International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Automatic Clustering Subspace for High Dimensional Categorical Data Using Neuro-Fuzzy Classification

R. Mahalingam[1], S. Omprakash[2]

[1]Department of Computer Science, Kovai Kalaimagal College of Arts & Science, Coimbatore-109, India

[2]Assistant Professor, Department of Computer Science, Kovai Kalaimagal College of Arts & Science, Coimbatore – 109, 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 sparsely 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: Classification, Neuro-Fuzzy, SVM, KNN, Dataset

1. Introduction

1.1 Overview of Data Mining

Advanced technologies have enabled us to collect large amounts of data on a continuous or periodic basis in many fields. The data present the potential for us to discover useful information and knowledge that we could not see before. This limitation demands automatic tools for data mining to mine useful information and knowledge from large amounts of data.

Clustering high dimensional data is an emerging research field. Subspace clustering or projected clustering group similar objects in subspaces, i.e. projections, of the full space. Clustering has been used extensively as a vital tool of data mining.

There are two approaches in Feature selection known as **Forward** selection and **Backward** selection. Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection methods can be decomposed into three broad classes

The Neuro-Fuzzy classification system, which is based on a built clustering algorithm reached recognition rates than other classifiers [1]. 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.

The objective of feature selection for unsupervised learning is to find the smallest feature subset that best uncovers clusters form data according to the preferred criterion. Feature selection for unsupervised learning can be subdivided into filter methods and rapper methods.

The aim of our work was to investigate the performance of different Neuro-Fuzzy classification methods for the distinction of benign and malign tissue in ultrasound prostate diagnosis. This study was done on segments with confirmed histology in small regions of interest within the area and it continuing to gather lymphoma data in order to result a data base of benign and malign tissue in genes.

1.2 Motivation and Goals

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends.

Such analysis can help provide us with a better understanding of the data at large. Most previous methods or algorithms are memory resident, typically assuming a small data size.

1.3 Objective

The goal of classification is to build a model that captures the intrinsic associations between the class type and the attributes so that an (unknown) class type can be accurately predicted from the attribute values. The clusters are unions of connected high density units within a subspace. Neural networks handle a classification efficiently along with the Fuzzy method that implicitly transforms the input space into another higher dimensional feature space.

1.4 Scope of the Research

The scope of the projects is to provide good classification results, but their behavior can also be explained and interpreted in human understandable terms, which provides the researchers with a better understanding of the data using fuzzy-Neuro system.

- To reduce the dimensionality of the feature space, to limit storage requirements and increase algorithm speed;
- To remove the redundant, irrelevant or noisy data.
- The immediate effects for data analysis tasks are speeding up the running time of the learning algorithms.
- To improve the data quality.
- To increase the accuracy of the resulting model.
- Performance improvement, to gain in predictive accuracy;

2. Review of Literature

Many important problems involve clustering large datasets. Although naive implementations of clustering are computationally expensive, there are established efficient techniques for clustering when the dataset has either;

- A limited number of clusters,
- A low feature dimensionality, or
- A small number of data points.

2.1 Unsupervised Clustering

Unsupervised clustering techniques have been applied to many important problems. By clustering patient records, health care trends are discovered. By clustering address lists, duplicate entries are eliminated. By clustering documents, hierarchical organizations of information are derived. There are three different ways in which the data set can be large:

- There can be a large number of elements in the data set,
- Each element can have many features, and
- There can be many clusters to discover.

Recent advances in clustering algorithms have addressed these efficiency issues, but only partially. For example, KD-trees provide for efficient EM-style clustering of many elements, but require that the dimensionality of each element be small. Another algorithm [2] efficiently performs K-means clustering by finding good initial starting points, but is not efficient when the number of clusters is large.

2.2 Evaluating Clustering in Subspace Projections of High Dimensional Data

Conclusive evaluation and comparison is challenged by three major issues.

- First, there is no ground truth that describes the true" clusters in real world data.
- Second, a large variety of evaluation measures have been used that reflect different aspects of the clustering result.
- Finally, untypical publications authors have limited their analysis to their favored paradigm only, while paying other paradigms little or no attention.

2.3 Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching

We present a new technique for clustering these large, high-dimensional datasets. The key idea involves using a cheap, approximate distance measure to efficiently divide the data into overlapping subsets we call *canopies*.

- Which descriptions refer to the same object, and
- What the best description of that object is.

We present experimental results for the domain of bibliographic citation matching. Another important instance of this classis the *merge-purge problem*. Companies often purchase and merge multiple mailing lists.

- Irrelevant features do not contribute to the predictive accuracy and
- Redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s).

2.4 Feature Subset Selection Algorithm

The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm, it involves;

- The construction of the minimum spanning tree from a weighted complete graph;
- The partitioning of the MST into a forest with each tree representing a cluster; and
- The selection of representative features from the clusters. In order to more precisely introduce the algorithm, and because our proposed feature subset selection framework involves irrelevant feature removal and redundant feature elimination.

2.5 Fast Clustering-Based Feature Subset Selection Algorithm For High-Dimensional Data

A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. Based on these criteria, a fast clustering-based feature selection algorithm (FAST) is proposed and experimentally evaluated in this paper. The FAST algorithm works in two steps.

- In the first step, features are divided into clusters by using graph-theoretic clustering methods.
- In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features.

To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study. The results, on 35 publicly available real-world high-dimensional image, micro array, and text data, demonstrate that the FAST not only produces smaller subsets of features but also improves the performances of the four types of classifiers. The algorithm involves;

- Removing irrelevant features,
- Constructing a minimum spanning tree from relative ones, and
- Partitioning the MST and selecting representative features.
- In the proposed algorithm, a cluster consists of features.

Filter methods in unsupervised learning are defined as using some intrinsic property of the data to select feature without utilizing the clustering algorithm. Entropy measure has been used as filter method for feature selection for clustering [3]. According to Kriegel, Kröger&Zimek (2009), four problems need to be overcome for clustering in high-dimensional data:

- Multiple dimensions are hard to think in, impossible to visualize, and, due to the exponential growth of the number of possible values with each dimension, complete enumeration of all subspaces becomes intractable with increasing dimensionality. This problem is known as the curse of dimensionality.
- The concept of distance becomes less precise as the number of dimensions grows, since the distance between any two points in a given dataset converges. The discrimination of the nearest and farthest point in particular becomes meaningless:
  $$\text{Lim}_{d\text{-}\infty}((\text{dist}_{max} - \text{dist}_{min})_{/\,(}\text{dist}_{min}) \rightarrow 0$$
- A cluster is intended to group objects that are related, based on observations of their attribute's values. However, given a large number of attributes some of the attributes will usually not be meaningful for a given cluster. But for different diseases, different blood values might form a cluster, and other values might be uncorrelated. This is known as the *local feature relevance* problem: different clusters might be found in different subspaces, so a global filtering of attributes is not sufficient.
- Given a large number of attributes, it is likely that some attributes are correlated. Hence, clusters might exist in arbitrarily oriented affine subspaces.

3. Methodology

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.

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

- Unrelated features do not donate to the predictive accuracy and
- Replicated features do not redound to getting a well predictor for that they offer almost information which is already available in other feature(s).

3.1 Proposed Algorithm

Neural networks and fuzzy systems can be combined to join its advantages and to cure its individual illness. Neural networks introduce its computational characteristics of learning in the fuzzy systems and receive from them the interpretation and clarity of systems representation. Thus, the disadvantages of the fuzzy systems are compensated by the capacities of the neural networks.

3.2 Neuro Fuzzy Classification

A Neuro-Fuzzy method is proposed in this paper for analyzing the gene expression data from micro array experiments. The proposed approach was tested on three benchmark cancer gene expression data sets. Experimental results show that our Neuro-Fuzzy method can be used as an efficient computational tool for micro array data analysis.

- An individual is a real object of reference.
- A universe of discourse is the set of all possible individuals considered.
- A variable V:⟶R is a function which maps into a predefined range R without any given function arguments: a zero-place function.
- A propositional function is "an expression containing one or more undetermined constituents, such that, when values are assigned to these constituents, the expression becomes a proposition" (Russell, 1919, S. 155).

The fuzzy robust principal component analysis algorithm used here and from where the nonlinear case is derived were introduced. A more thorough description can be found in and

$$E(U,w) = \sum_{i=1}^{n} u, e(x) + n \sum_{i=1}^{n} 1 - u, \qquad (1)$$

$Themeasuree(x_i)couldbee.goneofthefollowingdunctions:$

$$e_1(x_i) = \|x_i - w^j x_i w\|^2 \qquad (3)$$

The emergence of various new application domains, such as bioinformatics and e-commerce, underscores the need for analyzing high dimensional data. Researchers and practitioners are very eager in analyzing these data sets.

3.3 Classification of Cancer

DNA micro arrays are an exciting new technology with the potential to increase our understanding of complex cellular mechanisms. Micro array datasets [4][5] provide information on the expression levels of thousands of genes under hundreds of conditions. For example, we can interpret a lymphoma dataset as 100 cancer profiles with 4000 features where each feature is the expression level of a specific gene.

Lymphoma Dataset: The lymphoma cDNA array dataset comprises gene expression patterns of genes involved in different classes of lymphoma and normal cell lines from Alizadeh et al [6]. Cancer diagnosis has traditionally been carried out based on clinical and molecular evidence such as cell and tissue type, and heredity. Collect the gene expression in three classes of lymphoid malignancies: diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL), and chronic lymphocytic leukemia (CLL), and genes relevant to lymphocyte and/or cancer biology.

Lymphoma Data: Two types of classification are studied with the lymphoma data:

- Binary classification between cancerous and non-cancerous samples.
- Tissue type classification based on global gene expression.

Binary Classification of Cancerous / Non-cancerous Tissues

Table 3.1 shows the typical performance of Neuro -Fuzzy with the distance for cancer detection. These results on error average and variance are compared with results from Cai*et al.* using a K-means SVM and a KNN decision tree.

Table 3.1: Classification Error for Cancer Detection

LVQ Classification	True positive	True Negative	False Positive	False negative
Cancerous Tissues	72	22	2	0
Non-cancerous Tissues	22	72	0	2

In Table 3.2, the results from the Neuro –Fuzzy algorithm are based on 10 cross-validation experiments, and compare closely to the globally optimal SVM method. The Proposed algorithm Neuro-Fuzzy gives a much better and more stable performance compared to the C4.5 decision tree algorithm.

Table 3.2: Classification Error Comparison for Cancerous/Non-cancerous Cells

	Neuro-Fuzzy	*SVM*	*KNN*
Average error rate	2.08%	2.94%	8.55%
Error rate standard deviation	0.01%	0.03%	2.83%

3.4 Data Collection and Training

Once the optimal set of inputs and outputs has been identified, the next step is to collect sample input and output data to train the neural net. For this application, the training data was collected by measurements. Neuro Fuzzy requires several training parameters to be set before training.

These parameters are error (convergence) criterion, learning rate, and number of membership functions. These parameters can have significant effect on the final system solution. The network converges when the neural net learns to produce outputs within the specified error range for all training patterns. The number of membership functions chosen affects the level of accuracy achievable by the neural net.

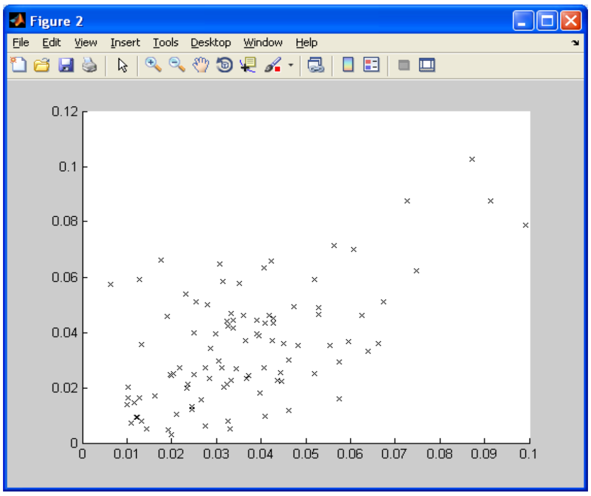


Figure 3.1: Cancer data classification

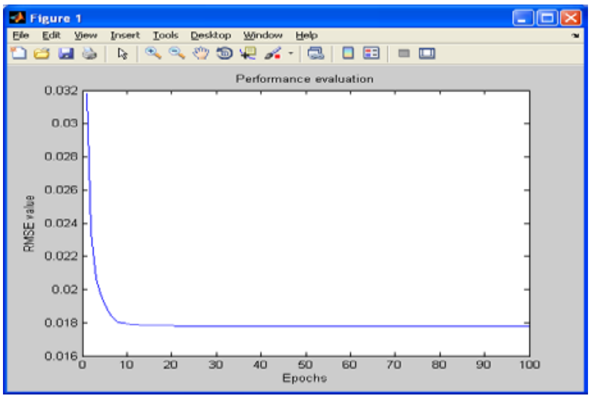


Figure 3.2: Performance of RMES value

4. Implementation of the Proposed System

The experimental results of our proposed Segmentation technique using lymphoma data set. Our proposed approach is implemented in MATLAB (MATLAB version 2013 a). In software the algorithm was implemented using the Dot Net(c#) language running on windows xp operating system[7]. MATLAB is a technical computing environment that is published by The MathWorks.

4.1 Neural Network Training Methodology

The figure below in 4.1 shows the methodology to follow when training a neural network. First you must collect or generate the data to be used for training and testing the neural network. Once this data is collected, it must be divided into a training set and a test set. If the error goal is met, training is complete. If the error goal is not met, there could be two causes:

- Poor generalization due to an incomplete training set.
- Over fitting due to an incomplete training set or too many degrees of freedom in the network architecture.

4.2 Fuzzy Neural Networks

If there is experiential input/output data from the relationship to be modeled, the fuzzy neural network's weights and biases can be trained to better model the relationship. This training can be performed with a gradient descent algorithm similar to the one used for the standard neural network.

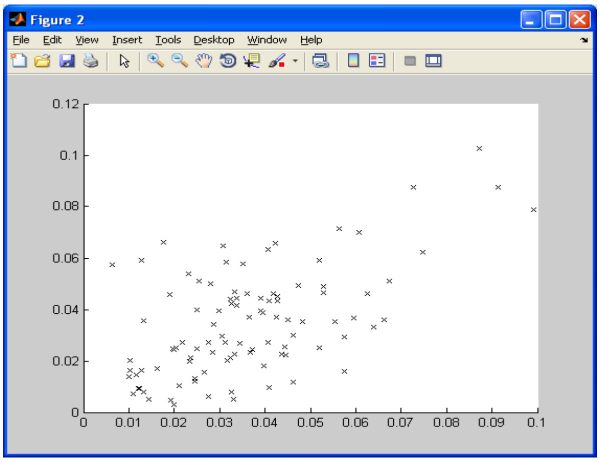


Figure 4.1: Lymphoma dataset

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

5. Experimental Results

In general the Bayes classifier and the KNN[2] classifier could not handle the massive data as good as the Neuro-Fuzzy classification systems. 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 95.85 %. This well-maintained a smooth relationship between sensitivity, specificity and produced maximum recognition rates.

- Number of selected genes,
- Predictive accuracy on selected genes,
- Extracted knowledge from the trained models.

Table 5.1: Comparison of the classification performance of different Classifiers on lymphoma cancer data set

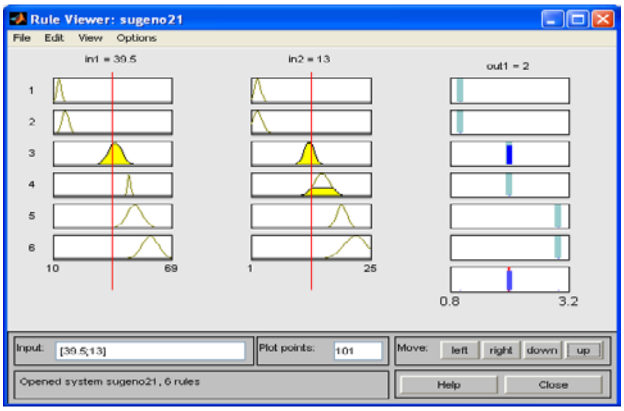
Methods	*Gene Selection*	*Lymphoma*
Neuro-fuzzy	Information gain	95.85
SVM	Information gain	N/A
	Information gain	72.64

6. Conclusion

Clustering large data sets is a ubiquitous task. Astronomers work to classify stars into similar sets based on their images. Search engines on the web seek to group together similar documents based on the words they contain or based on their citations. Marketers seek clusters of similar shoppers based on their purchase history and demographics.

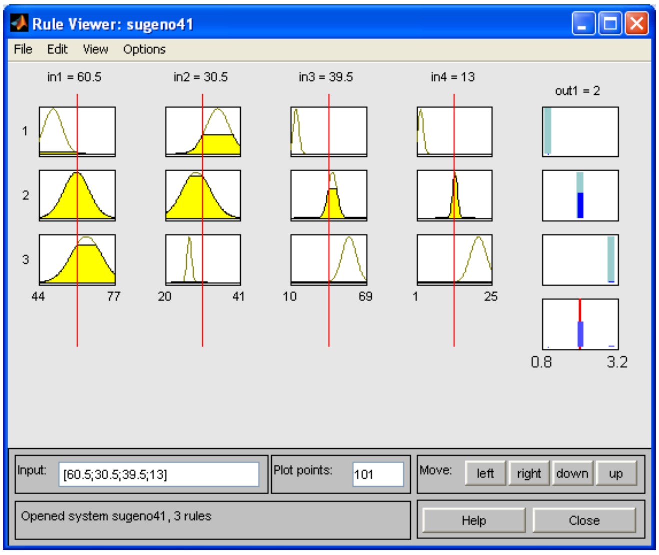
The task of a Neuro-Fuzzy classification is to provide with a computational solution to the feature selection problem motivated by a certain definition of *relevance*. This algorithm should be reliable and efficient. The many Neuro-Fuzzy classification is proposed in the literature are based on quite different principles (as the evaluation measure used, the precise way to explore the search space, etc) and loosely follow different definitions of relevance.

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.



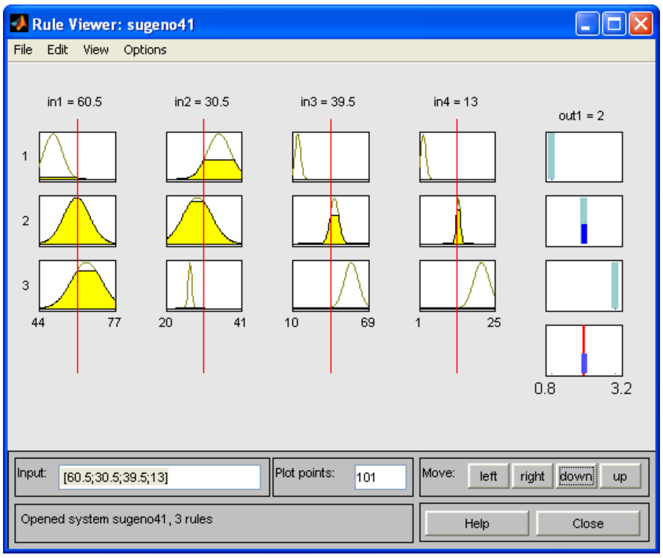
Input: [39.5;13]
Plot points: 101
Opened system sugeno21, 6 rules

Figure 6.1(a): Fuzzy Neuro classification



Input: [60.5;30.5;39.5;13]
Plot points: 101
Opened system sugeno41, 3 rules

Figure 6.2(b): Fuzzy Neuro classification



Input: [60.5;30.5;39.5;13]

Plot points: 101

Opened system sugeno41, 3 rules

Figure 6.3(c): Fuzzy Neuro classification

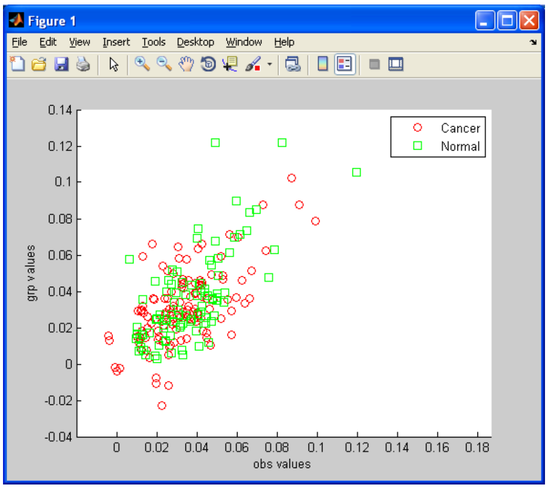


Figure 6.2: Overall data clustering

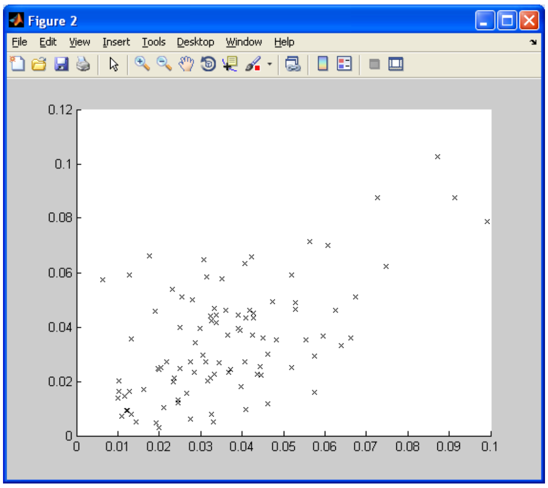


Figure 6.3: Cancer data classification

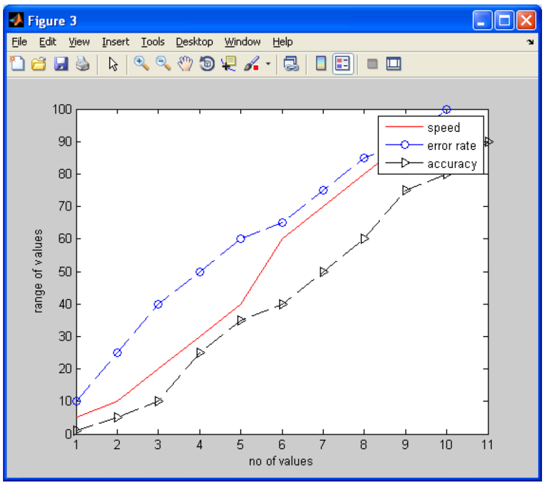


Figure 6.4: Metric range of the graph

7. Scope for Future Work

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] F. N. Afrati, A. Gionis, and H. Mannila.Approximating a collection of frequent sets. In Proc. 2004 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'04), pages 12–19, Seattle,WA, Aug. 2004.

[2] D.L. Boley.Principal direction divisive partitioning.*Data Mining and KnowledgeDiscovery*, 2(4):325–344, 1998.

[3] I.S. Dhillon, S. Malella, and R. Kumar.Enhanced word clustering forhierarchical text classification.In*KDD-2002*, 2002.

[4] S. Cho and H. Won, "Machine learning in dna microarray analysis forcancer classification," in *APBC*, vol. 34, 2003, pp. 189–198

[5] U. Fayyad and K. Irani, "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," Proc. 13thInt'l Joint Conf. Artificial Intelligence, pp. 1022-1027, 1993.

[6] D.H. Fisher, L. Xu, and N. Zard, "Ordering Effects in Clustering,"Proc. Ninth Int'l Workshop Machine Learning, pp. 162-168, 1992.

[7] Alizadeh, A.A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503-511 (2000).

Author Profile

Mr. R. Mahalingam, M.Sc., M.Phil., Department of Computer Science, Kovai Kalaimagal College of Arts & Science, Coimbatore-109, India.

Mr. S. Omprakash., M.Sc., M.Phil., Assistant Professor, Department of Computer Science, Kovai Kalaimagal College of Arts & Science, Coimbatore-109, India.