

Análise das secas na Amazônia utilizando técnica de mineração de dados e visualização em p-valores

Heloisa Musetti Ruivo_1¹, Fernando Manuel Ramos_2², Haroldo Fraga de Campos Velho_3²

¹Programa de Doutorado em Computação Aplicada – CAP
Instituto Nacional de Pesquisas Espaciais – INPE

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

Abstract. *The growing flood of data in the environmental sciences generates a bottleneck in its analysis and interpretation. This article shows that computational tools used to analyze large data sets of genome-wide studies can fruitfully be adapted and transposed to the environmental context. Here, we test this approach in the investigation of the causes of the Amazonian droughts of 2005 and 2010.*

Resumo. *O crescente aumento da quantidade de dados na área de ciências ambientais gera uma demanda de análise e interpretação. Este trabalho faz uso de dados meteorológicos e hidrológicos, e utiliza técnicas de mineração de dados na análise de variáveis climatológicas. O objetivo é apontar as variáveis mais relevantes em situações de secas severas para a região na Amazônia na primeira década do século XXI.*

Palavras-chave: *p-valor, mineração de dados, climatologia, classificação, análise estatística*

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 frequentes [IPCC 2007]. Observando as alterações climática ocorridas na última década na Amazônia, surge a preocupação na preservação da floresta. Sabe-se que a floresta Amazônica tem importante papel no clima do planeta, por várias razões, em particular devido ao grande estoque de carbono, tanto na biomassa como no solo, bem como o transporte de energia devido a intensa convecção registrada na região.

O progresso tecnológico ocorrido nas últimas décadas possibilitou que a grande quantidade de dados gerada aumentasse rapidamente. Com esse crescimento explosivo de dados armazenados, surgiu a necessidade do desenvolvimento de novas técnicas e ferramentas que pudessem transformar, de maneira inteligente e automática, os dados processados em informações úteis e em conhecimento. A mineração de dados surgiu da necessidade do emprego de técnicas e ferramentas que permitissem extrair informações de maneira automática de dados disponíveis.

Este trabalho fará uso dos dados armazenados e empregará uma técnica computacional de mineração e classificação de dados apontando as variáveis climatológicas responsáveis

por eventos climáticos extremos. A aplicação realizada investiga, quais foram os fatores climáticos associados as grandes secas ocorridas no Amazonas no século XXI. A metodologia de mineração de dados que será empregada no presente estudo classificará os dados através da ferramenta computacional BRB-ArrayTools [Simon and Lam 2006] e a visualização dos resultados, em campos de p-valores.

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, freqüentemente 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.

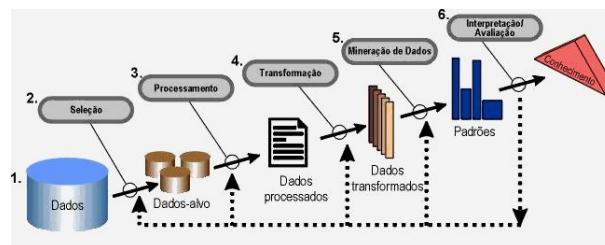


Figura 1. Visão geral das etapas que compõem o processo de KDD.

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 que será 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 utilizada neste trabalho provém da bioinformática. Trata-se do pacote **BRB-ArrayTools** versão 3.7.0 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 na página do 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 de uma amostra de células analisando a luminosidade de sinais fluorescentes. Hibridização é o processo bioquímico onde duas fitas de ácido nucléico com seqüências complementares se combinam.

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 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 intra-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].

3. Aplicação

Para esta análise foram utilizados dados climatológicos mensais que cobrem o período de janeiro de 1999 a janeiro de 2011 (145 meses). Os dados em grade foram extraídos do site <http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html> (40.500 dados). Trata-se de dados globais de reanálise com resolução espacial de $2.5^\circ \times 2.5^\circ$, com exceção da temperatura da superfície do mar, cuja resolução é de $2^\circ \times 2^\circ$. Da grade global selecionou-se uma subregião com coordenadas longitudinais 220 a 360, e latitudinais 40N a 40S. Todos os dados utilizados foram tabulados em termos de anomalias.

Fazendo uma analogia com a análise de MA, cada mês de dados representa um paciente, já cada variável climatológica, corresponde a um gene. Estendendo a analogia, a seca do Amazonas representa uma doença, e os fatores climáticos causadores do fenômeno, os genes reguladores ainda desconhecidos.

Para a análise as classes foram baseadas na precipitação média acumulada em uma região da Amazônia de coordenadas longitudinais 288 a 292, e latitudinais 8S a 4S. Estas foram divididas em 3 em relação a mediana da precipitação: baixa precipitação, média precipitação e alta precipitação. A classificação foi feita analisando-se baixa e média precipitação.

Os resultados são representados em campos de p-valores, ilustrados nas Figuras 2 e ???. Observa-se pelas imagens que regiões com tonalidades mais escuras compreendem os parâmetros que apresentam p-valor < 0.01 , ou seja, possuem uma probabilidade de 1% de ser um falso positivo. Em outras palavras, estas são as propriedades que se correlacionam com maior probabilidade, e potencialmente explicam as grandes secas na região da Amazônia analisada. Convém observar que [Marengo et al. 2008] e [Lewis et al. 2011] atribuem o aumento da temperatura da superfície do mar como um dos responsáveis pelas secas ocorridas, conforme se observa na Figura 2.

4. Conclusões

A crescente preocupação mundial com o meio ambiente é um fato real. A comunidade científica tem disponível dados de diversas origens para o estudo do impacto ambiental causado pela perturbação de fatores constituintes da atmosfera.

Neste trabalho, técnicas de classificação de dados, utilizadas atualmente na análise de experimentos de microarranjos de DNA, foram transpostas para a área ambiental. A aplicação ambiental investigou os fatores climáticos responsáveis pelas secas ocorridas na região Amazonica no século XXI. Para isso foram utilizados dados climatológicos disponíveis na internet pela comunidade científica. Resumidamente, os resultados obtidos

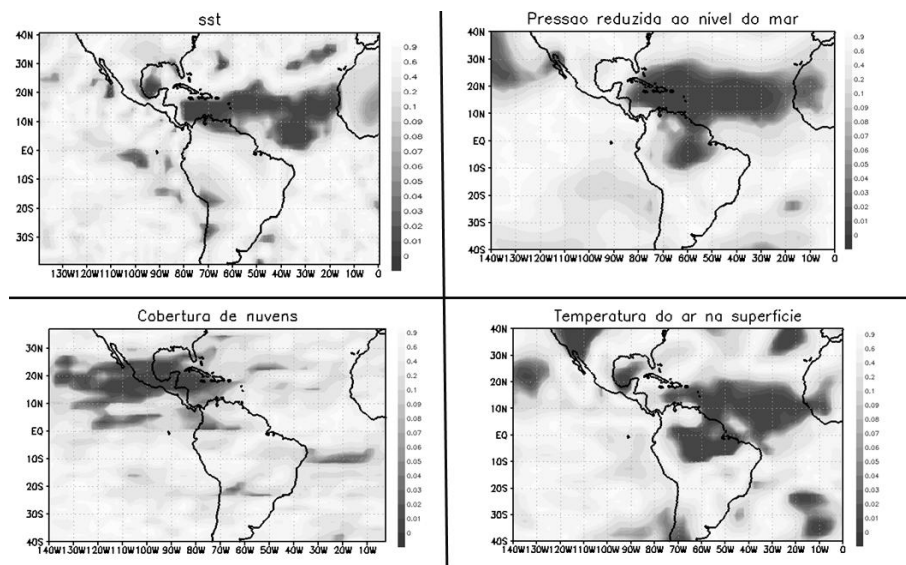


Figura 2. Representação em p-valores da influência das variáveis climatológicas nas secas ocorridas no Amazonas no século XXI.

indicam que as variações atmosféricas ocorridas no Atlântico Norte podem ser fatores relevantes na análise das secas em questão.

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.

Referências

- Amaratunga, D. and Cabrera, J. (2004). *Exploration and analysis of DNA microarray and protein array data*. Wiley Interscience, New Jersey.
- Fayyad, U.; Piatetsky-Shapiro, G. S. P. U. R. (1996). *Advances in Knowledge Discovery and Data Mining*. The MIT Press, California.
- 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.
- Lewis, S. L., Brando, P. M., and et al., O. L. P. (2011). The 2010 amazon drought. *Science*, 331(554–554).
- Marengo, J. A., Nobre, C. A., Tomasella, J., Oyama, M. D., Oliveira, G. S., Oliveira, R., Camargo, H., Alves, L. M., and Brown, I. F. (2008). The drought of amazonia in 2005. *Journal of Climate*, 21(3).
- Simon, R. and Lam, A. P. (2006). *BRB-ArrayTools - version 3.4 - User's manual*. National Cancer Institute.