Clustering Subspace For High Dimensional Categorical Data Using Neuro-Fuzzy Classification

Ms. K.Karunambiga#1,

#1M.phil Research Scholar

School of computer studies

RVS college of arts and science

Sulur,Coimbatore-402,TN,India

Mrs. M.Suganya *2

**2Assistant Professor*

School of computer studies

RVS college of arts and science

Sulur,Coimbatore-402,TN,India

Abstract

Clustering has been used extensively as a vital tool of data mining.Data gathering has been deliberated widely, but mostly all identified usual clustering algorithms lean towards to break down in high dimensional spaces because of the essential sparsity of the data points. Present subspace clustering methods for handling high-dimensional data focus on numerical dimensions.The minimum spanning tree based clustering algorithms, because they do not adopt that data points are clustered around centers or split by a regular geometric curve and have been widely used in training. The present techniques allow these algorithms to extend much more easily with both the number of instances in the dataset and the number of attributes. But the performance minimize soon with the size of the subspaces in which the groups are found. The important parameter needed by these algorithms is the density threshold and it is not easy to set, particularly across all dimensions of the dataset. The aim of this paper is proposed method investigate the performance of different Neuro-Fuzzy classification methods for the distinction of benign and malign tissue in genes.

Keywords : **ranking query, web database, immersed web**

I. INTRODUCTION

Data mining [5] a vast area in computer science, is the computational flow of identifying patterns in large data sets which holds methods at the centre point of artificial intelligence, statistics and database systems. The vital aim of the data mining is to explore information from a data set and transform it into an meaningful structure for research in future.

Mostly conventional clustering algorithms do not scale better to cluster high dimensional data sets in terms of success and proficiency, due to the natural sparsity of high dimensional data. In high dimensional data sets, there come across many problems. The distance between any two data points becomes exacts the same, so it is difficult to differentiate same data points from unlike data points. Clusters[1] are implicit in the subspaces of the high dimensional data space, and different clusters may available in various subspaces of different dimensions leads to another problem. Because of these problems, almost all conventional clustering algorithms fail to work well for high dimensional data sets.

The resemblance between objects is frequently processed using distance measures over the different dimensions in the dataset [4],[5]. Advanced Technology made data collection simple, quick, and results in larger, complex datasets with more objects and dimensions. As the datasets become larger and differs, alterations to existing algorithms are needed to hold cluster values and speed. A traditional clustering algorithm predicts all of the dimensions of an input dataset help in order to learn as much as potential about each object explained. In high dimensional data, most of the dimensions are frequently irrelevant.

The results must be carefully analyzed to ensure they are meaningful. The algorithms must be efficient in order to scale with the increasing size of datasets.

II. BACKGROUND STUDY

Feature subset selection can be analyzed as the process of identifying and removing as many unrelated and replicated features as possible. Due to following reasons 1) unrelated features do not donate to the predictive accuracy and 2) Replicated features do not redound to getting a well predictor for that they offer almost information which is already available in other feature(s). Among the feature subset selection algorithms, some can effectively reduce unrelated features but fail to get success in handling replicated features. But some of others can reduce the unrelated and also the redundant features. Our proposed Neuro fuzzy classification algorithm falls into the second group which reduce the unrelated and also the redundant features. Traditionally, feature subset selection research has focused on searching for relevant features.

Feature selection includes classifying a subset of the most functional features that generates well-suited results as the original entire set of features. A feature selection algorithm may be compared from both the competence and success points of view.

III. METHODOLOGY

Techniques for clustering high dimensional data have included both feature transformation and feature selection techniques. Feature transformation techniques attempt to describe a dataset in minimum dimensions by creating mixture of the exact attributes. These methods are very helpful in uncovering latent structure in datasets.

DNA microarrays are a sensational new technology with the ability to increase our requirement of complex cellular mechanisms. Microarray datasets enables information on the expression levels of thousands of genes under hundreds of conditions. For instance, it can make clear a lymphoma dataset as 100 cancer profiles with 4000 features. This feature is the expression level of a particular gene. This study allows us to discover many cancer subtypes depends upon relationships between gene expression profiles. Realizing the differences between cancer subtypes on a genetic level is vital to understanding which types of treatments are most preferable to be effective.

As a replacement it also can view the data as 4000 gene profiles with 100 features equivalent to specific cancer models. Patterns in the data expose information about genes whose products function altogether in pathways that do complicated functions in the organism. The analysis of these pathways and their relationships to one another can then be used to form a finished model of the cell and its functions, link the gap between genetic maps and living organisms.

In the present system, microarray data must be preprocessed to minimize the number of attributes before meaningful clusters can be discovered. Additionally, individual gene products have many different roles under various occurrences.

For instance, a particular cancer may be segmented along more than one set of features. There may be subtypes built on the motility of the cell as well as the cells tendency to split. Separating the models depend on motility would require experimenting one set of genes, while subtypes based on cell division would be explored when looking at a various set of genes. In order to detect such complicated relationships in enormous microarray datasets, more powerful and adaptable methods of studies are required. Subspace clustering is a hopeful method that extends the power of traditional feature selection by searching for identical subspaces for each cluster.

A. Neuro fuzzy classification

Subspace clustering is an expansion of characteristic selection that tries to find clusters in different subspaces of the same dataset.

The Neuro-Fuzzy classifiers[2] were tested on the well-known set of lymphoma data introduced by Fisher which consists of a three class problem based on four parameters of the genes, i.e. the petal length and width and the sepal length and width. One type of the genes can be separated linearly from the two other types whereas the other two types of the genes cannot be separated linearly from each other.

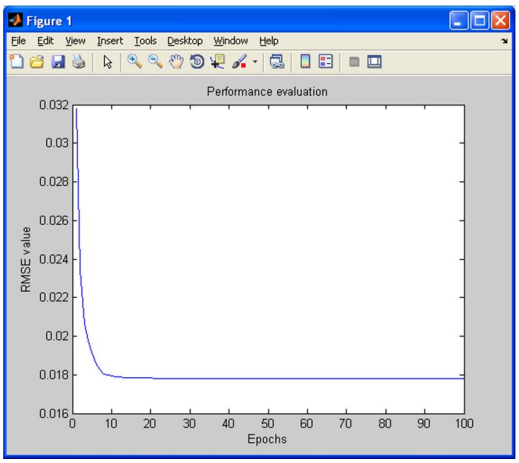

Fig 1:Performance of rmes value

Neuro-Fuzzy systems use a learning procedure to determine an appropriate set of Fuzzy membership function.

This set of membership functions can be expressed in linguistical terms and hence provides an understanding about the properties of the classification problem. Fuzzy [6] systems allow to incorporate a priori knowledge into the classification process which enables to include some of the experience of the physician into the classifier.

IV. EXPERIMENT RESULTS

In general the Bayes classifier and the KNN[3] classifier could not handle the massive data as good as the Neuro-Fuzzy classification systems. This helps our earlier analysis using artificial neural networks. This effect was not experimented with the lymphoma data which can be contributed to the different statistical properties of the two data sets. With the lymphoma data all system had problems to find the similar four outliers which limited the achievable recognition rate to 97.33 %. This well-maintained a smooth relationship between sensitivity, specificity and produced maximum recognition rates.

This study was done on segments with confirmed histology in small regions of interest within the area. Now we continuing to gather lymphoma data in order to result a data base of benign and malign tissue in genes.

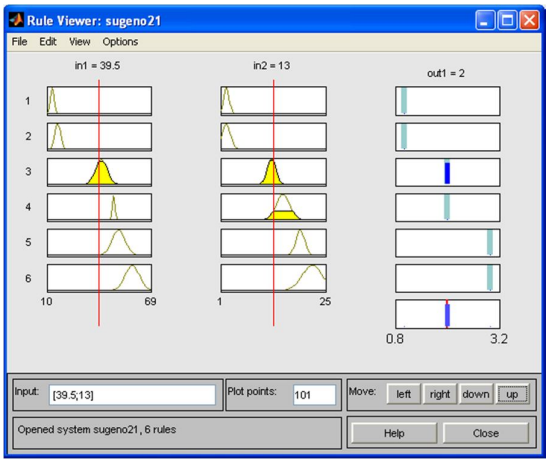

Fig 2: Neuro fuzzy classification(a)

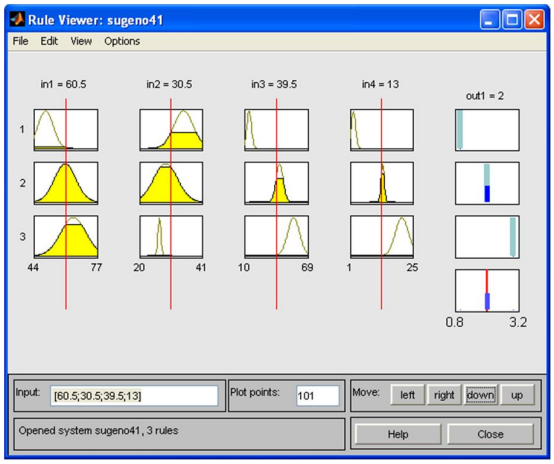

Fig 3: Neuro fuzzy classification(b)

V. CONCLUSION

High dimensional data is popularly common in many research areas. When the number of dimensions increased, many clustering methods start to suffer from the trial of dimensionality, humiliating the excellence of the results. In high dimensions, data becomes very thin and distance measures become meaningless. This problem has been experimented extensively and there are different solutions, each suitable for various methods of high dimensional data and data mining procedures. Subspace clustering tries to mis feature estimation and clustering in order to detect clusters in various subspaces

The Neuro-Fuzzy classification system, which is based on a built clustering algorithm reached recognition rates above in comparison to the Bayes classifier) and the KNN classifier. Our experiment results recommend that Neuro-Fuzzy classification algorithms have the capability a lot to progress common classification systems that can be used in ultrasonic tissue characterization.

With this proposed study we examined Neuro-Fuzzy classification algorithms which are based on different approaches to arrange and categorize biological data sets by the assembling of a Fuzzy interference system.

VI. SCOPE FOR FUTURE ENHANCEMENT

This proposed Neuro-Fuzzy classification algorithms based on various methods to arrange and classify biological data sets by the development of a interference system. These results prove that Neuro-Fuzzy algorithms have the capability to improve classification methods for the use in ultrasonic tissue characterization

In future this study can be stretched in many methods in order to provide better evaluations such as continuous data, missing values, and the use of combined evaluation measures.

REFERENCES

[1] A. K. Jain and R. C. Dubes. *Algorithms for clustering data.* Prentice-Hall, Inc., 1988.

[2] D. Nauck, R. Kruse, *"NEFCLASS - A Neuro-Fuzzy approach for the classification of data,"* presented at the Symposium on Applied Computing, Nashville, USA, 1995

[3] Z. Song and N. Roussopoulos. *K-nearest neighbor search for moving query point. In SSTD,* pages 79–96, 2001

[4] C. C. Aggarwal. *Towards meaningful high-dimensional nearest neighbor search by human-computer interaction. In Data Engineering, 2002*. Proceedings. 18th International Conference on, pages 593{604, 2002.

[5]. M. K. Jiawei Han. *Data Mining : Concepts and Techniques,* chapter 8, pages 335 393. Morgan Kaufmann Publishers, 2001.

[6] D. Nauck, *"Fuzzy Neuro Systems: An Overview," in Fuzzy Systems in Computer Science. Vieweg, Wiesbaden, 1994*

[7] G.H. John, R. Kohavi, and K. Pfleger, *"Irrelevant Features and the Subset Selection Problem,"* Proc. 11th Int'l Conf. Machine Learning, pp. 121-129, 1994.

[8] K. Kira and L.A. Rendell, *"The Feature Selection Problem:Traditional Methods and a New Algorithm,"* Proc. 10th Nat'l Conf. Artificial Intelligence, pp. 129-134, 1992.

[9] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, nos. 1/2, pp. 273-324, 1997.

ABOUT AUTHORS

[1] Mrs. Suganya M M.Sc(Cs).,M.Phil., is working as an Assistant Professor at RVS College of Arts and Science, Sulur, Coimbatore , India International and National level journals. Her area of interest is Netwoking.

[2] Ms. Karunambiga K M.Sc(IT)., is working as a Guest Lecturer in the Department of Computer Science in LRG Govt Arts College for Women, Tirupur,TN, India. At present pursuing M. Phil in RVS College of Arts and Science College, Sulur, Coimbatore, TN, India. Her area of interest is Data Mining.