

Metodologia para agrupamento de dados: versão VNS para linguagem R

Felipe Sousa Quintino

Universidade de Brasília - UnB
Instituto de Exatas/ Departamento de Estatística
Campus Darcy Ribeiro – Asa Norte – Brasília - DF
felipes.quintino2@gmail.com

Maria Amelia Biagio

Universidade de Brasília - UnB
Instituto de Exatas/ Departamento de Estatística
Campus Darcy Ribeiro – Asa Norte – Brasília - DF
mamelia@unb.br

RESUMO

Metodologias não-hierárquicas para agrupamento de dados têm sido bastante estudadas e utilizadas nas últimas décadas, muitas delas estimuladas pelos bons resultados demonstrados pela conhecida K-Means (MacQueen, 1967). Dentre as principais, destaca-se a metodologia H-Means (Howard, 1966), e mais recentemente sua forma não-degenerada H-Means+, as metodologias J-Means, J-Means+ e a metaheurística VNS (Hansen & Mladenovic, 2001), todas elas voltadas para a resolução do problema *MSSC* (*Minimum Sum of Squares Clustering*). No entanto, em ambiente computacional fortemente demandado por estatísticos, como a linguagem R de computação e estatística (R Core Team, 2014), estas metodologias, com exceção da primeira, ainda não estão disponíveis. O presente trabalho consiste no desenvolvimento computacional, através da linguagem R, das heurísticas de agrupamento não-hierárquico H-Means, J-Means, J-Means+ e a metaheurística VNS. Resultados computacionais são obtidos para o banco de dados Íris de Fisher e análise comparativa dos agrupamentos, segundo estas metodologias, é apresentada para distintos números de *clusters*. Os resultados apresentados são validados através dos valores ótimos já conhecidos para os testes realizados com o banco de dados Iris, dos quais pode-se observar a superioridade da metaheurística VNS, sobre as demais heurísticas, em situações em que o número de *clusters* é relativamente grande. Uma aplicação da metaheurística VNS ao agrupamento dos dias do ano de um banco de dados meteorológicos da região de Brasília, do ano de 2014, é apresentada, na qual é possível distinguir claramente dois grupos típicos de meses com características climáticas muito diferentes. Para sua validação, este último resultado é comparado com o resultado obtido pela heurística K-Means disponível no ambiente R. Nova fase deste trabalho deverá estar voltada à melhoria dos tempos computacionais observados para, *a posteriori*, buscar formas adequadas de disponibilização dos programas desenvolvidos aos usuários deste sistema.

Palavras chave: Agrupamento de Dados, Metaheurística VNS, Modelos Meteorológicos.

Main area: Statistics, Metaheuristics, Simulation.