Improving classification Accuracy of Neural Network through Clustering Algorithms

B.Madasamy[#1], Dr.J.Jebamalar Tamilselvi[*2]

[#]Assistant Professor & Research Scholar, Dept of MCA,
Agni College of Technology, Anna University, Chennai, India.
[*]Director & Professor, Dept of MCA,
Jaya Engineering College, Anna University, Chennai, India

Abstract— A **common problem in bio medical data using neural networks for classification purposes are complex nature of the data, high dimensionality, convoluted and overlapping classes with their remarkable ability to derive meaning from complicated data, can be extract patterns and trends are too complex. Bio medical classification is a complex process to make decisions. Classification performance of the neural network suffers dramatically. This paper proposes neural network and data mining techniques are combined to automate biomedical classification processes to support decision. To improve the classification ability and behavior of neural network is used by pre-processing and pre-clustered data with the help of Rule based induction, Multi-layer perceptron model, nearest neighbor, Radial basics function and back propagation learning algorithm is employed to classify such complex tasks. The proposed clustering algorithm applied to the bio medical dataset to reduce the amount of samples to be presented to the neural network. It improves accuracy and computation time when applied to the publicly available benchmark bio medical dataset.**

Keywords— **Neural Network, Data mining, back propagation, Multilayer perceptron.**

I. Introduction

Data The scope of data mining is the knowledge extraction from large databases. It is an interdisciplinary area of research in databases, pattern recognition, information retrieval, machine learning, parallel, visualization and distributed computing. There are numerous applications in data mining such as customer relationship management, market and industry characterization, pharmacology, medicine, stock management, and biology. The mission of bioinformatics as a new and critical research domain is to use them to extract accurate and reliable information in order to gain new biological insights. Bioinformatics demands and provides the opportunities for novel and improved data mining methods development.

Data mining classification is a process of identification in that raw data are converted to categorized meaningful information. It can be divided into two groups: supervised and unsupervised. Many classification techniques including decision tree, neural network (NN), and support vector machine (SVM) and other rule based classification systems have been proposed. Neural network classification is supervised practical approach with lots of success in several classification tasks. However, its classification efficiency and accuracy is generally a problem, which discusses in this paper. It can be used to model complex relationships between inputs and outputs or to find patterns, High-quality data, the "right" data; an adequate sample size and the right tool are required to data-mine effectively.

Neural networks represent a brain metaphor for information processing and how the brain actually functions. An artificial neural network (ANN), often just called a neural network. It is a mathematical model or computational model based on biological neural networks and biological neural system for data analysis. Neural networks provide a very general way of approaching problems. The construction of a Neural Network includes structure, encoding, and recall factors. Many Neural network models have been successfully applied to the solution of complex problems related to signal processing, classification, clustering, feature selection, data visualization, data mining, and information fusion. The aim is to synergistically merge the techniques with enhancing the overall performance of the neural network classification algorithm. The neural networks do not give explicit knowledge representation in the form of rules, large sample size, over fitting data, complicated relations between the input variables and the output variables, slow learning process. To avoid this kind of ill goes through clustering the information which present in data mining concepts.

II. RELATED WORK

The quality and size of the biomedical training samples are crucially important for classification. The more representative samples introduced to a classification process produces more accurate and reliable results. A small sample size is not enough for a neural network to recognize all classes, whereas a large number of sample patterns can make the network over specific and require more computation time for training. The popular neural networks can be trained to recognize the data directly, whereas in simple networks there is a chance of the system being complex and training may be difficult. The time taken and the accuracy of classification depend on the dimension of the input given and dimension in the training data.

Artificial Intelligence is a branch of science which deals with helping machines finds solutions to complex problems in a more human-like fashion. Artificial neural networks have been widely used for many classification purposes and generally proved to be more powerful than conventional techniques. ANNs require decisions on the part of the user which may affect the accuracy of the resulting classification. More accurate techniques based on Bayes theory, nearest neighbors, Rule based induction and neural networks were developed. Combined methods integrated with neural network or nearest neighbor approaches improve prediction accuracy. Many attempts have been made to speed up the convergence and improve the accuracy of neural network classification.

III. METHODOLOGY

In this study attempt to introduce a classification approach using the Multi-Layer Perceptron (MLP), Radial Basics Function, Rule Induction, K Nearest Neighbor with Back-Propagation learning algorithm and a feature selection using information gain are used to classify biomedical bench mark data. The limitations of applying neural networks to analyse massive data mining databases require excessive processing. The characteristics of the selected training data are importance for the performance of a supervised classification process. To represent sample set of data selection must be performed with large number of sample patterns can make the network over specific and require more computation time for training in the learning process.

The data have become available from different sources represents different characteristics which improves the accuracy of classification significantly. This classification has been made a lot of development, and the common algorithms for classification include K nearest neighbor algorithm (KNN), Bayes algorithm, Support Vector Machine algorithm (SVM), Decision Tree algorithms, Neural Network algorithm (Nnet), Boosting algorithm, etc. KNN is one of the most popular and extensive among these, but it still has many defects, such as great calculation complexity. In order to reduce the high calculation complexity, this paper used clustering method and chosen the cluster center as the representative points which made the training sets become smaller, so this algorithm reducing the complexity of traditional KNN algorithm, increase the efficiency and improve the accuracy of the algorithm.

The constructed Algorithm

```
1- Read the actual dataset of n attributes.
2- Divide the actual dataset into two subsets:
        a. The first subset includes only tuples
           without noisy values of n attributes.
        b. The second subset includes only tuples
           with missing values of n attributes.
3- Find the reduced subset of the first dataset.
4- Reduced second subset by keeping only the attributes that were resulted from the reduced first data set and store the result in second reduced set.
5- Merge the reduced datasets which passed into a ready-to-train dataset.
6- Apply different methods of training in order to find the conclusion. such that:
        Conclusion ← Train(RtTDS).
    a) Find the conclusion based on Rule Induction
       such that:
        (Conclusion)RI ← Train RI (RtTDS)
```

b) Find the conclusion based on K-Nearest Neighbor (KNN) such that:
(Conclusion)KNN←Train KNN(RtTDS)
c) Find the conclusion based on Radial Basics Function (RBF) such that:
(Conclusion)RBF ← Train RBF (RtTDS)
d) Find the conclusion based on Multi Layer Perceptron (MLP) such that:
(Conclusion)MLP ← Train MLP (RtTDS)

Clustering is a data mining technique that separates the data into groups whose component belongs mutually. Every object is assigned to one cluster which is most similar. Clustering does not require a prior knowledge of the groups that are formed and the members who must belong to it. Cluster analysis encompasses a number of different classification algorithms that can be used to organize observed data into meaningful structures. Each object iteratively assigned to one cluster, and the center of each cluster is the mean of its assigned objects. In general, the k-means technique will produce exactly k different clusters. The traditional KNN classification algorithm produces a high degree of calculation complexity. Improved KNN classification algorithm based on clustering center is proposed in the given training sets are clustered by k-means clustering techniques, and all cluster centers are taken as the new training samples, which indicate the importance of each training sample according to the number of samples in the cluster that contains this cluster center. The modified samples are used to accomplish KNN classification improves the accuracy of KNN algorithm.

Neural network methods provides highly accurate, noise-resistant of biomedical data recognition which enhance the learning capabilities and reduce the computation intensity of a competitive learning algorithm. The proposed model use multi-layered network architecture with a back propagation learning mechanism applied to the training dataset to reduce the amount of samples to be presented to the neural network by automatically selecting optimal samples. The attained results show that the proposed technique performs exceptionally in terms of both accuracy and computation time. This proposed algorithm used to divide into partitioning methods, in which the classes are commonly exclusive. Each object is a member of the cluster with which it is most similar; each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains with each object resides in its own cluster. To read the original data set and divides into two data sets which one contains without missing values and other contains with missing values. Reduce the two data subsets and merge these two reduced data set which is used for training data set. This training data set is used to make the decision. For clustering process make the centroid, similarity, threshold value, corresponding cluster, any objects remain to be clustered, repeat the procedures.

IV. RESULTS & DISCUSSION

Rule induction is a data mine system has to infer a model from the dataset that it may define classes rule induction, K-Nearest Neighbor (KNN) category, Radial Basics Function and Multi Layer Perceptron. The purpose of this algorithm is to classify a new data based on attributes and training samples. In processing elements in an ANN operate concurrently and collectively in a similar fashion to biological neurons. The obtained results will demonstrate the proposed technique performs exceptionally in terms of both accuracy and computation time when applied to the publicly available on benchmark biomedical dataset compared to a standard learning schema that use the full dataset. Clustering technique in neural network improves the accuracy and efficiency and easy to use. Cluster analysis encompasses a number of different classification algorithms that can be used to organize observed data into meaningful structures.

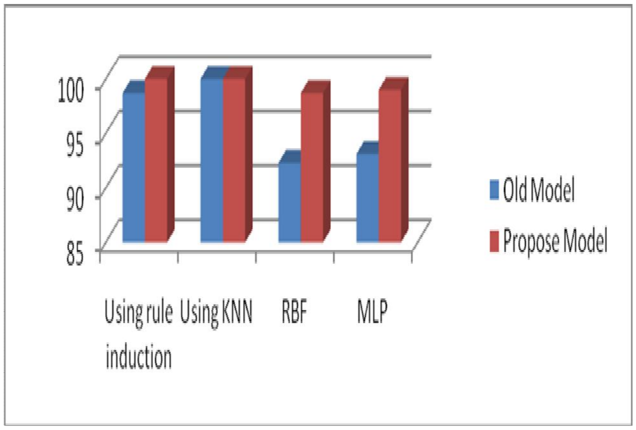

Fig. 1 Coverage values comparison results

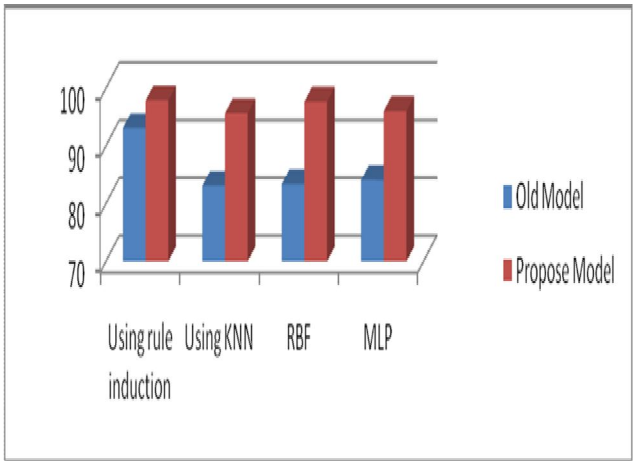

Fig. 2 Accuracy based comparison results

V. SUMMARY

The classification efficiency of neural network is analysed. Classification systems can help in increasing accuracy and reliability of the biomedical facts. Artificial Intelligence has led to the emergence of Decision Support Systems for medical applications. Neural networks able to derive knowledge from complicated or imprecise data used to extract patterns and trends are too complex to be noticed. It is to build a combined architecture of neural network and data mining techniques to automate the processes involving decision-making of large test sets. Neural Network is able to solve highly complex problems due to the inherent modularity of the structure makes it adaptable to applications.

Neural network closely follows the biomedical classifications to find the neural network having the best performance on data, the simplest approach to the comparison of different networks is to evaluate the error function using data which is independent of that used for training. The accuracy achieved by a supervised classification is largely dependent upon the training data. It is obtaining an increase of individual training classes by applying k-nearest neighbor which effectively simplifies the training data, removes outliers and inconsistencies. This approach can reduce the over fitting effect and increase the resulting classification accuracy. It includes network type, size and architecture, training step size, learning algorithms, and data representation. Multilayer is feed-forward neural networks trained with the standard back-propagation algorithm. It is supervised networks so they require a desired response to be trained. Various networks are trained by minimization of an appropriate error function defined with respect to a training data set. An important issue concerning the application of clustering methods in biomedical data is the assessment of cluster quality.

V. CONCLUSION & FUTURE WORK

The classification efficiency of the neural networks is compared. According to the arrived results, the performance of the supervised neural network cluster based classification approach provided significant performance. Neural Network is able to solve highly complex problems due to the non linear processing capabilities. In addition, the inherent modularity of the neural network structure makes it adaptable to a wide range of applications. The proposed algorithm improves accuracy and computation time when applied to the biomedical dataset. Furthermore, it is easy to use data from different sources in the ANN classification procedure to improve the accuracy of the classification.

REFERENCES

[1] S. Mitra, S. Datta, T. Perkins, G. Michailidis “Introduction to Machine Learning and Bioinformatics” Chapman & Hall/CRC Press, 2008.

[2] Drazen.S. “Estimation of difficult to Measure process variables using neural networks” Proceedings of IEEE MELECON 2004,May 12-15, Dubrovnik, Croatia .

[3] Moreno.L, “Brain maturation estimation using neural classifier” IEEE Transaction of Bio Medical Engineering” vol 42, no 2, April 1995.

[4] Tarassenko.L, Y.U.Khan, M.R.G.Holt, “Identification of inter-ictal spikes in the EEG using neural network analysis” IEEE

Proceedings –Science Measurement Technology, vol 145, no 6, November 1998.

[5] H.Demuth and M.Beale, "Neural network tool box: User's guide, Version 3.0" Natick, MA, 1998.

[6] K. M. Faraoun, A. Boukelif "Neural Networks Learning Improvement using the K-Means Clustering Algorithm to Detect Network Intrusions" International Journal of Computational Intelligence Volume 3 Number 2.

[7] Imdad Ali Rizvi, B.Krishna Mohan "Improving the Accuracy of Object Based Supervised Image Classification using Cloud Basis Function Neural Network for High Resolution Satellite Images" International Journal of Image Processing (IJIP), Volume (4) Issue (4),

[8] A. Dallali, A. Kachouri, M. Samet "Fuzzy c-means clustering, Neural Network, wt, and Hrv for classification of cardiac arrhythmia" ARPN Journal of Engineering and Applied Sciences VOL. 6, NO. 10, OCTOBER 2011.

[9] R. Alejo, V. Garcia, J.M. Sotoca, R.A. Mollineda, J.S. Sánchez "Improving the Classification Accuracy of RBF and MLP Neural Networks Trained with Imbalanced Samples" Springer-Verlag Berlin Heidelberg 2006.

[10] B.Madasamy, Dr. J. Jebamalar Tamilselvi, "Optimal Data mining Classification Algorithm for Bio Medicinal Facts" International Journal of Advanced Computing Engineering Applications (IJACEA) February -2013.

[11] B.Madasamy, Dr. J. Jebamalar Tamilselvi,"General Web Knowledge Mining Framework" International Journal of Computer Science and Engineering (IJCSE) October -2012.

[12] B.Madasamy, Dr. J. Jebamalar Tamilselvi,"Assessment of Freeware Data Mining Tools over some Wide Range Characteristics", SPRINGER Verlag Journal, ICIP - August 2012.

[13] B.Madasamy, Dr. J. Jebamalar Tamilselvi, "Wide Range Data Mining Application Framework using IT Decision Making" International Conference on Recent Trends in Computing Technology April 2013.

[14] B.Madasamy, Dr. J. Jebamalar Tamilselvi, "Taxonomy of Ageing in addition to non-Ageing Genes by means of General Data Mining Framework" International Journal of Computer Applications (IJCA) June - 2013.

[15] B.Madasamy, Dr. J. Jebamalar Tamilselvi, "General Framework for Biomedical Facts through Data Mining Practices" International Journal of Computer Trends and Technology (IJCTT) May – 2013.

[16] Guoqiang, Peter Zhang "Neural Networks for Classification: A Survey" IEEE transactions on systems, man, and cybernetics applications and reviews, vol. 30, no. 4, November 2000.

[17] Anchana Khemphila, Veera Boonjing "Parkinsons Disease Classification using Neural Network and Feature selection" World Academy of Science, Engineering and Technology 2012.

[18] Dr. K. Usha Rani "Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique" International Journal of Computer Applications Volume 10– No.3, November 2010.

[19] Chady El Moucary Marie Khair "Improving Student's Performance Using Data Clustering and Neural Networks in Foreign-Language Based Higher Education" The Research Bulletin of Jordan ACM , Vol 2.

[20] R.HariKumar, N.S.Vasanthi, M.Balasubramani "Performance Analysis of Artificial Neural Networks and Statistical Methods in Classification of Oral and Breast Cancer Stages" International Journal of Soft Computing and Engineering (IJSCE) Volume 2, Issue 3, July 2012

[21] Zhou Yong Li Youwen and Xia Shixiong "An Improved KNN Text Classification Algorithm Based on Clustering" Journal of computers, vol. 4, no. 3, march 2009.

[22] Zhenxiang Chen, Bo Yang, Yuehui Chen, Lin Wang, Haiyang Wang, Ajith Abraham and Crina Grosan "Improving neural network classification using further division of recognition space" International Journal of Innovative Computing, Information and Control ICIC International Volume 5, Number 2, February 2009.

[23] Dr. Yashpal singh, Alok singh chauhan "Neural networks in Data mining" Journal of Theoretical and Applied Information Technology" - 2009 JATIT. www.jatit.org

[24] Hyunsoo Yoon, Cheong-Sool Park, Jun Seok Kim, Jun-Geol Baek, "Algorithm learning based neural network integrating feature selection and classification" Expert Systems with Applications 40- 2013