Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Efficient and Robust Ensemble Method for High Dimensional Data Classification Using Radial Basis Functions Neural Network

Shweta B. Meshram, Prof. Sharmila M. Shinde
Computer Engineering Department, JSCOE,
Pune, Maharashtra, India

Abstract— As we are living in data age, the mountain of data growing exponentially. But the main issue is to identify useful information from it. Data mining provides two ways to recognize valuable information i.e., supervised and unsupervised learning methods. In supervised methods, ensemble methods performing exceptionally good.
This paper proposes here the ensemble method to classify high-dimensional data. This method will generate the independent feature subsets. Each independent feature subset is trained using base classifier. These results are combined by majority voting. The proposed method uses the radial basis function neural network to classify data of biomedicine.

Keywords—biomedicine, classification, ensemble method, high dimensional data, Radial basis function neural network.

I. INTRODUCTION

In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms. Unlike a statistical ensemble in statistical mechanics, which is usually infinite, a machine learning ensemble refers only to a concrete finite set of alternative models, but typically allows for much more flexible structure to exist among those alternatives.

Supervised learning algorithms are commonly described as performing the task of searching through a hypothesis space to find a suitable hypothesis that will make good predictions with a particular problem. Even if the hypothesis space contains hypotheses that are very well-suited for a particular problem, it may be very difficult to find a good one. Ensembles combine multiple hypotheses to form a (hopefully) better hypothesis. The term ensemble is usually reserved for methods that generate multiple hypotheses using the same base learner. The broader term of multiple classifier systems also covers hybridization of hypotheses that are not induced by the same base learner.

Evaluating the prediction of an ensemble typically requires more computation than evaluating the prediction of a single model, so ensembles may be thought of as a way to compensate for poor learning algorithms by performing a lot of extra computation. Fast algorithms such as decision trees are commonly used with ensembles (for example Random Forest), although slower algorithms can benefit from ensemble techniques as well.

The paper consists of mainly three parts. First parts describe the literature survey. Second part discusses about the proposed ensemble method in brief. Lastly result discussion is explained of the proposed method.

II. LITERATURE SURVEY

One of the ensemble method is bagging that was proposed by L. Breiman. It results in poor performance. There was a need to improve the performance of boosting. Hence AdaBoost.M1 and AdaBoost.M2 are proposed. Breiman proposed the "random forest" . But it cannot predict beyond the range of training data.

S.B. Cho, H. Won proposed ensemble of neural networks. The accuracy of the classification doesn't increase much. H. I. Elshazly, A.M. Elkorany, A. E. Hassanien proposed another ensemble method. It affects originality of features. Y. Piao, H. W. Park, C. H. Ji, K. Ho Ryu proposed the ensemble method that uses the Fast Correlation- Based Filter method (FCBF) to generate multiple feature subsets. The computation time of ensemble method is improved, but it affects the accuracy of the classification. Hence, new ensemble method is proposed to improve accuracy.

III. PROPOSED ENSEMBLE METHOD

There are many proposed ensemble methods that deals with high dimensional data. High dimensional data are the data that are characterized by few dozen to many thousands of dimensions i.e. nothing but features. Most of ecological data, data on health status of patients, movie rating data, climate data, bioinformatics data are examples of the high dimensional data.

The above figure represents the architecture of the proposed ensemble method. We are proposing the efficient ensemble method for high dimensional data classification. The ensemble method will first generate the different feature subsets, these subsets are trained using base classifiers. The proposed method use the radial basis function neural network as the base classifier. The majority voting is used to combine the decisions of classifiers and accordingly classes will be assigned.

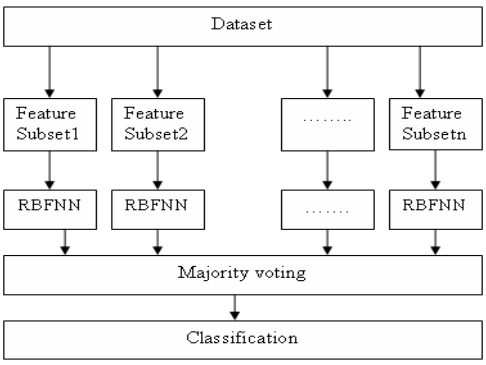

Fig. Architecture of proposed system

A. Subsets Generation:

There are many methods to generate feature selection. The proposed method uses the correlation based feature selection method which is the most accurate one when it is compared with other methods. At the heart of the CFS algorithm is a heuristic for evaluating the worth or merit of a subset of features. This heuristic takes into account the usefulness of individual features for predicting the class label along with the level of intercorrelation among them. The hypothesis on which the heuristic is based can be stated.

Good feature subsets contain features highly correlated with (predictive of) the class yet uncorrelated with (not predictive of) each other

other_

The CFS ranks the feature subsets instead of individual features by merit function

$$Merit_s = \frac{k\overline{C}_{cf}}{\sqrt{k + k(k-1)\overline{C}_{ff}}}$$

n. It is given as follows:

It is the merit function of the s feature subset with k features. $\overline{C_{cf}}$ represents the mean class-feature correlation and $\overline{C_{ff}}$ represent the mean feature-feature correlation.

B. Classification of each model:

Each feature subset is learned using base classifier. The proposed method uses the radial basis function neural network as base classifier. Each feature subset is trained using the radial basis function neural network as base classifier.

In the field of mathematical modelling, a radial basis function network is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have many uses, including function approximation, time series prediction, classification, and system control.

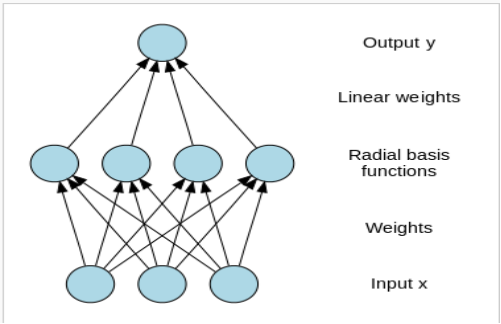

Fig. General architecture of the RBFNN

The idea of Radial Basis Function (RBF) Networks derives from the theory of function approximation. We have already seen how Multi-Layer Perceptron (MLP) networks with a hidden layer of sigmoidal units can learn to approximate functions. RBF Networks take a slightly different approach. Their main features are:

1. They are two-layer feed-forward networks.
2. The hidden nodes implement a set of radial basis functions (e.g. Gaussian functions).
3. The output nodes implement linear summation functions as in an MLP.
4. The network training is divided into two stages: first the weights from the input to hidden layer are determined, and then the weights from the hidden to output layer.
5. The training/learning is very fast.
6. The networks are very good at interpolation.

Hence the proposed system uses the radial basis function neural network as base classifier.

C. Combining the results of each classifier:

Majority voting is used for combining the results of each classifier. In this combining scheme, a classification of unlabelled is performed according to the class that obtains maximum votes. This method is also known as Plurality votes.

IV. RESULT DISCUSSION

The proposed system is robust as it can handle the binary as well as multiclass classification. The following table represents the comparison of the proposed system with the other existing system.

We are using the breast cancer wisconsin which is the binary classification problem. There are total 699 instances in dataset.

Table 1.Comparison of the proposed system

Sr. No.	Ensemble Method	Originality of Features	Computation Time
1	S.B. Cho, H. Won	Preserved	Large
2	H. I. Elshazly, A.M. Elkorany	Not preserved	Medium
3	Proposed method	Preserved	Less

The following column chart represents the accuracy of the proposed system with the other ensemble methods for analysis. We are using weka tool for analysis purpose.

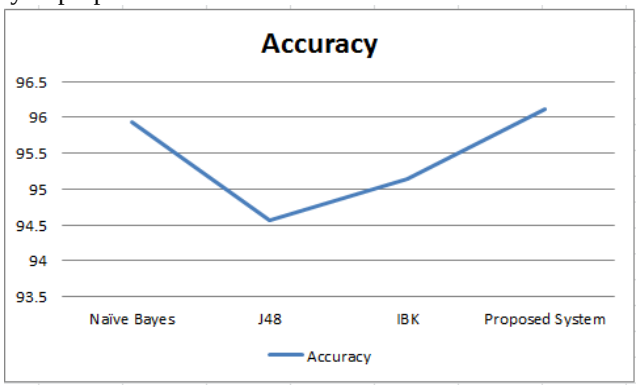

Fig. Comparison of the accuracy of the proposed method

IV. CONCLUSIONS

The aim of our method is to handle the classification task more efficiently and robustly with comparative accuracy. The proposed system preserves the originality of the features. Inconsistent and missing values were resolved before model construction but in real life that is not the case. The proposed system mainly consists of feature extraction, training and classification.

REFERENCES

[1] Yongjun Piao, Minghao Piao, Kiejung Park and Keun Ho Ryu," An Ensemble Correlation-Based Gene Selection Algorithm for Cancer Classification with Gene Expression Data", Bioinformatics Advance Access published October 11, 2012.

[2] Mohammad Ali Bagheri, Qigang Gao, Sergio Escalera,"A Framework towards the Unification of Ensemble Classification Methods", 2013 12th International Conference on Machine Learning and Applications.

[3] L. Breiman,"Bagging predictors", Mach. Learning. 24, 1996, pp.123- 140.

[4] Y. Freund, R.E. Schapire,"Experiments with a new boosting algorithm", International Conference on Machine Learning, 1996, pp.148-156.

[5] Leo Breiman," Random Forests", Statistics Department, University of California, Berkeley, CA 94720.

[6] Sung-Bae Cho, Hong-Hee Won,"Cancer classification using ensemble of neural networks with multiple significant gene subsets", published online: 12 November 2006 Springer Science & Business Media, LLC 2007.

[7] Yongjun Piao, Hyun Woo Park, Cheng Hao Ji, Keun Ho Ryu,"Ensemble Method for Classification of High-Dimensional Data", 978-1-4799-3919-0/14/ IEEE Big Comp 2014.

[8] Sung-Bae Cho, Hong-HeeWon,"Cancer classification using ensemble of neural networks with multiple significant gene subsets", published online: 12 November 2006 Springer Science+Business Media, LLC 2007.

[9] M. A. Hall,"Correlation based Feature Selection (CFS) for Discrete and Numeric Class Machine Learning", working paper series ISSN 1170-487X, My 2000.

