Comparação de métodos estatísticos para avaliação de relações causa-e-efeito aplicados a um banco de dados

Bráulia A. A. Perázio
Depto de Informática, CEE, UFV
36570-000, Viçosa, MG
E-mail: bperazio@bol.com.br

Mercio Botelho Faria
Depto de Matemática, CEE, UFV
36570-000, Viçosa, MG
E-mail: mercio@gmail.com

RESUMO

Os constantes avanços na área da Tecnologia da Informação têm viabilizado o armazenamento de grandes e múltiplas bases de dados. Sistemas gerenciadores de banco de dados, leitores de códigos de barras, dispositivos de memória secundária de maior capacidade de armazenamento e sistemas de informação em geral são exemplos de recursos que têm viabilizado a proliferação de inúmeras bases de dados de natureza comercial, administrativa, governamental e científica (GOLDSCHMIDT,2005). Surge então a necessidade da análise desses grandes bancos de dados, a fim de extrair informação e conhecimento para serem utilizados na solução dos diversos problemas.

Em muitas situações, o objetivo principal tem sido estudar a relação de dependência de variáveis em relação a diversas outras variáveis cabíveis de controle e, portanto, capazes de serem manipuladas e propiciarem valores mais adequados às primeiras.

Quando se avalia uma característica de interesse econômico, por exemplo,verifica-se muitas vezes que ela é dependente de vários fatores, por isso a necessidade de se utilizar um método estatístico que estime com qualidade esta relação de causa e efeito. Na estatística os métodos de análise de regressão auxiliam na obtenção desta relação. Entre eles são muito usada a análise de regressão linear múltipla, quando a resposta é quantitativa. E quando o objetivo é a classificação destes vários fatores: a análise discriminante.

Paralelamente, essas análises podem ser realizadas por meio das árvores de decisão, divididas em árvores de classificação para variáveis aleatórias discretas e de regressão para variáveis aleatórias contínuas.

Os classificadores baseados em árvores de decisão remontam aos anos 50, sendo uma referência o trabalho de Hunt, onde vários trabalhos de indução são apresentados(HUNT, 1966). Posteriormente, talvez o trabalho mais importante seja o extraordinário livro de Breiman, Friedman, Olshen e Stone, em que o algoritmo CART é apresentado (BREIMAN,1984). Também o trabalho de Quinlan em 1986 teve grande aceitação nesta área científica tendo servido de inspiração a muitos dos sistemas posteriormente apresentados em particular ao seu trabalho C4.5(QUINLAN,1993).

Nossos objetivos neste trabalho são: 1) Verificar a viabilidade da análise de dados por meio das árvores de decisão; 2) Comparar as estimativas obtidas por meio da árvore de classificação, e análise discriminante; 3) Comparar a eficiência das estimativas obtidas por meio da árvore de regressão e regressão linear múltipla;

Para isso utilizaremos o software R, a fim de simularmos diferentes conjuntos de dados compostos por duas variáveis (X1 e X2), com 400 observações, sob diferentes situações de relações entre as mesmas, a partir da distribuição normal. Serão simulados também os efeitos

normais do erro experimental. Para cada simulação serão realizadas 3 repetições. Serão usadas como medida de comparação o erro médio quadrático(EQM) e o erro percentual médio absoluto (MAPE).

Contudo, pretendemos verificar , através do banco de dados simulado, como mencionado acima, que o método de árvores de decisão é mais eficiente quando temos um grande conjunto de dados, quando comparado à análise discriminante e regressão linear múltipla.

Palavras-chave*: árvores de decisão, análise discriminante, regressão linear múltipla, simulação.*

Referências

[1] Breiman, L., Friedman, J. H., Olshen, R. A. e Stone, C. J. (1984). Classification and Regression Trees. Belmont, CA: Wadsworth.

[2] Hunt, E. B., Marin, J., & Stone, P. J. Experiments in induction. New York: Academic Press. 1966.

[3] Quinlan, J.R. (1986). Induction of Decision Trees. Machine Learning 1,1, pag. 81-106.

[4] Quinlan, J. R. (1987). Simplifying Decision Trees. Int. Journal of Man-Machine

[5] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo, CA:Morgan Kaufman.

[6] Shapiro, P. G. Knowledge discovery in real databases: A report on the IJCAI-89 Workshop. AI Magazine, v. 11, n. 5, Jan. 1991, Special issue, p.68-70.

[7] Goldschmidt, R. e Passos. Data Mining: um guia prático. Rio de Janeiro: Elsevier, 2005.