Enhancing BPN Performance using GA Identified Significant Features: A Case Study for Categorization of Heart Statlog Dataset

Asha Gowda Karegowda
Dept. of Master of Computer Applications,
Siddaganga Institute of Technology, Tumkur, Karnataka, India

ABSTRACT

Feature selection is an essential pre-processing method to remove irrelevant and redundant data. This paper presents the development of a model for classifying Heart Statlog. The model consists of two stages. In the first stage, genetic algorithm (GA) is used as random search method with Correlation based feature selection as fitness function for identifying significant features. The second stage a fine tuned classification is done using back propagation neural network using GA identified feature subset elicited in the first stage. Experimental results signify that the feature subset identified by the proposed filter when given as input to Back propagation neural network classifier, leads to enhanced classification accuracy.

Keywords

Feature selection, Filter approach, Genetic Algorithm, Correlation based feature selection, Back propagation neural network.

1. INTRODUCTION

Medical data mining has enormous potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. Data preprocessing is a significant step in the knowledge discovery process, since quality decisions must be based on quality data.

Data preprocessing includes data cleaning, data integration, data transformation and data reduction. These data processing techniques, when applied prior to mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining. The goal of data reduction is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Mining on the reduced set of attributes has additional benefits. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand. Further it enhances the classification accuracy and learning runtime [1].

This paper presents use of multivariate filters which used GA with correlation based feature selection as fitness evaluator. The relevant features are provided as input to neural network trained using back propagation method. Section 2 discusses wrapper and filter feature selection methods for both supervised learning algorithms.
For the sake of completeness Back propagation neural network has been described in Section 3. Section 4 describes Genetic search algorithm (GA) and Correlation based feature selection (CFS) as subset evaluating mechanism for GA followed by experimental results and conclusions in Section 5 and Section 6 respectively.

2. FEATURE SELECTION

Feature subset selection is of immense importance in the field of data mining. The increased dimensionality of data makes testing and training of general classification method difficult. In supervised learning, feature selection aims to maximize classification accuracy. The goal 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 in unsupervised learning is much harder problem, due to the absence of class labels.

For supervised learning the feature selection can be classified into two types, filter methods and wrapper methods. Filter method assess the relevance of the attributes based on data's intrinsic properties. Filter methods are independent of learning algorithm, hence once the significant features are identified by the filter can be provided as input to different learning algorithm.

Further filter methods can be classified as univariate filters and multivariate filters. Univariate filters are fast, scalable and simple. The downside of univariate filter is they ignore the feature dependencies which may affect the classification accuracy.
This downside is overcome in multivariate filters , but these are less scalable and slow compared to univariate filters[3,4]. Authors have proposed a multivariate filter which used GA with correlation based feature selection as fitness evaluator, and validations has been carried out by various classifiers like Decision tree, Naïve Bayes, Bayesian classifier, Radial Basis function and K-nearest neighbor algorithm for different medical data sets [5] .

In this paper the GA with correlation based feature selection has been used to provide the significant inputs to back propagation neural network. Wrapper method in supervised learning uses the method of classification itself to measure the importance of feature set, hence the features selected depends on the classifier model used i.e. the feature subset search algorithm is wrapped around the learning model. In other words, the wrapper approach usually conducts a subset search with the optimal algorithm and then a classification algorithm is used to evaluate the subset.
Hence the pertinent feature identified by a wrapper method cannot be provided as input to different learning methods [3, 4].

3. BACK PROPAGATION NEURAL NETWORKS

Neural network (NN) is an information-processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. Developing a neural network involves first training the network to carry out the desired computations. During the learning phase, training data is used to modify the connection weights between pairs of nodes so as to obtain a best result for the output nodes(s). The feed-forward neural network architecture is commonly used for supervised learning. Feed-forward neural networks contain a set of layered nodes and weighted connections between nodes in adjacent layers. Feed-forward networks are often trained using a back propagation-learning scheme. Back propagation learning works by making modifications in weight values starting at the output layer then moving backward through the hidden layers of the network. Neural networks have been criticized for their poor interpretability, since it is difficult for humans to interpret the symbolic meaning behind the learned weights. Advantages of neural networks, however, include their high tolerance to noisy data as their ability to classify patterns on which they have not been trained [1, 6-9].

4. FILTER APPROACH FOR FEATURE SELECTION USING GENETIC ALGORITHM

GA is a stochastic general search method, capable of effectively exploring large search spaces, which is usually required in case of attribute selection. Further, unlike many search algorithms which perform a local, greedy search, GAs performs a global search. A genetic algorithm mainly composed of three operators: reproduction, crossover, and mutation. Reproduction selects good string; crossover combines good strings to try to generate better offspring's; mutation alters a string locally to attempt to create a better string. In each generation, the population is evaluated and tested for termination of the algorithm. This procedure of selection, crossover and mutation is continued until the termination criterion is met[10].

In this paper WEKA GA is used as search method with CFS as subset evaluating mechanism (fitness function). The features selected by filter GA-CFS are provided as input to the classifier, back propagation neural network. The proposed method is shown in figure 1.

The downside of univariate filters like information gain is, it does not account for interactions between features, which is overcome by multivariate filters like CFS. CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Correlation coefficients is used to estimate correlation between subset of attributes and class, as well as inter-correlations between the features. Relevance of a group of features grows with the correlation between features and classes, and decreases with growing inter-correlation [11]. CFS is used to determine the best feature subset and is usually combined with search strategies such as forward selection, backward elimination, bi-directional search, best-first search and genetic search. Equation for CFS is given is equation 1.

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \qquad (1)$$

where r_{zc} is the correlation between the summed feature subsets and the class variable, k is the number of subset features, rzi is the average of the correlations between the subset features an the class variable, and rii is the average inter-correlation between subset features[11]

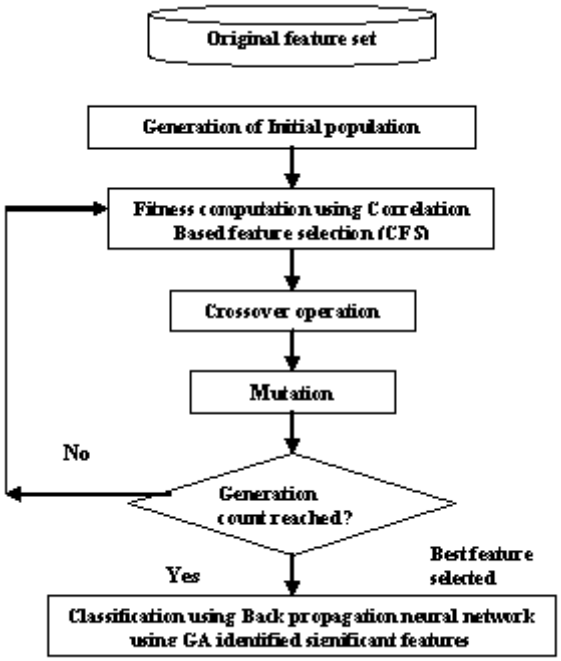

Figure 1. GA based filter approach for feature subset selection for Back propagation neural network classifier

5. RESULTS

The data used for the model is heart-statlog available in UCI machine learning dataset. Heart statlog includes 270 samples with the following attributes (13 attributes as input and last attribute as target variable) age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy and thal. The class label is Absence or presence of heart disease.

As a part of feature selection steps a multiivariate filter: Genetic algorithm with Correlation based feature selection as subset evaluating mechanism has been applied for Heart Statlog dataset. For GA, the population size is varied in the range of 10- 30, number of generation is the range of 10-30, crossover rate and mutation rate is set as 0.6 and 0.033 respectively.

GA with CFS resulted in features subset of 7 attributes namely chest pain type, resting electrocardiographic results,

exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy and thal. The GA identified 7 attributes were provided as input to back propagation neural network.

Experiments were carried out with back propagation neural network by varying the number of hidden nodes in the range of 5-15, number of epochs in the range of 100-500, with both all attributes and GA identified attributes.Figure 2 shows the comparative graph showing the classification accuracy obtained for best topologies by back propagation neural network with all attributes and GA identified attributes of heart statlog dataset for different number of epochs. The best performance of BPN with all features is found to be 90.37% with 13-8-1 topology for 200 epochs.

The best performance of BPN with GA identified significant features is found to be 85.19% with 7-5-1 topology for 100 of epochs. Figure 2 clearly depicts that classification accuracy of BPN with all features is less compared to superior classification accuracy of BPN with GA identified features.

The confusion matrix for BPN with all features and with GA identified features for Heart Statlog dataset is shown in Table 1. In addition the performance of BPN with all features and with GA identified features is measured using TP Rate, FP Rate, Precision, Recall, and F-Measure for both absent and present class label of Heart statlog dataset is as shown in Table 2.

6. CONCLUSIONS

This paper illustrated the importance of significant features identification for enhancing the performance of back propagation neural network. The filter approach for feature selection is carried out using Genetic algorithm with Correlation-based features subset for fitness evaluation. The experimental results clearly depict that employing feature subset selection using GA has improved the classification accuracy of back propagation neural network by an order of 5% for Heart statlog dataset.

7. REFERENCES

[1] J. Han And M. Kamber 2001. Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.

[2] Jennifer G. Dy, Feature Selection for Unsupervised Learning, Journal of Machine Learning ,pp845-889,2004

[3] Shyamala Doraisamy ,Shahram Golzari ,Noris Mohd. Norowi, Md. Nasir B Sulaiman ,Nur Izura Udzir A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music

[4] Y.Saeys, I.Inza, and P. Larrannaga, " A review of feature selection techniques in bioinformatics", Bioinformatics, 23(19),2207, pp.2507-2517.

[5] Asha Gowda Karegowda, M.A.Jayaram A.S .Manjunath, "Feature Subset Selection using Cascaded GA & CFS: A Filter Approach in Supervised Learning. International Journal on Computer Applications(IJCA) Volume 1 , 2011 June

[6] A. Roy, Artificial neural networks – a science in trouble, SIGKDD Explorations, 1:33-38, (2000)

[7] S. Haykin, Neural Networks- A comprehensive foundation, Macmillan Press, New York, (1994).

[8] D.E. Rinehart, G.E. Hinton, and R. J. Williams, Learning internal representations by error propagation, Parallel Distributed Processing, Cambridge, MA: MIIT Press, (1986).

[9] H. Lu, R. Setiono and H. Liu, Effective data mining using neural networks, IEEE Trans. On Knowledge and Data Engineering, 5: 8, (1996)

[10] D. Goldberg, Genetic Algorithms in Search, Optimization , and Machine learning, Addison Wesley, 1989.

[11] Mark A. Hall, Correlation-based Feature Selection for Machine Learning.

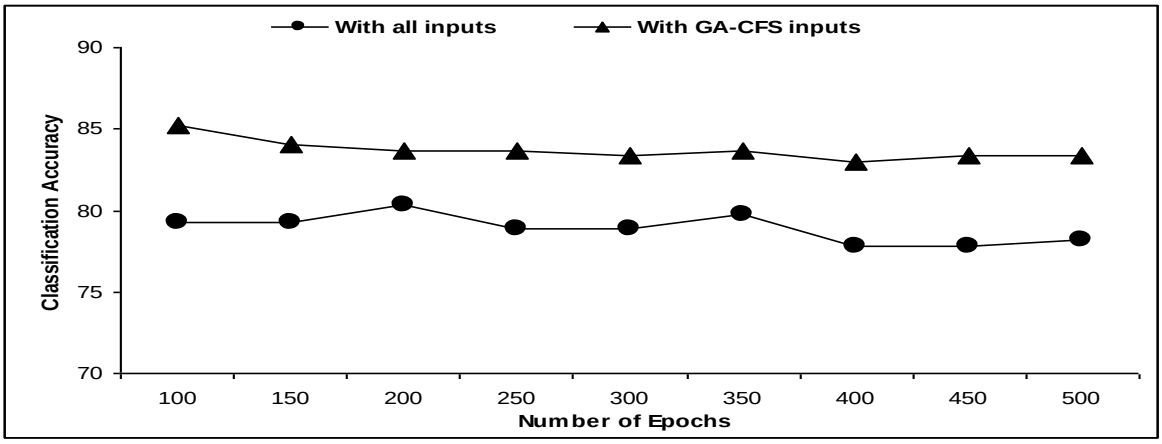

Figure 2. Classification Accuracy of BPN with all features and GA identified features for different epochs

Table 1. Confusion matrix for BPN with all and GA identified feature for Heart Statlog dataset.

Confusion matrix					
With All Features			With GA identified features		
	Absent	Present		Absent	Present
Absent	124	26	Absent	135	15
Present	27	93	Present	25	95

Table 2. Comparative performance of BPN with all and GA identified feature for Heart Statlog dataset.

Feature Selection	Number of features	Heart Statlog Class	TP Rate	FP Rate	Precision	Recall	F-Measure
With All features	13	Absent	0.827	0.225	0.821	0.827	0.824
		Present	0.775	0.173	0.782	0.775	0.778
With GA-CFS identified features	7	Absent	0.9	0.208	0.844	0.9	0.871
		Present	0.792	0.1	0.864	0.792	0.826