Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brasil, 21-26 abril 2007, INPE, p. 2871-2873.

Uso de Regras de Associação e Conjuntos Fuzzy na Mineração de Dados Sócio-Ambientais da Base de Dados Integrada (BDI) PIATAM/SIPAM

Fábio Roque da Silva Moreira [1]
Alexandre Gustavo Evsukoff [2]
Fernando Pellon de Miranda [3]

[1] Centro Brasileiro de Recursos Radarsat – CBRR/COPPE/UFRJ
Caixa Postal 68552 – Rio de Janeiro - RJ, Brasil
fmoreira@cbrr.coppe.ufrj.br

[2] Universidade Federal do Rio de Janeiro – UFRJ/COPPE
Caixa Postal 96 - 13416-000 – Rio de Janeiro - RJ, Brasil
alexandre.evsukoff@coc.ufrj.br

Petrobras – CENPES
[3] Ilha do Fundão, Qd 7, Prédio 20 – C. Universitária,
21949-900, Rio de Janeiro, RJ, Brasil

Abstract. This present work intends to acquire knowledge from the analysis of the social and environmental data of the PIATAM/SIPAM integrated database, employing a data mining model based in fuzzy association rules, where both binary (features) and quantitative (numerical) attributes of the database are mapped to fuzzy sets and the association rules are express as quantified sentences. Initially, a version of the algorithm proposed by Hu et al. (2003) have been programmed and tested. The algorithm consists of two parts: one to generate the large fuzzy grids, and the other to generate the fuzzy association rules. Preliminary results obtained from the test database demonstrated the effectiveness of the algorithm. However, as the PIATAM/SIPAM integrated database also presents spatio-temporal properties, further developments in the algorithm must be implemented to allow it recognized relationship patterns among attributes, determined by spatio-temporal characteristics.

Palavras-chave: *Data mining*, regras de associação, *lógica fuzzy*, e Amazônia Central.

1. Introdução

A integração de dados ambientais na Base de Dados Integrada (BDI) PIATAM/SIPAM sinaliza como uma fonte de informação e conhecimento para estudos científicos, de monitoramento, ou como input em processos de tomada de decisão, relacionados com as atividades de transporte de óleo e gás no trecho do Rio Solimões entre Coari e Manaus, Amazônia Central. Um modo de extrair esse conhecimento é através da aplicação de técnicas de mineração de dados (*data mining*) que constituem ótimas ferramentas para extração de conhecimento de massivos banco de dados.

Dentre as diversas técnicas de *data mining* a Regra de Associação, proposta primeiramente por Agrawal et al. (1993), aparece em destaque, sendo vastamente empregada em estudos de *data mining* . Os motivos dessa preferência estão relacionados à facilidade de compreensão da teoria e à alta capacidade de manipulação de bancos de dados com massivas quantidades de informação. Regras de associação fornecem meios de identificar e representar a dependência entre valores de atributos de objetos armazenados em banco de dados (Dubois et al., 2006).

Como exemplo, considere a regra $A \Rightarrow B$, obtida de um banco de dados comercial. O significado expresso pela regra é que se o produto A for comprado, então o produto B deve ser adquirido na mesma transação. Uma regra de associação é considerada interessante quando ela satisfaz um suporte mínimo e uma confiança mínima. Esses dois parâmetros

correspondem às medidas básicas de significância de uma regra de associação. O problema básico na mineração das regras de associação é gerar todas as regras de associação $A \Rightarrow B$ que tenham suporte e confiança maiores do que o limiar especificado pelo usuário (De Cock et al., 2005).

Por outro lado, tipicamente regras de associação apresentam uma abordagem *booleana* do problema, onde importa apenas se o item está presente ("1") ou não ("0") na transação, sendo desconsiderado os atributos quantitativos. Entretanto, normalmente banco de dados não são restritos a atributos binários, mas também contém atributos com valores numéricos. A BDI PIATAM/SIPAM se enquadra nessa condição uma vez que grande parte de suas informações são variáveis quantitativas.

Inicialmente, atributos numéricos eram representados por regras de associação quantitativas por meio de subconjuntos, tipicamente na forma de intervalos. Por exemplo, "Empregados com idades entre 30 e 40 anos tem salários entre $50.000 e $70.000". Entretanto, essa abordagem pode gerar certos efeitos indesejados relacionados aos limiares adotados que delimitam os subconjuntos (Sudcamp, 1995).

O uso de conjuntos *fuzzy* (Zadeh, 1965) combinado com regras de associação pode evitar esse tipo de problema, uma vez que, as fronteiras são definidas por intervalos graduais (*fuzzy*) ao invés de rígidos (*crisp*). A idéia é que os conjuntos *fuzzy* atuem como uma interface entre a escala numérica e uma escala simbólica (i.e. termos linguísticos), onde as regras descobertas são apresentadas de um modo linguístico e por consequência mais compreensivo e amigável (Dubois et al., 2006) (e.g. Empregados de meia-idade recebem salários consideráveis).

O algoritmo proposto por Hu et al. (2003) emprega de modo inteligente os conceitos das duas técnicas, regras de associação e conjuntos *fuzzy*, com a vantagem de precisar percorrer a base de dados apenas uma única vez para gerar as regras. Basicamente, o algoritmo é composto por duas fases. A primeira gera os grids *fuzzy* com suportes maiores do que o suporte mínimo especificado, e a segunda gera as regras de associação a partir desses grids.

Preliminarmente, um algoritmo baseado em Hu et al. (2003) foi programado e testado com o mesmo exemplo numérico utilizado pelo autor. A reprodução dos mesmos resultados obtidos por Hu et al. (2003) atesta o correto funcionamento do algoritmo que foi também aplicado aos dados de Índice de Picada por Homem/Hora (IPHH) do tema malária do projeto PIATAM. O Projeto PIATAM (Potenciais Impactos Ambientais no Transporte Fluvial de Gás Natural e Petróleo na Amazônia), idealizado para minimizar possíveis efeitos relacionados com as atividades da industria do petróleo na área de transporte de óleo e gás no trecho do Rio Solimões entre Urucu-Coari-Manaus, dispõe de dados de natureza ambiental e sócio-econômica, coletados em campo periodicamente (ciclos hidrológicos) por pesquisadores de diferentes áreas, assim como arquivos de imagens individuais de satélites e aeroportadas, além de mosaicos georreferenciados.

Os dados do tema malária foram coletados durante nove excursões realizadas entre os meses de Março/2004 e Abril//2006 nos diferentes períodos hidrológicos da bacia amazônica. Cada excursão levou em média 12 dias, onde 11 pontos (comunidades) foram visitados ao longo do rio Solimões. No total 96 coletas foram realizadas ao redor das habitações das comunidades no período das 18:00 às 22:00 horas, sendo um total de 52000 mosquitos de 29 diferentes espécies capturados. Entretanto, no presente estudo foram analisadas, pelo algoritmo, apenas as cinco espécies mais numerosas (Anopheles triannulatus, Culex (Mel.) sp., Mansonia amazonensis, Mansonia titillans e Mansonia humeralis) que juntas totalizam 97,0% dos indivíduos capturados. As cinco espécies foram particionadas em cinco conjuntos *fuzzy* (baixa, relativamente baixa, média, relativamente alta e alta quantidade de mosquitos) que juntos totalizando 25 variáveis linguísticas. Essas 25 variáveis foram analisadas pelo algoritmo de regras de associação *fuzzy* segundo diferentes limiares de suporte e confiança.

Embora regras de associação tenham sido obtidas (e.g. Se quantidade de Anopheles triannulatus é alta, Então a quantidade de Culex (Mel.) sp. é média, com suporte = 0.2764 e confiança = 0.6624; e Se quantidade de Anopheles triannulatus é alta, então quantidade de Mansonia amazonensis é baixa, com suporte = 0.3193 e confianca = 0.7653), sinalizando diferentes dependências que possam existir entre as quantidades de espécies de mosquitos, os resultados não foram suficientemente conclusivos.

Futuras discussões entre os especialistas do tema malária devem ser realizadas de modo que sejam definidas partições *fuzzy* com maiores significados semânticos, ou mesmo que sejam sugeridas para análise outras informações levantadas pela base de dados integrada PIATAM/SIPAM. Adicionalmente, avanços devem ser implementados no algoritmo de modo que este consiga apontar regras que tenham dependências em características espaço–temporais intrinsicas do dado. Isso se faz necessário, uma vez que as informações coletadas pelo projeto PIATAM apresentam como principal característica propriedades espaço-temporais, ou seja, as informações são coletadas em uma determindada localidade (i.e. espaço), e em uma determinada data ou período hidrológico (tempo).

Referências

Agrawal R, Imielinski T, and Swami A Mining association rules between sets of items in large databases. In **Proceedings...** The ACM SIGMOD International Conference on Management of Data, 1993,. p. 207–216.

De Cock, M.; Cornelis, C.; Kerre E. E. Elicitation of fuzzy association rules from positive and negative examples. **Fuzzy Sets and Systems,** n.149, p. 73-85, 2005

Dubois, D.; Hüllermeir, E.; Prade, H. A systematic approach to the assessment of fuzzy association rules**. Data Min. Knowl. Disc.**, n.13, p. 167-192, 2006.

Hu, Y-C.; Chen, R-S; , Tzeng, G-H. Discovering fuzzy association rules using fuzzy partition methods. **Knowledge-Based Systems,** n. 16, p. 137–147, 2003.

Sudkamp, T. Examples, counterexamples, and measuring fuzzy associations. **Fuzzy Sets and Systems**, n.149, p. 57–71, 2005.

Zadeh, L.A. Fuzzy sets. **Information and Control**, v.8, n.3, p. 338-353, 1965.