ANÁLISE DE REGRAS DE ASSOCIAÇÃO APLICADAS A VARIÁVEIS QUANTITATIVAS

Luis Eduardo dos Santos (PIC/Uem), Josmar Mazucheli (Orientador), e-mail: jmazucheli@gmail.com.

Universidade Estadual de Maringá / Centro de Ciências Exatas/Maringá, PR.

Ciências Exatas e da Terra / Probabilidade e Estatística.

Palavras-chave: MBA, estatística descritiva, mineração de dados.

Resumo:

O presente trabalho é resultado de uma pesquisa à respeito da análise de regras de associação em grandes conjuntos de dados contendo variáveis quantitativas. O objetivo é introduzir este conceito por meio de uma aplicação que permita exemplificar sua utilização em problemas desta natureza. Foram utilizados dados climatológicos referentes a marcações diárias realizadas na estação de Macapá – AP entre os anos de 1961 e 2014, para a obtenção de regras que permitam caracterizar o comportamento climático do local quando ocorre precipitação, levando em consideração que a grande presença de chuva em diferentes épocas do ano possibilita a utilização deste procedimento. O método escolhido na análise em questão foi o algoritmo Apriori.

Introdução

No estudo de diversos problemas envolvendo análise de dados, é de interesse bastante comum buscar padrões de comportamento entre as variáveis, que não podem ser identificados diretamente. Em virtude do crescimento constante da capacidade de coleta e armazenamento, em conjunto com o aumento da demanda por informação, os conjuntos de dados tornam-se mais volumosos ao longo do tempo, dificultando a detecção destes padrões e exigindo técnicas apropriadas para manipulação e retirada de informações.

Uma das principais técnicas para obter informações interessantes de grandes bases de dados é conhecida como Market Basket Analisys (MBA), que consiste em detectar regras de associação entre um conjunto de observações provenientes de um banco de transações.

Apesar de a teoria inicial ter sido fundamentada considerando dados transacionais, em que os elementos do banco são itens, ou seja, conjuntos contendo variáveis discretas e/ou categóricas, esta metodologia pode ser utilizada em casos que envolvam também variáveis quantitativas.

Encontrar características climáticas relacionadas à ocorrência de chuva em um local específico pode ser possível através de regras de associação, permitindo descrever o comportamento do fenômeno baseado em fatores que o influenciam.

Materiais e métodos

O banco de dados contém marcações diárias referentes ao clima da cidade de Macapá – AP, coletadas em 16655 dias entre 1961 e 2014. As variáveis observadas são precipitação (mm), temperatura máxima e mínima (ºC), insolação (h), umidade relativa média (%) e velocidade do vento média (m/s).

A precipitação foi categorizada como uma variável binária que indica a ocorrência de chuva (>0mm) e a não ocorrência (=0mm). As demais variáveis foram categorizadas em três níveis, sendo eles baixo, médio e alto, por meio de métodos não supervisionados (SIVARAMAKRISHNAN; MEGANATHAN, 2011). A ideia é construir categorias nas variáveis quantitativas considerando intervalos de valores numéricos (AUMANN; LINDELL, 2003) para que seja possível aplicar o procedimento. Para a detecção das regras, foi utilizado o algoritmo Apriori (AGRAWAL; SRIKANT, 1994) implementado no ambiente R (R DEVELOPMENT CORE TEAM, 2016).

Uma regra de associação é uma expressão na forma $X \Rightarrow Y$, que significa que as transações que contém X tendem a conter Y, sendo eles conjuntos de itens (AGRAWAL; MANNILA; SRIKANT et al, 1996). No caso aqui discutido, estamos considerando que as transações são os dias que foram realizadas as marcações, o conjunto de itens X é composto pelos três níveis de cada variável correlacionada à precipitação e o conjunto Y representa a resposta positiva quanto à ocorrência de chuva.

As regras foram selecionadas considerando um suporte mínimo, que é definido como a fração de transações que satisfazem a união da regra e

avaliadas de acordo com o coeficiente de confiança, que representa a proporção de transações que satisfazem *Y* dado que *X* é satisfeito (AGRAWAL; IMIELISNKI et al, 1993).

Resultados e Discussão

Ao aplicar o algoritmo Apriori no conjunto de dados categorizados, as regras referentes à ocorrência de precipitação foram detectadas e as mais relevantes foram selecionadas com base no suporte maior ou igual a 15%. As regras selecionadas apresentaram o coeficiente de confiança em torno de 80%.

As regras selecionadas destacaram quatro características climáticas que mais foram observadas nos dias chuvosos, dentre as quinze analisadas, sendo elas o nível alto na umidade relativa média, nível baixo de insolação, nível baixo de temperatura máxima e nível baixo de velocidade do vento média.

Além da presença individual destas características nas regras, combinações destas também foram selecionadas no procedimento, deixando mais evidente o padrão de associação. As regras selecionadas são apresentadas na Tabela 1.

Tabela 1 Regras de associação com suporte e confiança

Regra	Suporte (%)	Confiança (%)
URM = Alto ⇒ PRC = Sim	27,94	84,61
INS = Baixo ⇒ PRC = Sim	27,61	82,3
TM = Baixo ⇒ PRC = Sim	28,6	84,44
VVM = Baixo ⇒ PRC = Sim	25,67	72,06
INS = Baixo, URM = Alto ⇒ PRC = Sim	21,74	86,66
TM = Baixo, URM = Alto ⇒ PRC = Sim	21,93	87,92
URM = Alto, VVM = Baixo ⇒ PRC = Sim	16,39	85,96
TM = Baixo, INS = Baixo ⇒ PRC = Sim	22,3	87,18
INS = Baixo, VVM = Baixo ⇒ PRC = Sim	15,36	84,59
TM = Baixo, VVM = Baixo ⇒ PRC = Sim	15,1	87,44
TM = Baixo, INS = Baixo, URM = Baixo ⇒ PRC = Sim	18,84	88,4

URM = umidade relativa média, INS = insolação, TM = temperatura máxima, VVM = velocidade do vento média.

Conclusões

A partir das regras detectadas, foi possível identificar as principais características associadas à ocorrência de chuva no local, que é uma informação relevante no estudo de dados desta natureza.
Desta forma, a aplicação desenvolvida pôde mostrar a utilidade das regras de associação como ferramenta para análise descritiva de dados quando o problema em questão envolve um grande conjunto de dados.

Referências

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. **Mining association rules between sets of items in large databases**. ACM SIGMOD Record, ACM, v. 22, n. 2, p. 207–216, 1993.

AGRAWAL, R. et al. **Fast discovery of association rules**. Advances in knowledge discovery and data mining, AAAI/MIT Press, v. 12, n. 1, p. 307–328, 1996.

AGRAWAL, R.; SRIKANT, R. et al. **Fast algorithms for mining association rules**. Proc. 20th int. conf. very large data bases, VLDB, v. 1215, p. 487–499, 1994.

AUMANN, Y.; LINDELL, Y. **A statistical theory for quantitative association rules.** Journal of Intelligent Information Systems, v. 20, p. 255 – 283, 2003.

R DEVELOPMENT CORE TEAM. R: **A language and environment for statistical computing**. Versão 3.3.0. Vienna: R Foundation for Statistical Computing, 2016. Disponível em: <http://www.r-project.org>. Acesso em: 25 jul. 2016.

SIVARAMAKRISHNAN, T.; MEGANATHAN, S. **Association rule mining and classifier approach for quantitative spot rainfall prediction.** Journal of Theoretical and Applied Information Technology, v. 34, n. 2, p. 173–177, 2011.