International Journal of Information Technology and Knowledge Management
July-December 2010, Volume 2, No. 2, pp. 271-277

COMPARATIVE STUDY OF ATTRIBUTE SELECTION USING GAIN RATIO AND CORRELATION BASED FEATURE SELECTION

Asha Gowda Karegowda[1], A. S. Manjunath[2] & M.A.Jayaram[3]

Feature subset selection is of great importance in the field of data mining. The high dimension data makes testing and training of general classification methods difficult. In the present paper two filters approaches namely Gain ratio and Correlation based feature selection have been used to illustrate the significance of feature subset selection for classifying Pima Indian diabetic database (PIDD). The C4.5 tree uses gain ratio to determine the splits and to select the most important features. Genetic algorithm is used as search method with Correlation based feature selection as subset evaluating mechanism. The feature subset obtained is then tested using two supervised classification method namely, Back propagation neural network and Radial basis function network. Experimental results show that the feature subsets selected by CFS filter resulted in marginal improvement for both back propagation neural network and Radial basis function network classification accuracy when compared to feature subset selected by information gain filter.

Keywords: Feature Selection, Gain Ratio, Correlation based Feature Selection, Back Propagation Neural Network, Radial basis Function Network

1. Introduction

Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. 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. Data preprocessing is a significant step in the knowledge discovery process, since quality decisions must be based on quality data. 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 two filters namely information gain and correlation based feature selection methods for feature selection. The relevant features are provided as input to two supervised classification method: back propagation neural network and Radial basis function network. Section 2 discusses wrapper & filter feature selection methods, and also briefs about how decision tree uses gain ratio measure for feature selection. Further section 3 describes Genetic search algorithm (GA) with Correlation based feature selection (CFS) as subset evaluating mechanism. For the sake of completeness back propagation neural network and radial basis function network have been discussed in section 4 and section 5 respectively, followed by results in section 5 and conclusion in section 6.

[1,2,3]Siddaganga Institute of Technology, Tumkur-572103, Karnataka

Email: ashagksit@gmail.com, asmanju@gmail.com, ma_jayaram@rediffmail.com

2. Feature Selection

Methods used for data reduction can be classified into two types: Wrapper and filter method. Wrapper model approach uses the method of classification itself to measure the importance of features set, hence the feature selected depends on the classifier model used. Wrapper methods generally result in better performance than filter methods because the feature selection process is optimized for the classification algorithm to be used. However, wrapper methods are too expensive for large dimensional database in terms of computational complexity and time since each feature set considered must be evaluated with the classifier algorithm used. The filter approach actually precedes the actual classification process. The filter approach is independent of the learning algorithm, computationally simple fast and scalable. Using filter method, feature selection is done once and then can be provided as input to different classifiers. Various feature ranking and feature selection techniques have been proposed such as Correlation-based Feature Selection (CFS), Principal Component Analysis (PCA), Gain Ratio (GR) attribute evaluation, Chi-square Feature Evaluation, Fast Correlation-based Feature selection (FCBF), Information gain, Euclidean distance, i-test, Markov blanket filter. Some of these filter methods do not perform feature selection but only feature ranking hence they are combined with search method when one needs to find out the appropriate number of attributes. Such filters are often used

with forward selection, which considers only additions to the feature subset, backward elimination, bi-directional search, best-first search, genetic search and other methods [2,3,4].

2.1. Gain Ratio

A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes. The information gain measure is used to select the test attribute at each node of the decision tree. The information gain measure prefers to select attributes having a large number of values. The basic decision tree induction algorithm ID3 [5] was enhanced by C4.5 [6, 7]. C4.5 a successor of ID3, uses an extension of information gain known as gain ratio, which attempts to overcome this bias. The WEKA [8] classifier package has its own version of C4.5 known as J4.8. We have used J4.8 to identify the significant attributes.

Let S be set consisting of s data samples with m distinct classes. The expected information needed to classify a given sample is given by

$$I(S) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

where pi is the probability that an arbitrary sample belongs to class Ci and is estimated by s_i/s.

Let attribute A has v distinct values. Let s_{ij} be number of samples of class C_i in a subset S_j. S_j contains those samples in S that have value a_j of A. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = -\sum_{i=1}^{m} I(S) \frac{s_{1i} + s_{2i} + \ldots s_{mi}}{s} \quad (2)$$

The encoding information that would be gained by branching on A is

$$\text{Gain}(A) = I(S) - E(A) \quad (3)$$

C4.5 uses gain ratio which applies normalization to information gain using a value defined as

$$\text{SplitInfo}_A(S) = -\sum_{i=1}^{v} (|S_i|/|S|) \log_2(|S_i|/|S|)$$

The above value represents the information generated by splitting the training data set S into v partitions corresponding to v outcomes of a test on the attribute A.

The gain ratio is defined as

$$\text{Gain Ratio}(A) = \text{Gain}(A) / \text{SplitInfo}_A(S)$$

The attribute with the highest gain ratio is selected as the splitting attribute[1]. The non leaf node of the decision tree generated are considered as relevant attributes. The authors have integrated decision tree and neural network, which resulted in improved classification accuracy [9].

The summary of decision tree algorithm is given:

i. Choose an attribute that best differentiates the output attribute values.

ii. Create a separate tree branch for each value of the chosen attribute.

iii. Divide the instances into subgroups so as to reflect the attribute values of the chosen node.

iv. For each subgroup, terminate the attribute selection process if:

(a) The members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with the specified value.

(b) The subgroup contains a single node or no further distinguishing attributes can be determined. As in (a), label the branch with the output value seen by the majority of remaining instances.

v. For each subgroup created in (iii) that has not been labeled as terminal, repeat the above process.

2.2. Correlation based Feature Selection (CFS)

The downside of univariate filters for eg information gain is, it does not account for interactions between features, which is overcome by multivariate filters for eg 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 [4]. 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.

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \quad (4)$$

where r_{zc} is the correlation between the summed feature subsets and the class variable, k is the number of subset features, r_{zi} is the average of the correlations between the subset features an the class variable, and r_{ii} is the average inter-correlation between subset features[4].

Genetic algorithm is used as search method with Correlation based feature selection as subset evaluating mechanism. In this paper WEKA GA is used as search method with CFS as subset evaluating mechanism (fitness function). Authors have used the features selected by filter GA-CFS as input to the neural network classifier. It has substantially increased the classification accuracy of neural network [10]. 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. If the termination criterion is not satisfied, the population is operated upon by the three GA operators and then re-evaluated. This procedure is continued until the termination criterion is met [11].

3. BACK PROPAGATION NEURAL NETWORK (BPN)

BPN 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 [12-15].

$$o_i = f(net_j) \tag{5}$$

$$net_j = \sum_i w_{ji} o_i + \theta_j \tag{6}$$

Where w_{ji} is the weight of the connection from the ith node to the jth node, o_i is the output of the ith node, θ_j is a variable bias with similar function to a threshold, and a summation is over all units feeding into node j. The activation function that is used is given by

$$o_j = \frac{1}{1+e^{-net_j}} \tag{7}$$

The term back propagation refers to an iterative training process in which an output error E is defined by

$$E = \sum_p E_p = \frac{1}{2}\sum_p \sum_j (t_{pj} - o_{pj})^2 \tag{8}$$

Where summation is performed over all output nodes j and t_j is the desired or target value of output o_j for a given input vector. The direction of steepest descent in parameter space is determined by the partial derivatives of E with respect to the weights and bias in the network,

$$\frac{\partial E}{\partial w_{ji}} = -\sum_p \partial_{pj} o_{pj} \tag{9}$$

$$\frac{\partial E}{\partial \theta_j} = -\frac{\partial E_p}{\partial net_{pj}} \tag{10}$$

And it can be shown that,

$$\partial_{pj} = (t_{pj} - o_{pj})(1 - o_{pj}) o_{pj} \tag{11}$$

For output nodes j and

$$\partial_{pj} = (1 - o_{pj}) o_{pj} \sum_k \partial_{pk} w_{kj} \tag{12}$$

For all nodes in the intermediate layer where j refers to a node in one of the intermediate layers, and the summation is over all units k, which receive a signal from node j.

Further, it can be shown that the learning rule is given by

$$\Delta w_{ij}(t) = \eta \sum_p \partial_{pj} o_{pj} \tag{13}$$

$$\Delta \theta_j(t) = \eta \sum_p \partial_{pj} \tag{14}$$

Where t denotes a given instant of time and η is the learning parameter.

4. RADIAL BASIS FUNCTION NETWORK

RBF network consists of three layers, an input layer, which reads n inputs, a hidden layer that consists of m radial basis functions, and an output layer consisting of a linear additive function, which produces the response. The input neurons are linear and pass the input along to hidden neurons without any processing. The input feeds forward to each hidden neuron. Using radial basis function the hidden neuron

computes the signal and pass on these signals through weighted pathways to the linear output neuron which sums these up and generates an output signal. [16]. The difference between the standard feed forward neural network and the radial basis function network is that hidden neurons compute signals based on distances to training data points rather than inner-products with weight vectors, and the signals from each of the hidden neurons are linearly superposed at an output neuron using tunable weights. [17]. The m neurons of hidden layer have a linear threshold activation function that provides the network output as follows.

$$y_i = \sum_{j=0}^{m} w_{ij}\phi_j(x) \tag{15}$$

where w_{ij} is the connection weight between the hidden neuron j and the output neuron i and $\Phi_j(x)$ is the radial basis function applied at neuron j of the hidden layer to input x.

Some of the commonly used radial basis functions are as fallows.

(i) Gaussian Functions

$$\phi(x) = \exp\left(\frac{-(x-c)^2}{2\sigma^2}\right),\ \sigma > 0;\ x, c \in R \tag{16}$$

(ii) Multiquadrics:

$$\phi(x) = (x^2 + c^2)^{1/2},\ c > 0;\ x, c \in R \tag{17}$$

(iii) Inverse Multiquadrics:

$$\phi(x) = 1/(x^2 + c^2)^{1/2},\ c > 0;\ x, c \in R \tag{18}$$

The most commonly used radial basis function is Gaussian basis function as shown in equation number 16 where c is the center, and σ is the spread factor which has a direct effect on the smoothness of the interpolating function.

5. THE DATA MODEL

Diabetes mellitus is a disease in which the body is unable to produce or unable to properly use and store glucose (a form of sugar). Glucose backs up in the bloodstream causing one's blood glucose or "sugar" to rise too high. Poorly managed diabetes can lead to a host of long-term complications among these are heart attacks, strokes, blindness, kidney failure, blood vessel disease. Pima Indians Diabetes Database [18] includes the following attributes (1-8 attributes as input and last attribute as target variable):

1. Number of times pregnant;
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test;
3. Diastolic blood pressure (mm Hg);
4. Triceps skin fold thickness (mm);
5. 2-Hour serum insulin (mu U/ml);
6. Body mass index (weight in kg/(height in m)^2);
7. Diabetes pedigree function;
8. Age (years);
9. Class variable (0 or 1).

A total of 768 cases are available in PIDD. 5 patients had a glucose of 0, 11 patients had a body mass index of 0, 28 others had a diastolic blood pressure of 0, 192 others had skin fold thickness readings of 0, and 140 others had serum insulin levels of 0. After deleting these cases there were 392 cases with no missing values (130 tested positive cases and 262 tested negative) [19].

6. RESULTS

As a part of feature selection step we used two filter approaches (i) C4.5 decision tree with gain ratio as measure to select relevant attributes (ii) Genetic search with Correlation based feature selection as subset evaluating mechanism from Pima Indians Diabetes Database. C4.5 tree with 9 attributes (eight input and one output class attribute) gave an accuracy of 73.83%. The pruned tree gave number of times pregnant, plasma, pressure, mass, pedigree and age as significant attributes. On further removal of first attribute namely: number of times pregnant, it gave an improved accuracy of 75.91 %. Finally the following five relevant attributes namely Plasma Glucose concentration, diastolic blood pressure, Body mass index, diabetes pedigree function and age were used as inputs for BPN. The pruned J4.8 tree is shown in figure 1. The default K folds cross validation method with K = 10 was used for decision tree.

For GA, population size is 20,number of generation is 20, crossover rate is 0.6 and mutation rate is 0.033. GA with CFS resulted in features subset of 4 attributes namely plasma, insulin, mass and age which were given as input to neural network. For GA two methods, K fold cross validation with K = 10 and full data as training data was used. Both the methods with CFS resulted in same four-feature subset namely plasma, insulin, mass and age.

Various topologies were tried with original 8 inputs, and with features subsets selected by DT (5 inputs) and GA-CFS (4 inputs) with back propagation is shown in figure 2. Figure 3 shows the comparative graph showing the classification accuracy (%) obtained for best topologies by integrated model GA-CFS & BPN for 4-8-1 topology, integrated model DT & BPN for 5-15-1 topology, and BPN alone for 8-24-1 topology. Figure 3 clearly shows that classification accuracy for the unprocessed data (BPN alone) is the least with 72.88%. Classification accuracy of BPN with feature subset identified by decision tree was found to be 78.21%. Further the classification accuracy was highest

with 79.5% with feature subset identified using GA-CFS with classifier BPN.

Further the classification accuracy, root mean squared error, relative absolute error and root relative squared error achieved using RBF with original 8 inputs, and with features subsets selected by DT (5 inputs) and GA-CFS (4 inputs) is shown in table 1.Table 1 and figure 3 clearly depicts the improved classification accuracy of both BPN and RBF with the inputs identified by GA-CFS when compared to unprocessed input data & input attributes as identified by DT.

7. CONCLUSION

The downside of information gain in DT is, it does not account for interactions between features, which is overcome by 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. Subsets of features identified by CFS are highly correlated with the class while having low intercorrelation. Experimental results illustrates CFS identified feature subset have improved the BPN and RBF classification accuracy when compared to relevant input as identified by decision tree.

```
Relation:     pima_diabetes.
Instances:    768
Input Attributes: 5: plas, pres, mass, pedi, age

plas <= 127
|   mass <= 26.4: tested_negative (132.0/3.0)
|   mass > 26.4
|   |   age <= 28: tested_negative (180.0/22.0)
|   |   age > 28
|   |   |   plas <= 99: tested_negative (55.0/10.0)
|   |   |   plas > 99
|   |   |   |   pedi <= 0.561: tested_negative (84.0/34.0)
|   |   |   |   pedi > 0.561: tested_positive (34.0/9.0)
plas > 127
|   mass <= 29.9
|   |   plas <= 145: tested_negative (41.0/6.0)
|   |   plas > 145
|   |   |   age <= 25: tested_negative (4.0)
|   |   |   age > 25
|   |   |   |   age <= 61
|   |   |   |   |   mass <= 27.1: tested_positive (12.0/1.0)
|   |   |   |   |   mass > 27.1
|   |   |   |   |   |   pres <= 82
|   |   |   |   |   |   |   pedi <= 0.396: tested_positive (8.0/1.0)
|   |   |   |   |   |   |   pedi > 0.396: tested_negative (3.0)
|   |   |   |   |   |   pres > 82: tested_negative (4.0)
|   |   |   |   age > 61: tested_negative (4.0)
|   mass > 29.9
|   |   plas <= 157
|   |   |   pres <= 61: tested_positive (15.0/1.0)
|   |   |   pres > 61
|   |   |   |   age <= 30: tested_negative (40.0/13.0)
|   |   |   |   age > 30: tested_positive (60.0/17.0)
|   |   plas > 157: tested_positive (92.0/12.0)
```

Fig. 1: Weka.Classifiers.Trees.J48

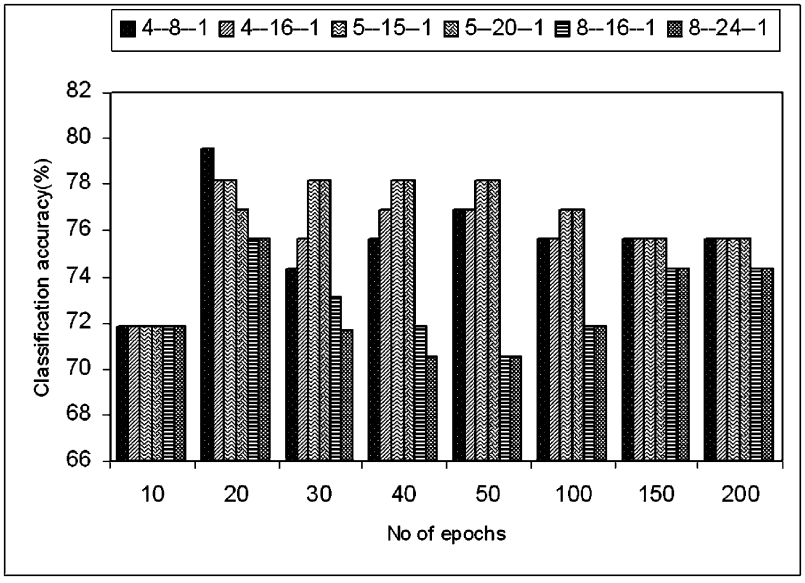

Fig. 2: Classification Accuracy (%) for different Topologies of BPN with 4,5 and 8 Inputs.

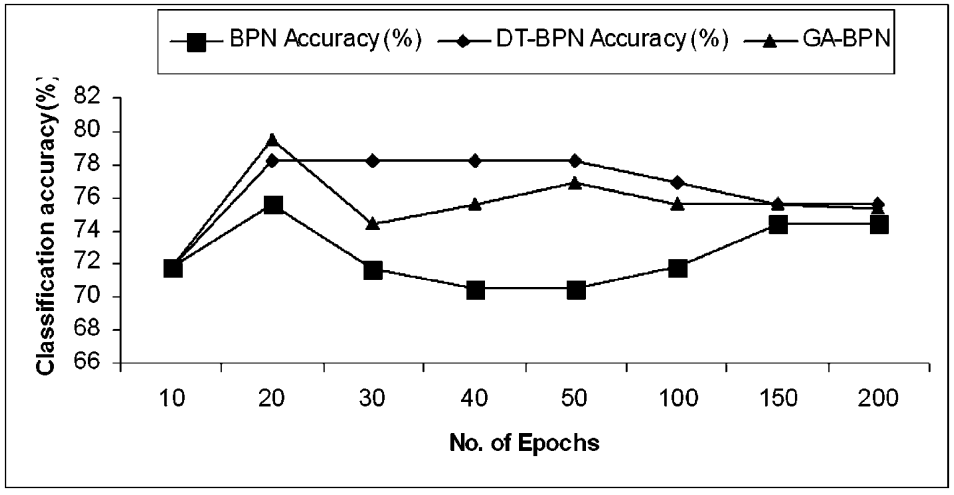

Fig. 3: Comparative Graph Showing the Classification Accuracy (%) Obtained by
(i) Integrated Model (GA and BPN)(4-8-1 Topology)
(ii) Integrated Model (DT and BPN) (5-15-1 Topology),
(iii) BPN (8-24-1 Topology)

Table 1
Classification Accuracy for Diabetic Dataset using RBF with,
(i) All the 8 Inputs Attributes, (ii) 5 Input Attributes Identified by DT,
(iii) 4 Input Attributes Identified by GA-CFS

Algorithm	No. of inputs	Classification Accuracy	Root mean squared error	Relative absolute error	Root relative squared error
RBF	8	81.20 %	0.37	66.14 %	82.22 %
DT & RBF	5	86.46 %	0.34	62.49 %	76.72 %
GA-CFS & RBF	4	88.00 %	0.33	59.57 %	73.87 %

REFERENCES:

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

[2] 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 ismir2008.ismir.net/papers/ISMIR2008_256.pdf(2008).

[3] Y.Saeys, I.Inza, and P. LarrANNaga, "A Review of Feature Selection Techniques in Bioinformatics", Bioinformatics, 23(19), pp.2507-2517, (2007).

[4] Mark A. Hall, Correlation-based Feature Selection for Machine Learning, Dept of Computer Science, University of Waikato. http://www.cs.waikato.ac.nz/~mhall/thesis.pdf.

[5] J.R. Quinlan, Induction of Decision Trees, Machine Learning 1: pp.81-106, Kluwer Academic Publishers, Boston, (1986).

[6] J.R. Quinlan, San Mateo, C4.5 Programs for Machine Learning: Morgan KaufmANN, (1993).

[7] J.R. Quinlan, Bagging, Boosting and C4.5, In Proc. 13th National Conf. Back Propagation Intelligence (AAAI'96), pp. 725-730. Portland, (Aug, 1996).

[8] http://www.cs.waikato.ac.nz/~ml/weka/.

[9] M.A.Jayaram , Asha Gowda Karegowda, Integrating Decision Tree and ANN for Categorization of Diabetics Data, International Conference on Computer Aided Engineering, IIT Madras, Chennai, India. (December 13-15, 2007).

[10] Asha Gowda Karegowda and M.A.Jayaram, "Cascading GA & CFS for Feature Subset Selection in Medial Data Mining", IEEE International Advance Computing Conference, Patiyala, India, (2009).

[11] D. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison Wesley,(1989).

[12] Roy A. Artificial Neural Networks – a Science in Trouble, SIGKDD Explorations, 1:33-38(2000).

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

[14] Rinehart D.E., Hinton G.E., and Williams R. J. Learning Internal Representations by Error Propagation. In D.E. Rumelhart and J.L. McClelland, Editors, Parallel Distributed Processing. Cambridge, MA:MIIT Press, (1986).

[15] Lu, H., Setiono R. and Liu , H.. Effective Data Mining using Neural Networks., IEEE Trans. on Knowledge and Data Engineering, (1996).

[16] P. J. Joseph, Kapil Vaswani, Matthew J. Thazhuthaveetil, A Predictive Performance Model for Superscalar Processors Microarchitecture, MICRO-39. 39th Annual IEEE/ACM International Symposium, on Volume, Issue, Page(s):161 - 170](Dec. 2006).

[17] Satish Kumar, Neural Networks A Classroom Approach, Tata McGraw Hill, (2006).

[18] http://www1.ics.uci.edu/~mlearn/MLSummary.html

[19] Joseph L.Breault, Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?, http://www.galaxy.gmu.edu/interface/I01/I2001Proceedings/Jbreault.