Aplicabilidade de Técnica de Data Mining sobre séries temporais de dados de satélites artificiais

Yasuo Kono
DSS / INPE
yasuo@dss.inpe.br

José Carlos Becceneri
LAC / INPE
becce@lac.inpe.br

Rafael Santos
IBTA
santos@ieee.org

Resumo

Técnicas de Data Mining têm sido utilizadas para obter regras que descrevem os registros de um banco de dados. Para alguns tipos de banco de dados as regras podem ser muito complexas e genéricas ou pouco aplicáveis. Neste artigo apresentamos alguns resultados experimentais da extração e classificação de regras obtidas de bancos de dados meteorológicos usando determinadas métricas.

Palavras-chave: *data mining*, regras de decisão, *ACO*

1. Introdução

O Brasil é um país de dimensões continentais e de diversidades ambientais que são necessários meios para medir as condições climáticas, meteorológicas e ambientais.

Para se medir as condições acima, são utilizados atualmente, as PCDs – Plataformas de Coleta de Dados.
Por serem automatizadas e auto-suficientes, fazem a coleta dos dados continuamente. Essas informações, são transmitidas ao CPTEC (Centro de Previsão de Tempo e Estudos Climáticos) do INPE (Instituto Nacional de Pesquisas Espaciais) em Cachoeira Paulista / SP, utilizando-se dos satélites artificiais, da série SCD (Satélite de Coleta de Dados) e do CBERS (China Brazil Earth Resource Satellite). A utilização de satélites é fundamental, pois resumidamente, serve como um "espelho", retransmitindo os sinais enviados pelas PCDs para as antenas de recepção.

Operando atualmente três satélites (SCD1, SCD2 e CBERS-2), o INPE possui uma infra-estrutura complexa composta de: Centro de Controle e Rastreio de São José de Campos (SP); Estações de Recepção de Cuiabá (MT), de Alcântara (MA) e de Natal (Natal/RN); rede de comunicação de dados que interligam as unidades e equipes de operação especializadas para controlar os satélites. O CPTEC em Cachoeira Paulista/SP, não faz parte da operação dos satélites, porém o processamento dos dados é realizado nos seus supercomputadores de última geração.

A quantidade de informações geradas pelos PCDs é enorme. É possível identificar alguma anormalidade dentre as informações utilizando algoritmos preparados especificamente para isso. Porém, podem existir inconsistências só percebidas em uma série temporal, isto é, as informações isoladamente parecem estar dentro dos padrões esperados, mas no decorrer do tempo mostram-se incorretos. Vários fatores podem gerar essas incorreções, desde falhas nos sensores até mudanças ambientais causadas por diversos motivos.

Portanto, existe uma necessidade de verificar se é possível a utilização de técnicas de "Data Mining" para identificar essas incorreções. Este artigo descreve a metodologia ACO (Ant Colony Optimization) para aplicação em Data Mining, ao invés das tradicionais metodologias atuais.

2. Mineração de Dados

Mineração de dados ou Data Mining é a extração não-trivial de informação implícita (nova ou previamente desconhecida) e útil a partir de bases de dados. [13]

O grande volume de dados disponíveis cresce a cada dia e desafia a nossa capacidade de armazenamento, seleção e uso. Esta tecnologia com suas ferramentas permitem a "mineração" destes dados a fim de gerar um real valor do dado transformando-o em informação e conhecimento.

É formada por um conjunto de ferramentas que através do uso de algoritmos de aprendizado ou baseados em redes neurais e estatísticas, são capazes de explorar um grande conjunto de dados, extraindo destes, conhecimento na forma de hipóteses e de regras. [14]
Basicamente "Data Mining" preocupa-se com a análise de dados e o uso de técnicas de *software* na procura de padrões em conjuntos de dados. É o computador que é responsável por procurar padrões, identificando as regras subjacentes nos dados.

Uma definição concisa que podemos adotar de *Data Mining* é "a pesquisa por informações valiosas em grandes volumes de dados" [11]. Essa definição aplica-se bem ao contexto deste trabalho: uma grande massa de dados (gigabytes, com crescimento mensal de cerca de 40 megabytes) com muitas informações potencialmente valiosas não exploradas até o início desta nossa pesquisa.

Dois dos objetivos que podem ser atingidos com *Data Mining* são predizer comportamentos futuros para tomada de decisões e descobrir padrões previamente desconhecidos de comportamentos. O interesse é no segundo objetivo: usaremos técnicas de *Data Mining* para extrair informações

desta massa de dados para tentar localizar exceções a regras aplicáveis aos dados. Estas exceções podem possivelmente caracterizar incoerências nos dados e na forma de coleta dos mesmos.

Diversas técnicas computacionais podem ser empregadas na pesquisa por informações valiosas. Segundo Thearling [9], as técnicas mais comumente usadas de *Data Mining* são:

1. Redes Neurais Artificiais [4]
2. Árvores de Decisões [8]
3. Algoritmos Genéticos [3]
4. Método do Vizinho Mais Próximo [6]
5. Regras de Indução [2]

Segundo [7], uma técnica que ainda é um campo de pesquisa inexplorado é a técnica *Ant Colony* [1]. Uma continuidade deste trabalho tentará empregar tal técnica para uma análise mais detalhada dos dados em questão.

3. Ant Colony Optimization

Ant Colony Optimization (ACO) é um paradigma para desenvolver algoritmos metaheurísticos para problemas de otimização combinatória. A peculiaridade essencial de um algoritmo ACO é a combinação da informação anterior da estrutura de uma solução promissora com a informação posterior da estrutura dos bons resultados obtidos previamente.[15]

A característica dos algoritmos ACO é o uso explícito de elementos de soluções anteriores. Ele direciona a construção de soluções de baixo-nível, como GRASP[16] realiza, mas incluindo em uma população de *framework* e randomizando a construção no modo Monte Carlo.

ACO[17, 18] é uma classe de algoritmos cujo primeiro membro, chamado Ant System, foi inicialmente proposto por Colorni, Dorigo e Maniezzo [19]. A idéia principal foi inspirada no comportamento de formigas reais, e em uma busca paralela sobre inúmeras *threads* computacionais baseadas em um problema de dados local e em uma estrutura dinâmica de memória contendo informações sobre qualidade dos resultados obtidos anteriormente.

Um algoritmo ACO é essencialmente um sistema baseado em agentes que simula o comportamento natural de formigas, incluindo o mecanismo de cooperação e adaptação. Em [21] o uso deste tipo de sistema como nova metaheurística foi proposto para solucionar problemas de otimização combinatória, e tem-se mostrado tanto robusto quanto versátil, em aplicações de leques diferentes de problemas de OC.

As idéias de ACO são baseadas em :

1. Cada caminho seguido por uma formiga é associada com uma solução candidata para um dado problema.
2. Quando uma formiga segue o caminho, a quantidade de feromônio depositado naquele caminho é proporcional à qualidade da solução candidata correspondente para o problema especificado.
3. Quando a formiga tem que escolher entre dois ou mais caminhos, os caminhos que possuem mais feromônio têm mais probabilidades de serem escolhidos pela formiga.

Isto resulta que as formigas convergem eventualmente para o caminho mais curto, significando em uma ótima ou próximo da solução ótima para o problema proposto, como explicado anteriormente no caso das formigas reais. Na essência, o projeto de um algoritmo ACO envolve as especificações de [22]:

1. Uma representação apropriada do problema, permite às formigas construir/modificar soluções de forma incremental durante a utilização de uma regra de transição probabilística, baseada em quantidades de feromônios na trilha e em um local, heurística que depende do problema
2. Um método que imponha a construção de soluções válidas, ou seja, soluções válidas às situações do mundo real que correspondam à definição do problema
3. Uma função heurística dependente do problema que possa medir a qualidade dos itens que possam ser agregados à solução parcial do problema
4. Uma regra para atualizar o feromônio, que especifica o quanto a trilha sofre alteração da quantidade de feromônio

Formigas artificiais contem características muito similares às formigas reais:

1. Formigas artificiais tem uma preferência probabilística por caminhos com grande quantidade de feromônio
2. Caminhos curtos tendem a ter maiores taxas de incremento em sua quantidade de feromônio
3. As formigas utilizam um sistema de comunicação indireta baseadas em quantidade de feromônio depositado em cada caminho.

4. Dados utilizados

Os dados utilizados são originados pelas Plataformas de Coletas de Dados (PCD), que após transmissão ao satélite que estiver ao alcance, são retransmitidos às antenas receptoras, pré-processados em rotinas chamadas de calibração, e finalmente armazenados e disponibilizados aos usuários/clientes.

O PCD [15] é uma estrutura composta de vários sensores que registram informações meteorológicas (temperatura, pressão, direção e velocidade do vento, umidade, etc). Opera autonomamente por possuir baterias e dispositivos de energia solar e por utiliza satélites para transmitir informações ao destino final no Centro de Missão de Coleta de Dados em Cachoeira Paulista (SP).

As informações coletadas são gerados pelos sensores: Temperatura do Ar, Temperatura Máxima do Ar últimas 24 H, Temperatura Mínima do Ar últimas 24 H, Umidade Relativa do Ar, Pressão Barométrica, Velocidade do Vento, Direção do Vento, Velocidade Máxima do Vento (Rajada), Direção do Vento na Velocidade Máxima, Radiação Solar Global, Radiação Solar Líquida, Precipitação Acumulada, Temperatura do Solo 100mm, 200mm, 400mm, Conteúdo Água no Solo 100mm, 200mm, 400mm, Fluxo de Calor no Solo. Sobre esses dados, utilizamos a criação de árvores de decisão, a extração de regras de classificação a partir das árvores de decisão e a análise de algumas métricas aplicáveis a estas regras. As métricas utilizadas no trabalho estão definidas em [5]:

1. **Acurácia** ou **Confiança**: esta medida corresponde ao percentual dos registros para os quais a predição da regra está correta, tomado sobre o total de registros para os quais o antecedente é aplicável e correto.
2. **Aplicabilidade**: esta medida representa o percentual de registros no banco de dados que podem ser avaliados por esta regra, ou seja, o percentual de registros para os quais o antecedente da regra é avaliado como sendo verdadeiro.
3. **Suporte**: esta medida, numericamente igual à aplicabilidade multiplicada pela acurácia, corresponde ao percentual de registros que são classificados corretamente quanto ao antecedente **e** ao conseqüente, em relação a todos os registros do banco de dados. Esta medida é freqüentemente usada com a medida de acurácia ou confiança para estabelecer a qualidade das regras individuais.
4. **Cobertura**: esta medida corresponde ao percentual dos registros que é classificado corretamente pela regra.
5. **Acurácia Padrão**: definida como a proporção dos registros do banco de dados que podem ser classificados com aquele conseqüente.

5. Resultados Esperados

Os resultados esperados são: a aplicabilidade das técnicas de *Data Mining* sobre o conjunto de dados oriundos das PCDs e retransmitidos pelos satélites da série SCD; aplicabilidade das métricas a base de dados do SCD1 e utilização do ACO para obter regras descritivas dos fenômenos meteorológicos.

Uma vez obtida alguma regra de decisão, observaremos seus antecedentes e seus conseqüentes. A verificação de alguns casos reais em que a regra for aplicada corretamente informará a aplicabilidade destas técnicas, comprovando sua utilização nestes tipos de dados.

Esperamos com isso conseguir qualificar as informações obtidas dos dados, permitindo assim, ter uma qualidade de informação melhor e a utilização mais racional destas.

6. Referências

[1] Bonabeau, E; Dorigo, M; Theraulaz, G. Swarm Intelligence, From Natural to Artificial Systems. Oxford University Press, 1999.

[2] Carvalho, D.R. *Data Mining através de indução de Regras e Algoritmos Genéticos*. Dissertação de Mestrado em Informática Aplicada, PUCPR, PR, 1999.

[3] Goldberg, D.E. *Genetic Algorithms in Search, Optimization and Machine Learning*. Boston: Addison-Wesley, 1989.

[4] Haykin, S.S. *Redes Neurais: Princípios e Prática*, Editora Bookman Companhia, 2000.

[5] de la Iglesia, B. *Mining rules from a database according to multiple measures of interest* in MOMH Multiple Objective Metaheuristic Workshop, Paris, 2002.

[6] Keller, J. M.; Gray, M. R.; Givens Jr., J. A. *A Fuzzy K-Nearest Neighbor Algorithm*, in "Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data", edited by J.C.Bezdek and S. K. Pal, IEEE Press, Piscataway, NJ, 258-264, 1992.

[7] Parpinelli, R; Lopes, H.S; Freitas, A.A. *Data Mining with an Ant Colony Optimization Algorithm*. http://www.ppgia.pucpr.br/~alex/pub_papers.dir/Ant-IEEE-TEC.pdf.

[8] Quinlan, J. R. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA. 1993

[9] Thearling, K. *An Introduction to Data Mining*. http://www.thearling.com.

[11]Weiss, S.M.; Indurkhya N. *Predictive Data Mining, A Practical Guide*. Morgan Kaufmann Publishers, Inc. San Francisco, California, 1998.

[12] Witten, I.H.; Frank, E. *Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, Inc. San Francisco, California, 2000.

[13] http://atlas.ucpel.tche.br/~loh/dm.htm

[14] http://www.utp.br/informacao/

[15] http://www.idsia.ch/~luca/aco2004.pdf

[16] T.A. Feo and M.G.C. Resende, *Greedy randomized adaptive search procedures*, Journal of Global Optimization 6, 1995, 109-133

[17] Dorigo, M. Ant colony optimization web page, http://www.iridiaulb.ac.be/mdorigo/ACO/ACO.html

[18] Dorigo, M. Di Caro, G. & Gambardella, L.M. *Ant Algorithms for Discrete Optimization*, Artificial Life, 5(2):137-172, 1999

[19] A. Colorni, M. Dorigo, and V. Maniezzo, *Distributed optimization by ant colonies*, Proceedings of ECAL'91, European Conference on Artificial Life, Elsevier Publishing, Amsterdam, 1991.

[20] http://www.ppgia.pucpr.br/~alex/pub_papers.dir/Ant-IEEE-TEC.pdf

[21] M. Dorigo and G. Di Caro, *The ant colony optimization meta-heuristic*, In: *New Ideas in Optimization*, D. Corne, M. Dorigo and F. Glover Eds. London, UK: McGraw Hill, pp. 11-32, 1999.

[22] E. Bonabeau, M. Dorigo and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*. New York, NY: Oxford University Press, 1999.