CHAPTER 9

CONCLUSION AND FUTURE WORK

9.1 CONCLUSION OF THIS THESIS

Data mining is the process of retrieving a data from extracting useful knowledge, to achieve the effective utilization of data resources. In this research the number of contributions that are brought in an effective way to accomplish the predicted results.

Chapter 1 introduces an overview of the clustering process. This introduction has been demonstrated by the cluster analysis process. The architecture for clustering process is also described in this chapter. Based on this chapter implementation of five new methods proposed and investigated in subsequent chapters.

Chapter 2 defined the literature review and description is provided to study existing works on data clustering. The existing analytical models still needs to be improved, based on the identification of number of clusters. Most of the researchers suggest multiple values of $k$ before starts the clustering process. This leads to affects the final clustering results, which are described briefly.

Chapter 3, describes improving the clustering performance by combining KM and PSO algorithms. The PSO clustering algorithm performs a global search in the entire solution space. The KM algorithm is the fastest method because of its simple and less number of iterations. But the
dependency of the algorithm on the initialization of the centers is a major problem and it usually gets stuck in local optima though it tends to converge faster than the PSO algorithm. Using the merits of the algorithms, KMPSO algorithm is implemented and discussed. It does not depend on the initial clusters and a single particle represents a set of cluster centers. Each particle represents one possible solution for clustering and the position of each particle is constructed. The experimental results and ROC curve are illustrated the proposed algorithm gives better local optimal number of clusters. Hence the proposed approach is focused on retrieving more efficient number of clusters in a given datasets.

Chapter 4 illustrates the new hybrid algorithm KMBA, to select the best initial centroid of each cluster. The KMBA algorithm locates the distance between data object and centroid based on the echolocation behavior of bat position and velocity. It computes the cluster center by using BA algorithm and it forms the clusters by using the KM algorithm. The new algorithm improves the convergence speed of BA and helps KM independent on the initial centers that are briefly described. The experimental results and ROC curves show the robustness and efficiency of the proposed algorithm.

Chapter 5 investigate a new MHKMA algorithm to identify the global optimal solution. In each iteration, Hill-climbing algorithm will adjust a single element and determine whether the change improves the value of the cluster. It may also lead to a pitfall of Hill-climbing called plateau where the results of some plateau level values of \( k \) produce the same results. The proposed algorithm is marginally efficient in terms of reduced number of iterations with global optimal solution which are explained briefly. The experimental results and ROC curves illustrate the robustness and efficiency of the proposed algorithm.
In chapter 6 described a new direct visual validation method $V^2$VAT, to identify the number of clusters in a graphical representation. It is used only for the pairwise divergence matrix instead of dissimilarity matrix where it increases the efficiency of the algorithm. From the divergence matrix the VAT image are generated to identify the number of diagonal sub block. It suggests the number of clusters in the given datasets, which are discussed and implemented. The simulation results and ROC curve show the proposed algorithm gives better accuracy than other proposed methods.

Chapter 7 illustrates a new MO clustering algorithm BATMClustMOO which is proposed for encoding the cluster centers instead of data points. It can be detected the appropriate number of clusters and the partitioning the given dataset with many different types of cluster structures. Each cluster is splitted into number of hyper spherical subclusters and various clusters are assigned into points. From these local points sub-clusters are considered separately. The objective functions are used to check sub-clusters which are properly merged into variable number of clusters. Three objective functions are used to reflect the total compactness of the clusters based on the Euclidean distance. After obtaining number of clusters, three diverse cluster validity indices Sym-index, Con-index, I-index are used. Next the BA algorithm is used to fix the exact number of clusters for getting effective clusters that are implemented and discussed. The experimental result shows the proposed method that gives better accuracy of all the datasets.

Chapter 8 demonstrates the performances of the proposed algorithms which are compared and analyzed in terms of accuracy, precision, recall and F-measure quality checking parameters. It helps to analyze the existing KM, PSO, BAT and VAT algorithms with proposed algorithms. The final BATMClustMOO algorithm has the higher performance compared to other methods, because of which it is used for MO optimization. Therefore
the proposed method BATMClustMOO is effective in all performance measures among other techniques. It is smartly increased from 85% to 99% of accuracy in all the proposed methods.

9.2 FUTURE SCOPE

The future scope of the research lies in applying subspace clustering techniques to high dimensional datasets. Normally data are divided into the two parts such as sparse data and dense data. In the KM clustering technique, the sparse data are not supported. In other words, for the sparse data the clustering accuracy is low by using the proposed clustering methods. For dense data it has the high clustering accuracy. But for the sparse data it gives the lower accuracy. Hence use a subspace clustering to a high dimensional datasets to improve the clustering accuracy.

In order to form the effective cluster in complex data types to use the higher order statistical method like linear moment (summarize the shape of probability distribution), pareto index (specifying the pareto distribution) and nonparametric skew (measures a skewness of random variable’s distribution).

This study also needs to focus on how to reduce the time complexity without compromising cluster quality and optimality. More experiments will be conducted with complex natural datasets with different features.

On similar lines several new investigations may be carried out in future and some of the ideas worth exploring are:

1. Fuzzification of input data.
2. Parallel distributed implementation of clustering algorithms to reduce the time taken by the computational process.

3. Applying unsupervised neural networks to compare the clustering results.

The above mentioned methods may be considered as a very fruitful research area. It is firm that the new algorithms proposed and formulated in this thesis have the potential to become trend setters in clustering problem and will attract many of the researchers to apply them in their respective field of interest.