CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Broad conclusion

This thesis entitled, “DESIGN AND DEVELOPMENT OF METRICS AND AN ALGORITHM TO ENHANCE CLUSTER QUALITY IN DATA MINING WITH STRUCTURAL AND ATTRIBUTE SIMILARITIES” consists of six chapters. The first chapter explains the concepts of landmark dimensionality reduction algorithm, FEDRA, various clusters and computational metrics along with their applications. The major purpose of this work is to improve the quality of clustering algorithms for high dimensional datasets by using FEDRA. Both structural and attribute similarities have been balanced by clusters and that integrated by Neighbourhood random walk distance measures. Metrics are playing effective roles in data mining as well as in cluster quality. The process leads to produce ineffective clusters whenever not using any such metrics.

The second chapter is devoted to the past contributions of various researchers on FEDRA, clustering similarities and metrics.

The third chapter describes the concept and application of FEDRA in dimensionality reduction of datasets. The distance between landmark points in both original space and projection space are considered and the points are used to study their projection. The role of landmark dimensionality reduction algorithm is discussed. Algorithms for FEDRA, Projection Heuristic and distributed algorithms are established. Using these algorithms, the values of stress and purity are estimated.
based on UCI Machine learning repository and Pascal large scale challenge respectively to variants of FEDRA and various methods for different number of dimensions. Based on the evaluation on large scale data sets the behaviours of FEDRA are validated. On comparing the four variants, the best variant as well as the high quality variant are identified. As in the case of comparing FEDRA against LMDS, RP and PCA through the values of stress and purity, the best method is judged. Another experiment is performed and evaluated for the time requirements of k-Means of the variants of FEDRA and the clustering quality is measured.

In the fourth chapter, graph clustering has been considered in data mining. Structural and attribute clustering algorithm has been modified. The technical terms for this study are described. Distance measures are used to obtain the similarity or dissimilarity between any pair of objects. Unified Neighbourhood random walk distances can be obtained through transition probability matrix and formed Neighbourhood random walk distance matrix. The expressions for influence function, density function and average weight are derived. The values for densities and entropies of different clusters are obtained and exhibited in the graphs. These estimated densities and entropies identify the best quality clusters. Based on modified SAC algorithm, the objects are classified into several clusters.

In the fifth chapter, the concepts of data mining and clustering are presented. The Roles of metrics in both data mining and cluster quality are explained. For focusing cluster quality, the most powerful techniques and issues of Metrics are described with suitable illustrations. The illustrations direct the researcher to identify the best quality clusters.
The Sixth chapter contributes the summary of the chapters from 1 to 5 and presented the suggestions for future work.

If any other technique other than FEDRA of landmark dimensionality reduction algorithm is applied, the researcher may achieve best results and get quality clusters. Likewise, in MSAC algorithm, the new densities and entropies may be obtained and well processed clusters may be classified through canonical correlation analysis and also by kernel PCA. Other than stated metrics, any researcher may use some mathematical or statistical metrics, the cluster size, coupling, complexity and total cost will be obtained.

6.2 Contributions

i. Clustering process is a vital one in data mining and its major achievement is to identify cluster quality. An important problem in clustering is how to decide what is the best set of clusters for a given dataset. In the literature, there are numerous techniques for studying cluster quality. It is impossible to use all techniques in this study. Therefore a few selected prime techniques are adopted to assess the quality of clusters.

ii. First of all, FEDRA and its variants are used. The distance measure between two points in the projection space is derived in an inequality form. The algorithms are established. The stress and purity values for the variants of FEDRA are computed by using the well framed algorithms. These values identify the best variant of FEDRA and high quality cluster. On comparing different methods, the analysis reveals that PCA is the best method and it reflects quality cluster. The time requirements of k-means are discussed for
various variants. The variant which requires minimum time period that identify cluster quality.

iii. Graph clustering has been studied by using advanced distance measures and related matrices. The objects are classified into several clusters by using SAC and Modified SAC algorithms and inferred that MSAC algorithm is better. In addition that density, entropy and clustering efficiency for S, SA and W clusters corresponding to different cluster sizes are compared and identified the best cluster.

iv. The cluster quality is also analysed based on some metrics and their components. The mathematical expressions and their relevant numerical values for cluster size, density, coupling and cost are presented and also exhibited graphically. These measures reflect the cluster quality. Apart from these methods and methodologies, suppose one may apply any other technique, he/she may get better cluster quality.

6.3 Future enhancement

The purpose of research on data mining is to study the cluster quality. The process of classification is an essential task for this type of analysis. Classification must be performed according to the size, shape, character and other properties of the objects of data set. Here, the following methods are suggested for future enhancement.

Bayes’ technique for classification, Canonical analysis, Factor analysis, Resource analysis in network, Line of balance are technically leading methods to
classify the data set into several clusters. These methods are based on either probability theory or correlation analysis and its principles which help to estimate the measures like Eigen values, Eigen vectors, density, entropy, various standard metrics, cost due to clustering and so on.

The above methods and measures lead to identify cluster quality. If a researcher may apply one or more of the above mentioned methods and walk through the measures, then the researcher may easily get the quality clusters.