World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:12, No:07, 2018

An Improved K-Means Algorithm for Gene Expression Data Clustering

Authors : Billel Kenidra, Mohamed Benmohammed

Abstract : Data mining technique used in the field of clustering is a subject of active research and assists in biological pattern recognition and extraction of new knowledge from raw data. Clustering means the act of partitioning an unlabeled dataset into groups of similar objects. Each group, called a cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Several clustering methods are based on partitional clustering. This category attempts to directly decompose the dataset into a set of disjoint clusters leading to an integer number of clusters that optimizes a given criterion function. The criterion function may emphasize a local or a global structure of the data, and its optimization is an iterative relocation procedure. The K-Means algorithm is one of the most widely used partitional clustering techniques. Since K-Means is extremely sensitive to the initial choice of centers and a poor choice of centers may lead to a local optimum that is quite inferior to the global optimum, we propose a strategy to initiate K-Means centers. The improved K-Means algorithm is compared with the original K-Means, and the results prove how the efficiency has been significantly improved.

Keywords : microarray data mining, biological pattern recognition, partitional clustering, k-means algorithm, centroid initialization
Conference Title : ICDM 2018 : International Conference on Data Mining
Conference Location : Dublin, Ireland
Conference Dates : July 23-24, 2018