

MELHORANDO A REPRESENTATIVIDADE DE PROTÓTIPOS POR MEIO DA ANÁLISE DE ÍNDICE DE VALIDADE INTERNA.

Thomaz Araujo Ruckl (thomaz-araujo@outlook.com)
Alexandre Szabo (alexandreszabo@gmail.com)

O avanço da tecnologia tem favorecido o aumento da produção e do armazenamento de dados em áreas como biologia, medicina, transações comerciais, dentre outras. A área que trata desse assunto é denominada Descoberta de Conhecimento em Bases de Dados, sendo a Mineração de Dados (MD) uma de suas principais etapas. Dentre as tarefas de MD destaca-se a tarefa de Agrupamento de dados, cujo objetivo é descobrir grupos de dados com características similares entre si. O que se deseja é obter grupos compactos e bem definidos, tal que elementos do mesmo grupo sejam similares entre si e dissimilares em relação aos demais grupos. Assim, uma das maneiras de se avaliar a qualidade de uma solução de agrupamento é fazer uso de índices internos de validação, os quais usam distância intra/inter grupo. Essas medidas fazem uso da posição de representantes de grupos (protótipos) para obter as (dis)similaridades ao invés de comparar todos os elementos entre si. Tais índices fornecem a solução da partição e não a qualidade individual dos grupos, e um protótipo mal posicionado no grupo pode comprometer a solução de agrupamento como um todo. Assim, conhecendo a representatividade de protótipos é possível ajustá-los para melhorar a solução global e aumentar a confiabilidade na análise dos grupos encontrados. Dessa maneira, esse trabalho propôs um método para obter o índice de validade em cada grupo, identificar aquele com baixa representatividade de protótipo e ajustar sua posição. Experimentos foram desempenhados com o Algoritmo k-Médias, com o índice da Soma do Erro Quadrático, o qual avalia quão compactos são os grupos obtidos, com uma base de dados sintética e com dez bases de dados da literatura. Para medir a representatividade de um grupo usando o índice interno, seu valor é normalizado ((0,1]) em relação ao grupo de melhor índice (índice dos grupos individuais/índice do melhor grupo) e então o valor normalizado é comparado com um limiar para decidir sobre o ajuste de protótipos: se índiceNormalizado < limiar, então ajuste o protótipo do grupo. O ajuste do protótipo é realizado adicionando um novo protótipo ao redor deste e em seguida o k-Médias é executado novamente para ajustar a posição de protótipos nos respectivos grupos. Concluímos que aumentar o valor do limiar também aumenta a qualidade da solução global e não aumenta significativamente a quantidade de protótipos. O ganho de qualidade de solução com o método proposto em relação ao k-Médias foi entre 8,17% e 34,56%. Estatisticamente, o limiar de 0,8 é o mais interessante porque mantém a qualidade da solução com menor probabilidade de aumento do número de protótipos na partição.

Os autores agradecem à Universidade Federal da Grande Dourados pelo apoio financeiro.