Mineração de dados para identificar eventos meteorológicos extremos no Brasil

Heloisa Musetti Ruivo[1], Fernando M. Ramos[2], Haroldo F. de Campos Velho[2]

[1]Programa de 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

Abstract. *This study uses data mining methodologies to analyze extreme weather events. The goal is to identify the relevant climatological factors that influenced such events. Two methods are evaluated: statistical classification of DNA microarrays, and decision trees. These techniques were tested to analyze the flood occurred in 2008 in Santa Catarina, where some climatological parameters responsible for the event can be pointted .*

Resumo. *Este trabalho utiliza metodologias de mineração de dados para analisar fenômenos meteorológicos extremos. O objetivo é identificar os fatores climatológicos relevantes que influenciaram tais eventos. Duas metodologias foram avaliadas: classificação estatística de microarranjos de DNA; e árvores de decisão. Estas técnicas foram testadas para análise da enchente ocorrida em Santa Catarina em 2008, onde se pode apontar alguns parâmetros climatológicos responsáveis pelo evento.*

Palavras-chave: *mineração de dados, análise estatística, árvore de decisão, entropia, classificação.*

1. Introdução

Há evidências experimentais do aquecimento do sistema climático global, através do monitoramento das temperaturas médias globais do ar e dos oceanos. Em um planeta mais aquecido, os fenômenos climáticos e meteorológicos extremos como secas, inundações, tempestades severas, ventanias e incêndios florestais se tornam mais freqüentes [IPCC 2007].

Técnicas de mineração de dados têm sido usadas para estudos de eventos meteorológicos severos como secas intensas. Em [Ruivo 2007], tais tecnologias foram utilizadas para investigar quais os fatores climáticos associados à grande seca de 2005 na Amazônia e as variáveis físico-químicas que controlam a emissão de gases do efeito estufa em reservatórios de hidrelétricas. Este trabalho está voltado para estudo e desenvolvimento de ferramentas para avaliação de chuvas intensas. O estudo de caso estará voltado para a região de Santa Catarina, onde há registros de eventos severos como ventos fortes, chuvas de granizo, enchentes, inundações e até mesmo tornados. Um dos fenômenos mais frequentes que causam desastres naturais em Santa Catarina são as inundações com 1.215 ocorrências em um período de 21 anos [Marcelino et al. 2005].

Em Santa Catarina, chuvas intensas afetaram 1.5 milhões de pessoas resultando em 120 vítimas e 69.000 pessoas desabrigadas. Segundo [Dias 2008], não há registro de um novembro tão chuvoso nas regiões da Grande Florianópolis, Vale do Itajaí e Litoral Norte

como observado em 2008, quando diversos recordes históricos foram quebrados. Em Blumenau e Joinville, os totais do mês ficaram em torno de 1000 mm (equivalente a 1.000 litros/m^2), para uma média climatológica mensal de aproximadamente 150 mm [Marengo 2009].

Neste trabalho, faz-se uso de dados armazenados e emprega-se técnicas computacionais de mineração e classificação de dados com o objetivo de apontar as variáveis climatológicas responsáveis por eventos climáticos extremos. A aplicação realizada aqui investiga quais foram os fatores climáticos associados a enchente ocorrida em Santa Catarina. As metodologias de mineração de dados que serão empregadas no presente estudo são: classificação de dados através da ferramenta computacional BRB-ArrayTools [Simon and Lam 2006] (desenvolvida para aplicações em bioinformática), e algoritmos de árvore de decisão disponíveis no pacote WEKA [Witten and Frank 2000].

2. Mineração de dados e classificação

Mineração de dados é uma fase na descoberta de conhecimento em bancos de dados (KDD) que procura por uma série de padrões escondidos nos dados, frequentemente envolvendo uma aplicação iterativa e repetitiva de métodos de mineração de dados particulares. O objetivo de todo o processo de KDD é tornar os padrões compreensíveis às pessoas, visando facilitar uma melhor interpretação dos dados existentes [Fayyad 1996]. As seis etapas que compõem a descoberta de conhecimento estão ilustradas na Figura 1.

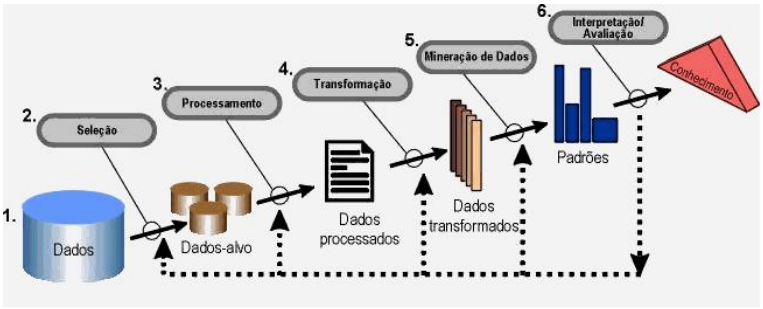

Figura 1. Visão geral das etapas que compõem o processo de KDD.

As várias tarefas desenvolvidas em mineração de dados têm como objetivo primário a predição e/ou a descrição. A predição usa atributos para predizer os valores futuros de uma ou mais variáveis de interesse. A descrição contempla o que foi descoberto nos dados sob o ponto de vista da interpretação humana.

A tarefa mais significativa abordada neste trabalho é a classificação de dados. Essa tarefa consiste em classificar um item de dado como pertencente a uma determinada classe dentre várias classes previamente definidas. A ferramenta de classificação utilizada neste trabalho provém da bioinformática. Trata-se do pacote BRB-ArrayTools versão $4.2.1$ desenvolvido pelo Biometric Research Branch of the Division of Cancer Treatment and Diagnosis of the National Cancer Institute. É um software livre, voltado para análise de dados de Microarranjos de DNA, e está disponível no site *http://linus.nci.nih.gov/ brb/download.html*. Mais detalhes podem ser encontrados em [Simon and Lam 2006].

A tecnologia de microarranjos (MA) de DNA consiste em medir o nível de expressão de milhares de genes simultaneamente. A idéia fundamental é comparar níveis de expressão do gene entre duas amostras de tecidos, uma normal e outra com tumor [Hautaniemi 2003]. A tecnologia de MA é um processo baseado em hibridização que possibilita observar a concentração de mRNA (ácido ribonucleico mensageiro) de uma amostra de células analisando a luminosidade de sinais fluorescentes. O RNA pode catalisar importantes reações biológicas, por isso analisar sua concentração (quantidade) após a hibridização com células normais, fornece grandes informações para a área biomédica (maiores informações em [Hautaniemi 2003]). Hibridização é o processo bioquímico onde duas fitas de ácido nucléico com seqüências complementares se combinam. A abordagem estatística utilizada para determinar quais genes são mais expressivos na análise, está explicado no item 2.1..

Outra feramenta utilizada consiste em árvores de decisão. Métodos de árvore de decisão representam um tipo de algoritmo de aprendizado de máquina que utilizam uma abordagem dividir-para-conquistar para classificar casos usando uma representação baseada em árvores. Esta filosofia baseia-se na sucessiva divisão do problema em vários subproblemas de menores dimensões, até que uma solução para cada um dos problemas mais simples possa ser encontrada.

Uma árvore de decisão é um modelo representado graficamente por nós e ramos, parecidos com uma árvore, mas invertida. O nó raiz é o primeiro nó da árvore e fica no topo da estrutura. Cada nó contém um teste sobre um ou mais atributos (parâmetros) e os resultados deste teste formam os ramos das árvores [Witten and Frank 2005]. Cada nó folha, nas extremidades da árvore, representa um valor de predição para o atributo meta [Meira 2008].

Depois de construída, a árvore pode ser usada para classificar exemplos cuja classe é desconhecida. Para classificar um exemplo, testam-se os valores de seus atributos segundo a árvore de decisão. Um caminho é traçado a partir do nó raiz, descendo pelos ramos de acordo com os resultados dos testes, até chegar a um nó folha, que representa a classe de predição exemplo [Han 2001].

O critério para escolha do atributo que divide o conjunto de exemplos em cada repetição é um dos aspectos principais do processo do método. Entre os critérios mais conhecidos e usados tem-se o ganho de informação e a razão de ganho, definidos com base na teoria da informação (explicado no item 2.2.) [Quinlan 1993]. O ganho de informação é uma medida usada para selecionar o atributo de teste em cada nó de decisão de uma árvore. O atributo com maior ganho de informação é escolhido como atributo de teste. Este atributo minimiza a informação necessária para classificar os exemplos das partições resultantes da divisão. Tal abordagem ligada à teoria da informação minimiza o número de testes esperados para classificar um exemplo e garante que uma árvore simples seja encontrada [Han 2001].

2.1. Análise estatística

Um importante objetivo no estudo de MA de DNA é a identificação de genes que são diferentemente expressos entre classes pré-definidas. Esta identificação com funções desconhecidas pode levar a um melhor entendimento das funções destes genes. A análise estatística empregada neste trabalho é obtida pela opção de *Class Comparisson* do pro-

grama BRB-ArrayTools. A teoria estatística utilizada permite estimar a probabilidade de se ver esta diferença tão grande quanto observada. O método mais comumente usado é o t-estatístico que mede a razão da variação de expressão do gene entre o entre-classes e o interior-classes. O t-estatístico é então convertido para probabilidade, conhecido como p-valor, que representa a probabilidade de se observar em hipótese nula, um t-estatístico tão grande quanto observado no dado real. Maiores detalhes da formulação podem ser obtidos em [Amaratunga and Cabrera 2004].

2.2. Árvores de Decisão

Existem várias implementações utilizando algoritmos de indução (construção) baseados em árvores de decisão conhecidos na literatura. O algoritmo utilizado neste trabalho é o J48 (obtido no programa WEKA), que constrói um modelo de árvore de decisão baseado num conjunto de dados de treinamento, e usa esse modelo para classificar outras instâncias num conjunto de testes [Quinlan 1993].

Para a escolha dos atributos a serem testados, o J48 utiliza uma grandeza chamada "taxa de ganho" para selecionar o atributo que tenha o maior poder de discriminação entre as classes para cada nó. A taxa de ganho mede a quantidade de informação gerada pelo teste de um atributo específico que seja relevante para classificação de um objeto. Assim o algoritmo seleciona os atributos que irão gerar uma árvore simples e eficiente.

A taxa de ganho de informação mede a redução da entropia causada pela partição dos exemplos de acordo com os valores do atributo [Eijkel 1999]. A entropia de Shannon [Shannon 1948] visa medir a incerteza sobre um espaço desordenado de um modo geral. A construção de uma árvore de decisão é guiada pelo objetivo de diminuir a entropia. A introdução da entropia no processo de construção de árvores de decisão visa a criação de árvores menores e mais eficazes na classificação.

3. Aplicação

Para esta análise, foram utilizados dados climatológicos mensais que cobrem o período de janeiro de 1999 a dezembro de 2010 (144 meses). Os dados em grade foram extraídos do site *http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html*. Trata-se de dados globais de reanálise com resolução espacial de 2.5° x 2.5°. A região de análise selecionada compreende uma subregião com coordenadas $30^\circ W$ a $60^\circ W$, e $20^\circ S$ a $50^\circ S$. Como a enchente foi um evento de curta duração, optou-se por trabalhar com pentadas, ou seja, foram calculadas as médias de 5 dias dentro do período. Sendo assim, temos 73 valores para cada variável por um ano. Em seguida foram calculadas as anomalias dentro do período, ou seja, calculou-se a média para cada pentada no ano (73 médias), e em seguida subtraiu-se o valor da pentada da média. Todos os dados utilizados foram tabulados em termos de anomalias (isto é, o valor corrente menos a média da pentada para os 12 anos considerados).

Dentro do espírito da analogia com a análise de MA, cada mês de dados representa uma amostra, ou um "paciente", e uma coluna na base de dados. Já cada variável climatológica, corresponde a um "gene", e uma linha na base de dados. Estendendo a analogia, pode-se imaginar a enchente em Santa Catarina como uma "doença", e os fatores climáticos causadores do fenômeno, os genes reguladores ainda desconhecidos.

A extração do conhecimento do banco de dados é realizada através de "projetos" que necessitam da definição das "classes" que nortearão as operações de classificação. A classificação foi baseada na precipitação bruta medida em uma região fortemente afetada pela enchente em SC (região demarcada por um círculo vermelho nas imagens). Os dados de precipitação são provenientes das estações pluviométricas administradas pela Agência Nacional de Águas (ANA) obtidos diretamente do endereço *http://ana.gov.br/portalsnirh/*. A classificação foi baseada no intervalo entre o menor e o maior valor da série em anomalia gerada (Figura 2). Para uma análise mais específica, dividiu-se o intervalo de anomalia em 3 sub-intervalos: maior anomalia e 8 ([43,5, 8]) - representando chuva abundante, [8, 0] - representando chuva moderada, 0 e menor anomalia ([0, -11,2]) - representando chuva fraca. Os resultados são representados em campos de p-valores (Figuras 3, 4, 5 e 6) e correspondem a comparação entre chuva abundante x chuva moderada. As isolinhas coloridas representam o dado em anomalia da pentada com maior índice de chuva que foi de 22 a 26 novembro 2008 (intervalo da série entre barras vermelhas na Figura 2).

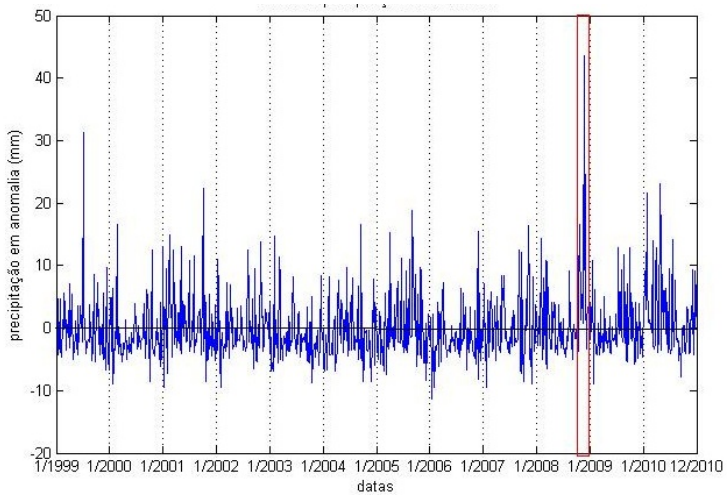

Figura 2. Média da precipitação em Santa Catarina.

Observa-se pelas imagens que regiões com tonalidades mais escuras compreendem os parâmetros que apresentam p-valor < 0.01, ou seja, possuem uma probabilidade menor de 1% de ser um falso positivo. As ilustrações do omega nas Figuras 3 e 4, mostram uma densa nuvem escura que se estende do Oceano Atlântico até o litoral de SC. Observa-se pelas isolinhas que no período da enchente, omega está negativo no continente (movimento vertical ascendente) e positivo no oceano (movimento vertical descendente). Este valor de omega ascendente no continente resulta em precipitação. Nas Figuras 5 e 6, as setas representam a resultante do vento zonal e meridional medido também de 22 a 26 novembro 2008. Segundo [Dias 2008], a localização de um anticiclone de bloqueio no oceano Atlântico (com ventos que giram no sentido anti-horário no Hemisfério Sul, como esquematizado na Figura 6) determinou a ocorrência de ventos de leste sobre boa parte da costa da Região Sul. Esses ventos, devido à orientação Norte-Sul da costa, incidiram mais diretamente sobre o litoral de SC, transportando, portanto, a umidade típica do oceano para o continente. Os ventos persistentes e úmidos vindos do mar foram levantados pela serra catarinense causando o esfriamento e a condensação do ar. Como

consequência disso, chuvas de fraca ou moderada intensidade atingiram continuamente a região litorânea de SC.

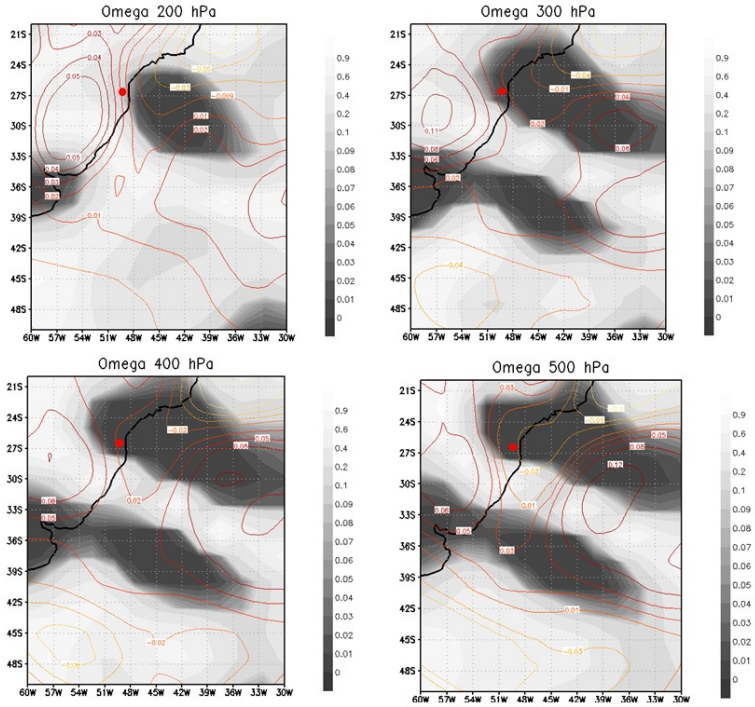

Figura 3. Representação em p-valores da influência das variáveis climatológicas na enchente de Santa Satarina.

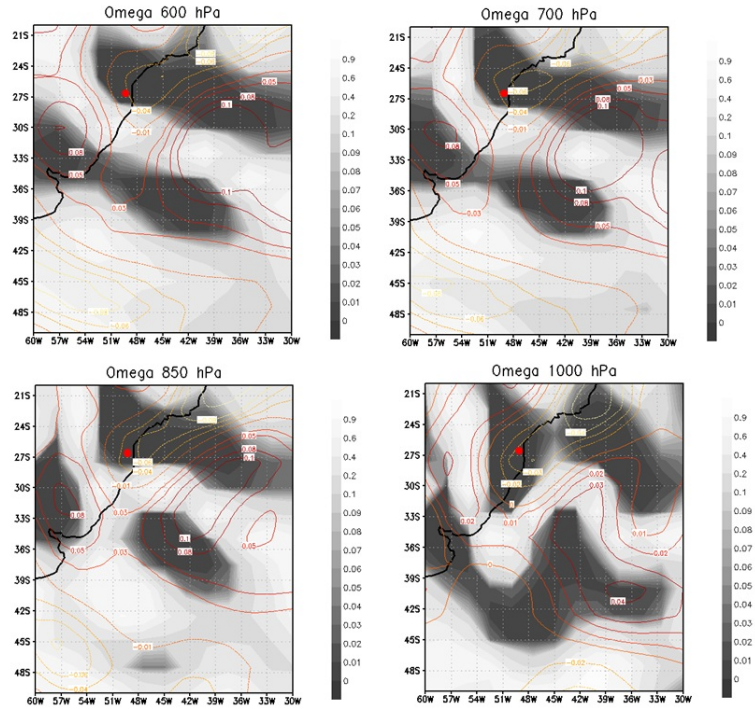

Figura 4. Representação em p-valores da influência das variáveis climatológicas na enchente de Santa Satarina.

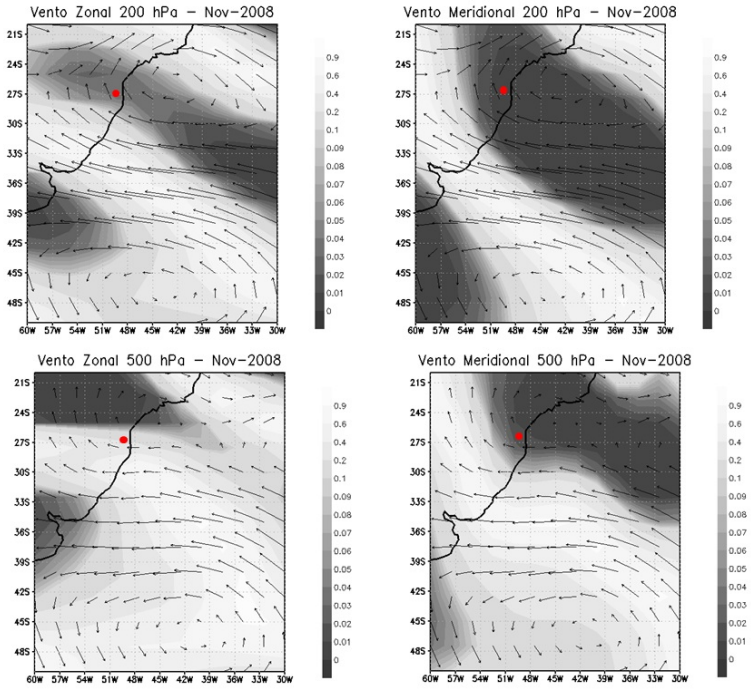

Figura 5. Representação em p-valores da influência das variáveis climatológicas na enchente de Santa Satarina.

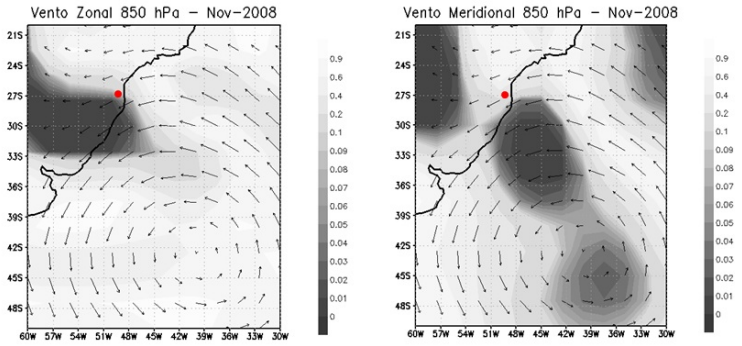

Figura 6. Representação em p-valores da influência das variáveis climatológicas na enchente de Santa Satarina.

A árvore de decisão (Figura 7) foi gerada utilizando as 50 variáveis com menor p-valores identificadas na classificação do BRB. O resultado ilustrado na árvore de decisão foi obtido adotando como classificatória a série de precipitação dividida em 2 classes de acordo com a mediana: pouco (valores abaixo da mediana) e muito (valores acima da mediana), representando respectivamente pouca chuva e muita chuva. Nesta análise foram usadas as séries de 2000 a 2007 como conjunto de treinamento, e 1999, 2008 a 2010 no conjunto de teste. A escolha destes conjuntos foi devido ao fato de que em 1999 houve também uma chuva abundante na região, que pode ser observada na Figura 2. A árvore apresentada identificou corretamente as chuvas intensas ocorrida em outubro/2008 como também a chuva em meados 1999. Podemos observar que a variável omega foi apontada como fator determinante nos eventos.

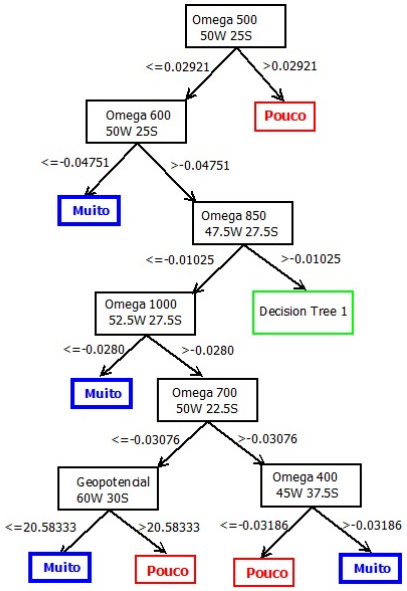

Figura 7. Árvore de decisão gerada: treinamento 2000 a 2007; teste 1999, 2008 a 2010.

4. Conlusões

Neste trabalho, técnicas de classificação de dados, utilizadas atualmente na análise de experimentos de microarranjos de DNA, foram utilizadas na área ambiental. Para isso, variáveis ambientais substituíram os genes na análise. Como a tecnologia de MA aponta os níveis de expressão gênica mais relevante, neste estudo foram apontadas as variáveis climatológicas que mais influenciaram o evento. Em seguida com as variáveis mais significativas na análise, foi gerada uma árvore de decisão para avaliar qual a variável que o classificador aponta como mais importante, e o respectivo limiar. A aplicação ambiental investigou os fatores climáticos responsáveis pela enchente ocorrida em Santa Catarina em 2008. Para isso, foram utilizados dados climatológicos disponíveis na internet pela comunidade científica.

O trabalho apresentou resultados satisfatórios e inéditos por consolidar dados de diversas origens que até hoje não foram analisados de forma integrada. A originalidade se encontra também no fato de mostrar que métodos da bioinformática que utilizam análise estatística podem ser aplicados na área ambiental. A metodologia estatística empregada neste trabalha serviu como redutor de dados para aplicação em árvores de decisão. Além disto, através das imagens geradas, pode-se observar a magnitude de uma dada variável climatológica em seu espaço geográfico.

Com as duas metodologias empregadas, pretende-se analisar outros fenômenos meteorológicos extremos. A ideia principal é automatizar os processos de KDD que até então foi feito passo a passo devido a complexidade dos dados, e à "descoberta de conhecimento" que se adquiriu durante o procedimento. Todo o processo foi amplamente e minuciosamente analisado tornando viável o aumento do banco de dados e a análise mais detalhada de parâmetros classificatórios.

Referências

Amaratunga, D. and Cabrera, J. (2004). *Exploration and analysis of DNA microarray and protein array data*. Wiley Interscience, New Jersey.

Dias, M. A. F. S. (2008). As chuvas de novembro de 2008 em santa catarina: um estudo de caso visando à melhoria do monitoramento e da previsão de eventos extremos. INPE, INMET, EPAGRI.

Eijkel, G. C. V. D. (1999). *Intelligent Data Analysis*. Springer-Verlag, New York.

Fayyad, U.; Piatesky-Shapiro, G. S. P. U. R. (1996). *Advances in Knowledge Discovery and Data Mining*. The MIT Press, California.

Han, J.; Kamner, M. (2001). *Data mining: concepts and techniques*. Morgan Kaufmann Publishers, San Francisco.

Hautaniemi, S. (2003). *Studies of microarray Data analysis with applications for human cancers*. Tampere University of Technology, Tampere, Finland.

IPCC (2007). *Cambio climático 2007: Informe de síntesis*. Grupo Intergubernamental de Expertos sobre el Cambio Climático [Equipo de redacción principal: Pachauri, R.K. y Reisinger, A. (directores de la publicación)], Ginebra, Suiza.

Marcelino, I. P. V. O., Nascimento, E. L., and Ferreira, N. J. (2005). Tornadoes in santa catarina state (southern brazil): event documentation, meteorological analysis and vulnerability assessment.

Marengo, J. A. (2009). Impactos de extremos relacionados com o tempo e o clima - impactos sosiais e econômicos. volume 8. Mudanças Climaticas - INPE.

Meira, C. A. A. (2008). Processo de descoberta de conhecimento em baes de dados para a análise e o alerta de doenças de culturas agrícolas e suas aplicações na ferrugem do cafeeiro.

Quinlan, J. R. (1993). *C4.5: programs for machine learning.* Morgan Kaufmann Publishers, San Francisco.

Ruivo, H. M. (2007). Análise integrada de dados ambientais utilizando técnicas de classificação e agrupamento de microarranjos de dna.

Shannon, C. E. (1948). A mathematical theory of communication. *Reprinter with corrections from The Bell System Technical Journal*, 27:379–423.

Simon, R. and Lam, A. P. (2006). *BRB-ArrayTools - version 3.4 - User's manual*. National Cancer Institute.

Witten, I. H. and Frank, E. (2005). *Data mining: practical machine learning tools and techniques.* Morgan Kaufmann Publishers, San Francisco.

Witten, I. H. and Frank, E. S. (2000). *Data mining: Practical machine learning tools and techniques with java implementation.* Morgan Kaufmann Publishers, California.