Extended Abstract

Review and Research: Journal of Engineering And Technology

2021

Vol.10 Issue 7

Algorithms for a minimum sum of squares clustering

Pierre Hansen
GERAD and HEC Montreal, Canada.

Abstract:

Cluster analysis aims at solving the following very general problem: given a set of entities, find subsets (also called clusters) which are homogeneous and well separated. Homogeneity means that similar entities should belong to the same cluster. Separation means that dissimilar entities should belong to different clusters. This concept can be expressed mathematically in many ways. Hence, many heuristics and exact algorithms have been proposed. This problem is the main one in data mining with applications to a large variety of fields. Using the sum of squares of errors in clustering was already proposed by Steinhaus and co-workers in 1935. Since then, many heuristics and exact algorithms have been proposed for its resolution. Perhaps the most well-known most studied and most often applied in data mining is the minimum sum of squares clustering. Progress in the design of heuristics and, more recently, of exact algorithms has been substantial. All the algorithms are a branch and bound type (Ed-wards and Cavalli-Sforza 1965, Koontz, Narendra and Fukunaga 1975, Diehr 1985). A very important contribution is the K-means heuristic due to Llyod (first developed in 1957, published only 1982) and independently to Forgy (1965) and to MacQueen (1967). It works as follows: (1) choose an initial set of entities as cluster centers; (2) assign each entity to the closest center; (3) update the current centers by considering the centroids of the clusters; (4) return to step (2) as long as there is a modification in the centers. Despite a very substantial success (over 8525 citations of Lloyd's paper, according to Google Scholar), there are some difficulties in its application: (i) the number of clusters is not known A PERIORI, (ii) there is a strong dependency of the results on the initial choice of cluster centers, (iii) some clusters may disappear during the process. Many proposals have been made to alleviate these difficulties. Recent progress includes the use of metaheuristics: Variable Neighborhood Search Hansen and Mladenovi´c 1997, 2001), Tabu Search (Glover 1989) and GRASP (Fea and Resende 1989, 1995). Exact methods for the minimum sum of squares problems have also been developed. They include column generation (du Merle et al. 1999), cutting plans (Peng and Xia 2005), dynamic programming (Van Os and Meul-man 2004, Jensen 1969), DC programming (Tao 2007), and concave minimization (Bagirov 2008), Relation Linearization Technique (Sheraldi and Desai 2005). Merging branch and bound with an adaptation of a location problem, i.e.: Weber's problem with maximum distance led to substantial progress in exact resolution: the size of the largest instance solved exactly was raised from 220 to 2392 entities. The minimum sum-of-squared error clustering problem is shown to be a concave continuous optimization problem who's every local minimum solution must be integer. We characterize its local minima. A procedure of moving from a fractional solution to a better integer solution is given. Then we adapt Tuy's convexity cut method to find a global optimum of the minimum sum-of-squared error clustering problem. We prove that this method converges in finite steps to a global minimum. Promising numerical examples are reported.

This work is partly presented at 7th International Conference on Big Data Analytics & Data Mining | September 26-27, 2018 | Chicago, USA