo dos dados que se deseja modelar. Assim, optou-se por utilizar também uma técnica recentemente proposta em TCA, conhecida como *particionamento aleatório dos dados de treinamento* [Gupta et al. 2005], que provê uma estratégia de divisão da base de dados visando à diminuição da complexidade no cálculo das reduções, uma vez que trata cada partição separadamente, permitindo a construção de um classificador com melhor desempenho, conforme

demonstrado pelos resultados.

A classificação da base de dados meteorológicos foi estendida (em relação ao trabalho supracitado) para os meses de janeiro de fevereiro de 2007 a 2011, compondo uma base de dados única que foi dividida aleatoriamente em 1, 8, 16 ou 32 partições. Os resultados mostram o bom desempenho da abordagem proposta. Abordam-se a seguir a descrição dos dados meteorológicos empregados, a metodologia adotada neste trabalho, compreendendo a TCA, o algoritmo de Johnson, e o esquema de particionamento aleatório e, finalmente, os resultados e comentários finais.

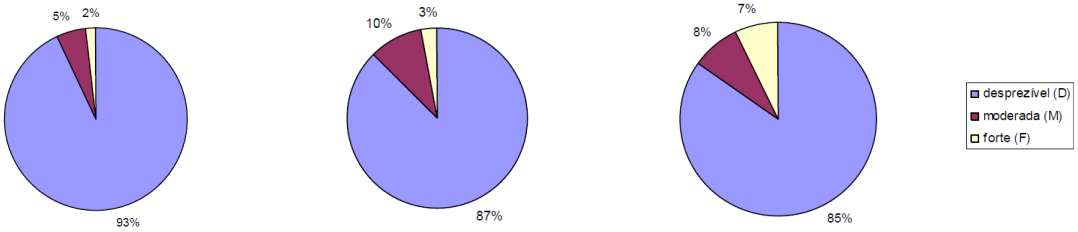
2. Dados e Metodologia

2.1. Dados empregados

Os dados utilizados para representar o estado da atmosfera foram gerados pelo modelo de previsão do tempo Eta e foram fornecidos pelo Centro de Previsão de Tempo e Estudos Climáticos (CPTEC/INPE), sendo referentes aos meses de janeiro e fevereiro dos anos de 2007 a 2011. Os dados empregados do modelo Eta são constituídos de 58 variáveis, selecionadas com auxílio de meteorologistas, sendo 9 variáveis unidimensionais e 7 variáveis em níveis (1000, 925, 800, 700, 500, 300 e 250 hPa), tem uma resolução espacial de 20 km e suas saídas são correspondentes aos horários de 0, 6, 12 e 18hs UTC. Já os dados de descargas elétricas, contendo os registros individuais em formato ASCII foram gerados pela Rede Integrada Nacional de Detecção de Descargas Atmosféricas (RINDAT), sendo também cedidos 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. A densidade de ocorrência de descargas, que constitui o atributo de decisão foi então discretizada em 3 faixas com base numa avaliação feita para casos de atividade convectiva severa conhecidos. Para uma melhor análise optou-se por analisar regiões pequenas, denominadas mini-regiões, que possuíam bons históricos de eventos severos. As 3 mini-regiões foram definidas com uma extensão de 1 grau de latitude e 1 grau de longitude, ou um total de 36 pixels considerada a resolução de 20km do modelo Eta:

A) Pantanal Sul Matogrossense: latitudes 18,4° S a 19,4° S e longitudes 56,4° O a 57,4° O.
B) Alta Sorocabana Paulista: latitudes 21,4° S a 22,4° S e longitudes 49,4° O a 50,4° O.
C) Vale do Paraíba e Litoral Norte: latitudes 23° S a 24° S e longitudes 45° O a 46° O.

A Figura 1 mostra a distribuição de classes para cada mini-região, em termos percentuais. O conjunto de dados é constituído de 42588 elementos.



Mini-região A Mini-região B Mini-região C

Figura 1. Distribuição de classes (densidade de descargas elétricas)

Estes dados foram analisados no software ROSETTA (*Rough Set Toolkit for Analysis of Data*) que é um ambiente para mineração de dados voltados para a TCA [Øhrn 1999].

2.2. Teoria dos Conjuntos Aproximativos (TCA)

A TCA é uma extensão da teoria dos conjuntos, cujo enfoque é o tratamento de imprecisão e incerteza nos dados. Foi inicialmente desenvolvida por Zdzislaw Pawlak [Pawlak 1982] no início da década de 80. Entretanto, a TCA difundiu-se somente na década de 90 graças ao aumento da capacidade de processamento e à difusão do uso dos computadores pessoais.

O conhecimento em TCA é representado na forma tabular, representado pelo par ordenado $S = (U; A)$, chamado de sistema de informação, onde U é um conjunto finito não-vazio de objetos chamado de universo e A é um conjunto finito não-vazio de atributos condicionais ou condições, tal que a: $U \rightarrow V_a$ para todo $a \in A$. O conjunto V_a é chamado de conjuntos de valores de a.

Uma forma particular do sistema de informação é obtida adicionando-se um atributo distinto dos atributos condicionais, com o objetivo de criar classes. Essa forma do sistema de informação é chamada de sistema de decisão, onde o $S = (U; A \cup \{d\})$, onde $d \notin A$ é o atributo de decisão.

A TCA baseia-se na noção de indiscernibilidade, que é uma relação de equivalência entre objetos, considerando-se um subconjunto de atributos, que permite particionar o universo U. Segue a definição formal:

Dado o sistema de informação $S = (U; A)$, então, para qualquer subconjunto de atributos $B \subseteq A$ existe uma relação de equivalência $IND_A(B)$, ou simplesmente $IND(B)$ se:

$$IND(B) = \{(x, x') \in U \mid \forall a \in V, a(x) = a(x')\} \quad (1)$$

$IND(B)$ é chamada de relação de B-indiscernibilidade. Se $(x, x') \in IND(B)$, então os objetos x e x' são indiscerníveis ou iguais considerando-se o subconjunto de atributos B. A classe de equivalência da relação determinada pelos objetos x que verificam $IND(B)$ é denotada $[x]_B$ [Komorowski 1999].

Na teoria de TCA, com base na relação de indiscernibilidade, é possível reduzir os atributos considerados supérfluos, obtendo-se as chamadas reduções. Uma redução é um subconjunto $B \subseteq A$, onde são preservadas as classes de equivalências definidas pelo conjunto de todos os atributos, ou seja, $IND(A)$. Dessa forma, se para um subconjunto de atributos $B \subseteq A$, verifica-se que $IND(B) = IND(A)$, então B é dito redução de A, ou $RED(B)$. As reduções constituem então um método de seleção de atributos, possibilitando também a compactação da base de dados.

Computar reduções é um problema do tipo NP-difícil [Komorowski 1999]. Consequentemente, este tipo de problema requer o uso de metas-heurísticas para busca de soluções, tais como algoritmos genéticos, busca tabu, recozimento simulado ou colônia de formigas. [Hedar et al. 2008]. Neste trabalho, optou-se pelo algoritmo de Johnson, que é uma meta-heurística gulosa, com duas propriedades convenientes: rapidez na computação e retorno de uma única redução.

2.2.1. Cálculo das reduções na TCA

Na TCA o modo clássico de se calcular as reduções é por meio da função de função de discernibilidade, cuja definição é dada a seguir. Dado um sistema de decisão *S* com *n* objetos, sua correspondente matriz de discernibilidade de *S* é uma matriz simétrica *n* x *n* na qual cada elemento c_{ij} consiste no conjunto de atributos que possuem valores diferentes para os objetos x_i e x_j, com $i \neq j$, sendo, portanto a diagonal desta matriz vazia:

$$c_{ij} = \{a \in A \mid a(x_i) \neq a(x_j)\} \text{ para } i,j = 1, ..., n \qquad (2)$$

A função de discernibilidade f_A para um sistema de informação *S* é uma função de *m* variáveis Booleanas: $a_1^*,...,a_m^*$, as quais são correspondentes aos atributos $a_1,..., a_m$, definida por:

$$f_A(a_1^*,...,a_m^*) = \wedge\{\vee c_{ij}^* \mid 1 \leq j \leq i \leq n, c_{ij} \neq \varnothing\} \qquad (3)$$

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

As reduções são obtidas, no modo clássico da TCA, através da simplificação da função f_A.

2.2.2. Algoritmo de Johnson

O algoritmo de Johnson, como dito anteriormente, tem a característica retornar uma única redução *B*, por meio da função de discernibilidade [Johnson 1974]. Denotando-se por *I* o conjunto formado por cada termo *i* da função de discernibilidade f_A e $w(i)$ seu correspondente o peso, o algoritmo pode ser descrito pelo pseudocódigo abaixo:

1. $B=\varnothing$; (assume-se inicialmente um conjunto vazio de atributos)
2. Seja *a* o atributo que maximiza $\sum w(i)$, para todos os conjuntos *i* em *I* que contenham *a*.
3. $B \leftarrow a$ (adiciona-se o atributo *a* à redução *B*)
4. Removem-se todos os conjuntos i de *I* que contenham *a*.
5. Se $I=\varnothing$ retorne *B, c*aso contrário, vá para o passo 2.

Em geral os atributos são computados com pesos unitários, o que implica que, na prática, seleciona-se no passo 2 o atributo que ocorre com mais frequência dentro de *I*.

2.2.3. Particionamento Aleatório do Conjunto de Treinamento - PACT

Neste trabalho, o desempenho da classificação melhorou consideravelmente com adoção da abordagem proposta por Gupta et al. (2005). Utilizam-se partições aleatórias do conjunto de treinamento (*Randomized Training Set Partitions*) na seleção de atributos e aplica-se o algoritmo de Johnson para cada partição. Uma vez que a complexidade algorítmica desse algoritmo varia com o quadrado do número de elementos, obtém-se uma redução de tempo significativa na seleção de atributos e no treinamento. Essa abordagem consiste basicamente em dividir aleatoriamente o conjunto de treinamento em *n* partições de mesmo tamanho, aplicar o algoritmo de Johnson em cada partição. Isso fornece uma redução para cada partição, da qual é derivado um conjunto de regras. Finalmente, o classificador é composto pela união dos *n* conjuntos de regras das *n* partições. Adotou-se esta abordagem com uma única ou mais partições, conforme apresentado adiante.

3. Resultados e Discussões

A seguir são mostrados os resultados de classificação para as 3 mini-regiões (A, B e C), com dados de janeiro e fevereiro de 2007 a 2011, utilizando o particionamento aleatório do conjunto de treinamento para 1, 8, 16, ou 32 partições (aqui denotados por 1P, 8P, 16P ou 32P, respectivamente). O algoritmo gerador das reduções é o de Johnson, que retorna somente uma redução. Então neste caso para o esquema 16P, são geradas 16 reduções. A partir das reduções calculadas é induzido um conjunto de regras, para classificação. O esquema de amostragem dos dados foi de o *holdout*, onde 80% dos dados foram destinados ao treinamento e 20% para teste.

A seguir, as figuras 2, 3, 4 e 5 mostram as matrizes de confusão obtidas para cada esquema de particionamento aleatório e para cada mini-região. As classes "D", "M" e "F" correspondem a atividade convectiva desprezível/fraca, moderada e forte, respectivamente. Nestas matrizes, cada elemento (i,j) fora da diagonal corresponde ao total de elementos da classe i que foram incorretamente classificados como sendo da classe j (com i,$j \in$ {D,M,F}).

	D	M	F
D	7907	3	0
M	406	61	1
F	120	0	20

Mini-região A

	D	M	F
D	7423	24	0
M	656	167	11
F	165	9	63

Mini-região B

	D	M	F
D	7161	6	0
M	704	58	4
F	546	1	38

Mini-região C

Figura 2. Matrizes de Confusão para PACT com 1 partição.

	D	M	F
D	7856	52	2
M	11 2	348	8
F	29	8	103

Mini-região A

	D	M	F
D	7356	89	2
M	217	588	29
F	23	39	175

Mini-região B

	D	M	F
D	7106	44	17
M	221	498	47
F	104	37	444

Mini-região C

Figura 3. Matrizes de confusão para PACT com 8 partições

	D	M	F
D	7871	38	1
M	97	361	10
F	9	9	122

Mini-região A

	D	M	F
D	7349	96	2
M	159	645	30
F	10	41	186

Mini-região B

	D	M	F
D	7100	51	16
M	149	563	54
F	37	47	501

Mini-região C

Figura 4. Matrizes de confusão para PACT com 16 partições

	D	M	F
D	7893	17	0
M	194	266	8
F	46	1	93

Mini-região A

	D	M	F
D	7374	72	1
M	196	610	28
F	18	43	176

Mini-região B

	D	M	F
D	7116	29	22
M	235	484	47
F	64	33	488

Mini-região C

Figura 5. Matrizes de confusão para PACT com 32 partições

Para o caso de partição única (1P) o classificador errou bastante nas atribuições da classe M e em especial da classe F. Isso se deve à cobertura das regras não abrangente gerada a partir de uma única redução. Utilizando o esquema de particionamento múltiplo, a classificação melhora, uma vez que o conjunto de regras é mais abrangente. Isso pode ser facilmente comprovado por métricas de avaliação dos classificadores, mostrados na Tabela 1, que contêm os índices de acurácia ou exatidão global (Acc) e

Kappa (κ). A acurácia por vezes mascara um desempenho ruim do classificar, em função do desbalanceamento das classes, como mostrado na Figura 1. Neste caso o índice Kappa expressou melhor o desempenho do classificador.

Tabela 1. Acurácia e índice Kappa da classificação.

	1P		8P		16P		32P	
Região	**Acc**	**κ**	**Acc**	**K**	**Acc**	**κ**	**Acc**	**κ**
A	0,94	0,22	0,98	0,80	0,98	0,85	0,97	0,72
B	0,9	0,33	0,95	0,78	0,96	0,82	0,96	0,80
C	0,85	0,12	0,95	0,78	0,96	0,85	0,95	0,80

O esquema de particionamento 16P obteve o melhor desempenho de classificação, apresentando índice Kappa acima dos 0,8 para todas as mini-regiões, assim como uma acurácia no mínimo igual, ou melhor, aos demais esquemas.

A Tabela 2 mostra o número de regras geradas (ou cardinalidade, que é denotada por |RUL|) e o tempo médio (*t*) para cálculo de cada redução em um computador pessoal de 4 núcleos. Deve-se enfatizar que o uso de uma partição implica em tempo de execução muito superior a qualquer um destes tempos, constituindo outra vantagem significativa do PACT.

Tabela 2. Cardinalidade das regras e tempo médio (s).

	1P		8P		16P		32P	
Região	**|RUL|**	***t***	**|RUL|**	***t***	**|RUL|**	***T***	**|RUL|**	***t***
A	30939	14400	170609	160	261062	40	317409	35
B	29642	18000	185286	180	367629	50	452522	40
C	32467	25200	212776	600	354098	180	502172	120

Foi verificado um melhor desempenho de classificação para particionamento múltiplo, verificando-se que o conjunto de regras induzido por uma única redução apresentou baixo desempenho de classificação.

4. Comentários Finais

Este trabalho apresentou resultados da mineração de dados meteorológicos aplicada a eventos convectivos severos em 3 mini-regiões do território brasileiro. Foram empregados dados selecionados do modelo meteorológico Eta como atributos condicionais e dados de densidade de descargas atmosféricas como atributo de decisão, assumindo que alta densidade de descargas seja indicativa de atividade convectiva severa. Foram empregados o algoritmo de Johnson e o particionamento aleatório do conjunto de treinamento para obtenção das reduções e das correspondentes regras, para a construção de classificadores baseados na Teoria de Conjuntos Aproximativos.

Os resultados foram expressivos, mostrando que a abordagem proposta pode ser viável para a previsão de ocorrências de eventos convectivos severos por meio de um classificador que monitore as saídas de um modelo numérico de previsão meteorológica. Houve uma clara evolução em relação ao trabalho anterior [Pessoa et al. 2010], que usava uma base de dados relativa apenas ao ano de 2007. A presente abordagem demonstrou ser robusta, mesmo para uma base de dados relativa ao período 2007-2011, devendo-se considerar que existem aspectos sazonais que fazem com que padrões de atividade convectiva possam variar de ano para ano.

Referências

Caetano, M.; Escobar, G.C.J.; Stephany, S.; Menconi, V.E.; Ferreira, N.J.; Domingues, M.O. e Mendes Junior, O. (2009).Visualização de campo de densidade de ocorrências de descargas elétricas atmosféricas como ferramenta auxiliar no nowcasting. In: *Proceedings of XIII Latin American and Iberian Congress on Meteorology and X Argentine Congress on Meteorology*, Buenos Aires, Argentina.

Gupta, K. M.; Moore, P. G.; Aha, e Pal, S. K. (2005). Rough set feature selection methods for case-based categorization of text documents. In: *Proceedings of the First International Conference on Pattern Recognition and Machine Intelligence*, pp. 792-798,. Kolkata, India: Springer.

Hedar, A. R.; Wang, J. e Fukushima, M. (2008). Tabu search for attribute reduction in rough set theory. *Journal of Soft Computing - A Fusion of Foundations, Methodologies and Applications* 12 (9): 909-918. doi:10.1007/s00500-007-0260-1.

Johnson, D.S. (1974). Approximation algorithms for combinatorial problems. *Journal of Computer and Systems Sciences*, 9:256-278.

Komorowski, J.; Pawlak, Z.; Polkowski, L.; e Skowron, A. (1999). Rough sets: a tutorial. In: *Rough fuzzy hybridization: A new trend in decision-making,* (S.K. Pal e A. Skowron, eds.), Springer-Verlag, Singapore.

Mesinger, F.; Janjic, Z. I.; Nickovic, S.; Gavrilov D. e Deaven D. G. (1988) The step-mountain coordinate: Model description and performance for cases of Alpine lee cyclogenesis and for a case of Appalachian redevelopment. *Monthly Weather Review*, 116, 1493-1518.

Øhrn, A. (1999) .*Discernibility and Rough Sets in Medicine: Tools and Applications*. Tese de Doutorado. Norwegian University of Science and Technology, Department of Computer and Information Science, NTNU.

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

Pessoa, A.S.A.; Silva, J.D.S.; Stephany, S.; Strauss, C.; Caetano, M. e Ferreira, N.J. (2010). Mineração de dados meteorológicos associada a eventos severos no Pantanal Sul Matogrossense. In: *XXXIII Congresso Nacional de Matemática Aplicada e Computacional*, 2010, Água de Lindóia, SP. Anais do XXXIII CNMAC, v. 1. p. 1-8.

Politi, J. (2005). *Implementação de um Ambiente para Mineração de Dados Aplicada ao Estudo de Núcleos Convectivos*. Dissertação de Mestrado em Computação Aplicada, INPE, INPE-14165-TDI/1082.

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

Strauss, C.; Stephany, S. e Caetano, M. (2010). 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*, 2010, Água de Lindóia, SP. Anais do XXXIII CNMAC, v. 1. p. 1-8.