An Enhanced K-Means Clustering Algorithm to Improve the Accuracy of Clustering using Centroid Identification Based on Compactness Factor

Dr.M.Sakthi
Associate Professor and Head
Department of Computer Science
NGM College
Pollachi, Coimbatore (Dist.), Tamil Nadu.

Abstract---The Researchers find it difficult to extract information from a large data set through a standard function. It is found insufficient of standard functions to extract needed information. It has been considered that the k-means algorithm in the situation where the data is too enormous to be stored in main memory and must be retrieved sequentially, such as from a disk, and where it must be used as slight memory as possible. The k-means clustering also converges very quickly when it is employed to obtain data from huge data collections. It is also on other hand, k-means has some disadvantages too, and it includes affluent computation by getting cluster centers which are randomly selected at initial. It influences the two factors, performance of the algorithm and number of clusters initialization. In this paper an improved k-means algorithm in terms of data clash strainer mechanism is given. The data clash strainer mechanism is implemented through a function Regional Centroid Component (RCC) mechanism which is added to the standard k-means algorithm. This density based recognition mechanism is built on the properties of clash data. The clustering result is effectively enhanced by ignoring the clash data prior to the process of data clustering. Hence, the improved algorithm offers a great accuracy when compared to other existing cluster algorithms.

Keywords: Cluster, Data clash strainer, K-means, RCC.

1. Introduction

A vast amount of data is dealt in various fields and those big data are handled using data mining techniques to retrieve information. "We are living in the information age" is a popular saying; however, it is like actually living in the data age. Terabytes or petabytes of data pour into our computer networks, the World Wide Web, and various data storage devices every day from business and which is needed to be extracted in a useful manner to infer knowledge from it [1]. The technique of data mining involves the cluster analysis which is one of the main focuses of the present-day researchers.

Clustering is a fundamental method for appreciative and interpreting data that seeks to partition input objects into groups, known as clusters, such that objects within a cluster are similar to each other, and objects in different clusters are not. A clustering invention called k-means is simple, intuitive, and widely used in practice. Given a set of points S in a Euclidean space and a parameter k, the objective of k-means is to partition S into k clusters in a way that minimizes the sum of the squared distance from each point to its cluster center [3]. This circumstance causes the