Journal of Theoretical and Applied Information Technology
15th December 2012. Vol. 46 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ATTRIBUTE SELECTION USING ARTIFICIAL NEURAL NETWORKS – A CASE STUDY OF ISCHEMIC HEART DISEASE

[1] K.RAJESWARI, [2] Dr.V.VAITHIYANATHAN

[1] Research Scholar,School of Computing,SASTRA University,Thanjavur 613 401, India

[2] Associate Dean and CTS Chair Professor, School of Computing, SASTRA University, Thanjavur 613401 India

E-mail: [1]raji.pccoe@gmail.com, [2]vvn@it.sastra.edu

ABSTRACT

Attribute selection also called as feature selection is a preprocessing technique to select a set of features or subset of features from the available large collection of features. An artificial neural network is the simulation of a human brain which learns with experience. Efficiency of a model or a system in terms of cost, time and accuracy will greatly improve if proper features of a system are selected. This proposed method uses Artificial Neural Network for selecting the interesting or important features from the input layer of the network. A Multi Layer Perceptron Neural Network is used for selection of interesting features from a Ischemic heart data base with 712 patients. Initially the number of attributes was 17 and after feature selection the number of attributes was reduced to 12. All combination of features is attempted as Inputs of a Neural Network. When the input features become 12 the predicted accuracy during training is high as 87.36% using 10 – fold cross validation. Further removal of features lowers the accuracy and hence the interesting attributes selected for prediction is concluded to be as 12 for this Ischemic Heart Disease data set.

Keywords: *Data Mining,;Feature Selection;Multi Layer Perceptron, Neural Network;Ischemic Heart Disease(IHD)*

1. INTRODUCTION

IHD is a cardiovascular disease (CVD) which is alarmingly on rise throughout the world. Especially Indian population with IHD and death due to this is dangerously increasing.The largest cause of death according to [1] by 2020 will be CVD. It is predicted that nearly 2.6 million Indians are about to die due to coronary heart disease. This constitutes 54.1% of all CVD deaths. Mostly this is expected to occur in individuals with age 30 to 69 years. Hence research on IHD has become significant. This paper focuses on reducing the number of features examined for IHD identification thereby reducing the doctor's time, patient's time and cost. The research objectives are (1) to observe neural networks, especially back propagation neural networks (2) to study about IHD and (to try various combinations of parameters or features for IHD identification. The organization of the paper is as follows. Chapter 2 discusses briefly about back propagation algorithm. Chapter 3 discusses about IHD and the parameters used to identify IHD. Chapter 4 discusses about methods and materials used for the research. Chapter 5 discusses about the findings and Chapter 6 gives the conclusion and scope for further research.

2. BACKPROPAGATION ALGORITHM

To identify a person with IHD, the feed forward neural network trained with back propagation algorithm is used. And further the output is classified as high, medium and low risk levels. Back propagation is a neural network learning algorithm where the network will have a set of input nodes connected to the nodes in output layer through the nodes in the hidden layer. Here each connection has a weight value associated with it and is shown in Fig. 1[13].

Journal of Theoretical and Applied Information Technology
15th December 2012. Vol. 46 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

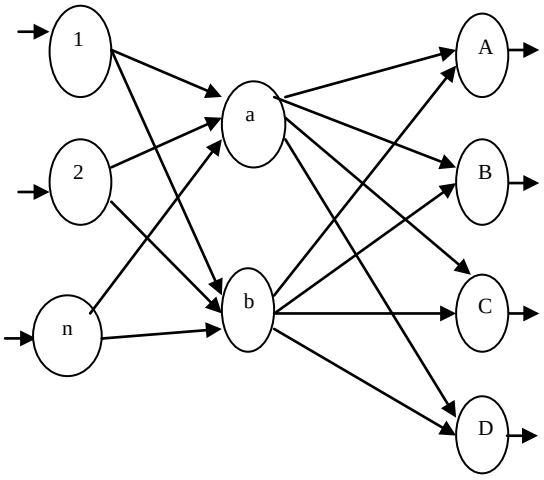

Fig 1.A Feed Forward Neural Network

1) There are two phases

1. Learning phase where the network learns by modification of weights.
2. Testing phase where an unknown input is tested for proper learning of neural network.

In Fig.1, the nodes 1, 2, 3,.., n represents the input nodes of input layer. These n nodes represent the features or parameters determining IHD. Nodes a,b represents nodes of hidden layer. Nodes A, B, C, D represents nodes of output layer.

The Back-propagation algorithm [13] is as follows.

Input

- The samples used for training
- The rate of learning - l
- A feed – forward multi layered fully connected network

Method

Network is initialized with random values of weights and biases.

Repeat till termination conditions are met

{

for each sample S in **the set of training samples**

{

// Propagate the inputs forward

for each hidden or output layer unit j

{

- Sum of products of weight and output of a particular node with the bias assumed is the input value for the next layer
- For each unit j, compute the output using exponential activation function.
- Back propagate the errors with the following steps
- for each unit j in the output layer
- $Error_j = Out_j (1- Out_j) (Tru_j – Out_j)$;
- for each unit j in the hidden layers , from the last to the first hidden layers
  - $Error_j = Out_j (1- Out_j) \Sigma_k Error_k w_{jk}$;

//Compute the error wrt the next higher layer , k

- for each weight wt_{ij} in network
  - Increment the weight, bias values
  - $wt_{ij} = wt_{ij} + \Delta wt_{ij}$;
- for each bias θ_j in network
  - $\Delta \theta b_j = (lr)Err_j$;
  - $\theta b_j = \theta_j + \Delta \theta b_j$;

}}

Fig 2. Backpropagation Algorithm

3. ISCHEMIC HEART DISEASE

A huge body of data pertaining to the IHD patients is existing in many publications and University of California, Irvine (UCI) data base. But, only few studies like in Chandigarh, screening patients over the age 30 and in Haryana were the occurrence is 65.4 pre1000 males and 47.8 per females in city population [2].In epidemiology, the occurrence of a disease in a population is defined as the full number of cases of the illness in the population at a given time, or the total number of sick cases in the people, divided by the total quantity of persons [15]. Our study was carried out for a database collected from Madras Medical College, Chennai in India for the age group 30 and above.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Indians are more prone to heart disease that lead to worse outcomes like IHD a condition characterized by reduced blood supply to the heart[14]. The pattern of IHD in India has been reported to be as follows:

- In India, IHD appears in earlier ages compared with developed countries. The tough period is attained between 51-60 years of age
- Males are affected more than females.
- Hypertension[9] and diabetes account for about 40 percent of all cases.
- Heavy smoking is responsible aetologically in a good number of cases [3][4][5][6].
- A family history of IHD. Children of parents with heart disease are more likely to develop it themselves.
- Sedentary life style, a casual life with no physical activity [10] .
- Type A individuals who are time conscious, tightly bound to job and are restless [11].
- Higher alcohol intake, defined as 75g [7] Research has revealed an association between moderate alcohol consumption and lower risk for CHD [16][17].

4. MATERIALS AND METHODS

A. Data Preparation

The following data were collected and analyzed for Indian heart risk score prediction based on extensive study and expert opinions. from doctors with respect to Indian body conditions, life style and eating habits. After discussion with cardiologists a three-stage questionnaire was prepared. Diagnosis was done through data collection for each individual patient as given in tables I, II and III.

Stage I includes physical examination parameters. Stage II includes Co Morbid features collection and Stage III includes attributes on personal habits and hereditary.

Table I. Stage One Diagnosis

Stage 1						
Age	Gender [F – Female M- Male]	Menopause [0 for pre menopause]	Ht in cm	Wt in kg	Body Mass Index in kg/m2	Waist Measure in cm
30	F	0	165	65	19.69	75

If sex is female, pre or post menopause details are recorded. Waist circumference is noted as it is proved to be an important risk determinant factor for CHD[12].

Table II. Stage Two Diagnosis

Stage 2				
Co Morbid Factors				
SBP	DBP	Diabetes	Cholesterol	Thyroid
120	80	1	1	0
140	90	0	1	0

SBP is Systolic Blood Pressure and DBP is Diastolic Blood Pressure.

Table III. Stage Three Diagnosis

Stage 3				
Personal Habits	Family History	Genetic Factors	Type A Personality	Sleeping Disturbance
1	0	0	1	1
0	1	1	1	1

Based on the data collected, the immediate risk analysis was classified as ‘No Risk’, ‘Low Risk’, ‘Medium Risk’ or ‘High Risk’. Table V data is collected from experienced doctors in cardiology based on the importance of every feature collected in tables III, IV and V. Three expert opinions were collected. Two identical opinions were taken into consideration for deriving conclusion. Varied opinion data of a patient was removed from the dataset.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table IV. Output

Absolute Risk for CAD			
No Risk	Low Risk	Medium Risk	High Risk

B. Attribute Selection - Pruning

In the IHD 17 features from 712 patients were collected. After collection of these features Expert doctor's opinion on Risk of Ischemic heart disease was obtained after ECG was taken. Multilayer perception network was used. In the dataset 10 fold cross validation is used in the same way as in Table 1, every single feature was removed subsequently and the accuracy during training and testing period was noted. During the first iteration it was noted that the accuracy was worse when the following attributes were removed

1. Age
2. Gender
3. Menopause
4. Diabetes
5. Cholesterol
6. Smoker/Alcohol.

Hence according to our data set, we can conclude in iteration 1- the above attributes contribute for identifying IHD. Table IV below shows the accuracy obtained when the features F1 to F17 were removed one by one. The description of Features are given in the first column

Table IV. Sample Pruning

	~F1	~F2	~F3	~f4	~F5	~F6	~F7	~F8	.	~F 17
Age	X									
Gender		X								
Menop			X							
Ht				X						
Wt					X					
BMI						X				
WC							X			
SBP								X		
DBP										
Diab										
Chol										
Thyroid										
Personal habits										
Family History										
Genetic										
Type A										
Sleep										X

During the second iteration, every time two attributes are removed and the accuracy is checked using a classifier. For example{{F1,F2} are removed, followed by {F1,F3}, {F1,F4},{F1,F5},……..{F2,F3} ,{F2,F4}, {F2,F5}……{F3,F4},{F3,F5},{F3,F6} ,……{F4,F5},{F4,F6},{F4,F7},…… ……{F5,F6}, {F5,F7},{F5,F8},…….....{F16,F17}}i s checked for accuracy. We found that accuracy was worse when the following pairs were removed.

{F1, F2]

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Thus second iteration identified Age and Gender as important features.

During the third iteration, three features at a time were removed. For example: F1, F2, F3 together were removed from the list of independent variables or factors or as inputs. Accuracy is noted. The process continues with 4 feature removal, 5 feature removal, and when 6 features are removed accuracy is worse with any combination. At this stage we stop removing further combination of features.

Genetic algorithm(GA) is used to select the attributes. It is an optimization algorithm, and resulted in 14 attributes selected as important attributes. GA has left gender. When compared to Genetic search, this method of using a classifier, to retrieves important features is significant as it classifies the output according to the selected attributes.

With filtered subset evaluator, only 3 attribute selected namely, cholesterol, diabetes mellitus and type of personality with sleeping disturbance, decides the output cadre obtaining 76% of accuracy using neural network classifier.

5. RESULTS AND DISCUSSION

The Feed forward Neural Network has 17 data inputs and 4 outputs. A total of 17 variables were taken for study after discussion with experts. The Software SPSS tool is used for training and testing the data set with the different combination of features. It is concluded that the 12 features that are important to determine the risk level of IHD are as follows: {Age, gender, menopause, body mass index, waist circumference, systolic blood pressure, diastolic blood pressure, diabetes, cholesterol, hereditary, personal habits and typer A personality}. The accuracy of classification with these 12 attributes is 87.36%. Kappa statistic is 82.77% which gives an assurance about the quality of inputs. The summary of confusion matrix is as below in table V.

Table V. Accuracy

a	b	c	d	Classified as
213	14	1	0	a=0
25	156	9	0	b=1
3	6	205	10	c=2
0	0	22	48	d=3

6. CONCLUSION

This paper focuses on attribute reduction using Artificial Neural Networks. Comparison with GA and subset evaluator is done. The 17 features are reduced to 12 using back propagation algorithm and the accuracy is verified usindg parameters like sensitivity, specificity, Area under curve, kappa statistic, precision and recall measures. This work can be widespread to different zones to find its legitimacy.

REFRENCES:

[1]www.whoindia.org/LinkFiles/NMH_Resources_National_CVD_database-Final_Report.pdf

[2] Dewan, B.D. et al (1974), Indian heart j.,26:68.

[3] I Sinha,B.C (1970). Jr. Ind. Med. Assoc,55 : 171.

[4] Slone, D. et al(1978). N. Eng. J. Med. 298:1273.

[5] R Sharper A.G. et. Al (1981). Brit. Med. J. 283:179.

[6] Bain, C. et. Al(1978). Lancet, 1:1087.

[7] WHO (1985). Primary Prevention of CHD EURO Rep and Studio 98. Copenhagen.

[8] Keys, A. 91980). Seven Countries : a multivariate analysis of death and CHD, Harvard University Press, Cambridge, M.A.

[9] WHO (1985). Tech. Rep. Ser., 727.

[10] Miller, N.E. et.al (1979) Lancet, 1:111.

[11] Jenkins, C.D.et a(1974). N. Eng. J. Med., 290:1271.

[12] Palla Venkataramana, Palakuru C Reddy (2002).Association of overalland abdominal obesity with coronary heart disease risk factors : comparison between urban and rural Indian men', Asia pacific Journal Clin Nutr 11(1):66-71

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[13] Jiawei Han and Micheline Kamber, "DataMining Concepts and Techniques", Elsevier Second Edition

[14]http://timesofindia.indiatimes.com/city/delhi/TOI-campaign-against-heart-disease-a-success/articleshow/7193581.cms

[15] http://en.wikipedia.org/wiki/Prevalnece

[16]http://pubs.niaaa.nih.gov/publications/aa45.htm

[17]Hennekens, C.H. Alcohol and risk of coronary events. In: Zakhari, S.,and Wassef, M., eds. Alcohol and the Cardiovascular System. NIAAA Research Monograph No. 31. NIH Pub. No. 96-4133. Washington, DC: U.S. Govt. Print. Off., 1996. pp. 15-24.