UM ALGORITMO DE AGRUPAMENTO APLICADO À MANIPULAÇÃO DE BASES DE DADOS DE GRANDE PORTE

Adriel Costa Maia
Escola Nacional de Ciências Estatísticas
Rua André Cavalcanti, 106, Centro – Rio de Janeiro – RJ.
adrielcmaia@gmail.com

José André de Moura Brito
Escola Nacional de Ciências Estatísticas
André Cavalcanti, 106, sala 406, Centro – Rio de Janeiro – RJ.
jambrito@gmail.com

Em função do grande avanço tecnológico observado nas últimas décadas, e com desenvolvimento dos meios de comunicação, atualmente são produzidas, diariamente, milhares de informações, dando origem ao desafio de manipular bases de dados cada vez maiores e extrair informações úteis. Tal fato tem demandado, com frequência, a necessidade do desenvolvimento de novas técnicas voltadas à análise de dados.

Essas técnicas, por sua vez, possibilitam a manipulação de bases de dados de grande porte, produzindo soluções de boa qualidade em um tempo computacional factível. Algo importante, tendo em vista que muitas técnicas tradicionais possuem limitações em relação à quantidade de registros (ou objetos) pertencentes à base. Neste sentido, a mineração de dados é uma tecnologia que agrega várias técnicas tradicionais de análise de dados, com algoritmos sofisticados e que são utilizados para processar grandes volumes de dados. Em particular, no presente trabalho, foi estudada a técnica de análise de agrupamentos, que agrega um conjunto de métodos que são aplicados às bases de dados com o objetivo de dividir os registros em grupos homogêneos, considerando a utilização de alguma métrica. No presente projeto foram estudados, inicialmente, vários métodos agrupamento, optando-se pela utilização de um método de agrupamento que resolve o problema dos k-medoids. Neste problema, dado um conjunto de n objetos com f atributos, e fixado o número k de grupos, deve-se selecionar, dentre os n objetos, k objetos denominados medoids. Os (n-k) objetos restantes são alocados ao medoid correspondente mais próximo, segundo uma medida distância. Em análise de agrupamento há dois métodos que resolvem tal problema, quais sejam: PAM e CLARA. A partir do estudo desses métodos e da metaheurística algoritmos genéticos de chaves aleatórias viciadas, desenvolveu-se um novo método de agrupamento para o problema dos k-medoids.

O objetivo é aplicar tal método em n amostras selecionadas de uma base de dados, considerando que cada amostra corresponde a subconjunto de registros da base. Mais especificamente, aplica-se o método em cada amostra, são definidos os grupos e os objetos que não foram considerados em cada amostra são alocados ao seu medoid mais próximo. Dessa forma, são produzidas n soluções, tomando-se a solução com menor valor da função objetivo correspondente ao medoid. Procedimento similar a este é utilizado no método CLARA. Não obstante, com a utilização do novo método combinado com o procedimento de seleção de várias amostras, foi possível trabalhar com bases de dados de grande porte e produzir soluções de excelente qualidade, quando comparadas àquelas produzidas pelos métodos PAM e CLARA.

De forma avaliar a eficiência e a eficácia do novo método e do procedimento, foram realizados experimentos computacionais com várias bases de dados de porte variado, produzindo-se vários resultados interessantes que são apresentados e discutidos neste trabalho.

PALAVARAS CHAVE: Algoritmos Genéticos, k-Medoids, Agrupamento.