ABSTRACT

Clustering process is one of the vital aspects in data mining. In this thesis a few selected prime techniques are adopted to assess the quality of clusters. Fast and Efficient Dimensionality Reduction Algorithm (FEDRA) is described with its applications. A class of datasets of FEDRA as well as its variants are discussed. The inequality satisfying the lower and upper bounds of the distance between two points in the projection space is derived. Stress and purity values are computed through the algorithms. These values lead to identify the best variant and the highest quality cluster. On comparing different methods, it is concluded that principal component analysis is the best. The time requirements of k-means are evaluated and the time minimization is pointed out.

Graph clustering in data mining has been analyzed by using advanced distance measures. Structural and attribute cluster algorithm has been modified and objects are classified into several clusters. Density, Entropy and cluster efficiency for S, SA and W clusters corresponding to cluster size are computed and discussed.

The concepts of various metrics with their components are studied. The theoretical descriptions and mathematical expressions for metrics along with graphical representations are presented. The cluster quality is measured through these metrics. The remarkable suggestions are given to improve the clustering process and cluster quality.