CHAPTER 6
CONCLUSIONS AND FUTURE STUDY

The present study proposes new clustering methods to find optimal clusters from the given data set. The proposed clustering methods are extensions to k-means. These are aimed at finding optimal cluster structures automatically by overcoming the draw backs in the most of the existing clustering algorithms. Various cluster validity measures are implemented to evaluate clustering results of proposed algorithms (SPSS, AGMFI, AMOC and AMSOS). Evaluation tool like FatiGO is used to determine the gene similarity of a micro array clusters. The present investigation on various data sets proved that the proposed algorithms in this thesis are robust, automatic and efficient in finding optimal cluster structures over the existing ones.

Findings of Present study

1. The SPSS algorithm is an extension to k-means++ to find optimal centroids for k-means type clustering algorithms. The SPSS determines initial centroids in single pass. The clusters obtained in this process are unique where as the existing algorithms result different structures in different runs.

2. The average improvement over k-means is 10%, over k-means++ is 3% and over fuzzy-k is 2.8%. On an average there is nearly 5% improvement with robust solution in a single pass.

3. The quality of SPSS is 82.58% in terms of Rand measure.

4. The proposed AGMFI approach clearly specifies that it can find the optimal clusters and performs well on all datasets compared to existing algorithms but is sensitive to the input parameter k (number of clusters). It works well when k is close to the number of clusters in population.

5. The proposed AMOC algorithm is more efficient than AGMFI algorithm in terms of optimal clusters for any large k. The experimental results using seven validity measures and error rate on eleven data sets show that the AMOC algorithm produces nearly optimal clusters compared to existing algorithms.

6. The results of AMOC show that average error rates is equally good when compared to those of k-means, k-means++, fuzzy-k and SPSS.
7. The results of AMOC show that they are far better when compared to ACDE (a recent automatic clustering method) in most of the cases.

8. The best error rate over 40 runs of AMOC is very much comparable to the existing algorithms.

9. The maximum error rate over 40 runs of AMOC appears to be the least when compared to those of existing algorithms.

10. The quality of AMOC in terms of Rand index is 70%.

11. The proposed AMSOS algorithm produces unique clustering structure automatically. The misclassification rate of single clustering solution of the AMSOS is either as same as the minimum error rate found in 40 independent runs or nearly equal with the means of different solutions of the existing algorithms in most of the cases.

12. The proposed AMSOS nearly equally good with existing algorithms k-means, k-means++, fuzzy-k, SPSS and far better with ACDE, an automatic clustering method.

13. The proposed AMSOS shows an improved performance of 30% over ACDE in terms of error rate.

14. The average performance improvement of AMSOS is 8% over AMOC in terms of error rate.

15. The AMSOS produces 82.21% qualitative clusters in terms of Rand validity measure.

16. The quality of AMSOS on all data sets is 72.15% in terms of silhouette measure.

17. It is evident from the heat maps of the microarray cluster profiles that the expression profiles of the genes of a cluster are similar to each other and they produce similar color patterns.

18. As determined by FatiGO (a web tool to evaluate microarray clusters using GO), the proposed algorithms increased the enrichment of genes of similar function within the cluster.

19. The smaller P values (nearer to zero) in the FatiGO results indicate that the majority of genes in a cluster belong to one category and a particular GO term is associated with the group of genes.

20. Recently Sudhakar Jonnalagadda and Rajagopalan Srinivasan (2009) developed a method that determined 5 clusters from yeast data set where as the existing finds as 4. The AMOC proposed in this study also finds 5 clusters from yeast2 dataset.
Scope for further work

The novel cluster analysis models developed for the selection of optimal clusters is useful for future research to generate new clustering techniques that

- combine feature subset selection in selection of clusters to reduce dimensionality
- new validity measures that measure cluster accuracy more accurately
- incorporate biological knowledge while clustering gene expression data to produce more appropriate co-expressed gene groups and coherent patterns
- aimed at multi objective clustering for micro array data and
- find soft clustering solutions.

The new cluster analysis models can be used to identify gene regulatory networks, such as diabetes micro array data sets to determine biologically relevant groups of genes and to reveal the co-expression of genes which were uncharacterized previously. These models help to find the genes that cause a specific disease by comparing with normal cells to facilitate drug design and therapy planning.