Global Journal of Computer Science and Technology
Neural & Artificial Intelligence
Volume 13 Issue 3 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection

By M. Akhil Jabbar, B.L Deekshatulu & Priti Chandra
AEC, Bhongir, India

Abstract - Now a day's artificial neural network (ANN) has been widely used as a tool for solving many decision modeling problems. A multilayer perception is a feed forward ANN model that is used extensively for the solution of a no. of different problems. An ANN is the simulation of the human brain. It is a supervised learning technique used for non linear classification Coronary heart disease is major epidemic in India and Andhra Pradesh is in risk of Coronary Heart Disease. Clinical diagnosis is done mostly by doctor's expertise and patients were asked to take no. of diagnosis tests. But all the tests will not contribute towards effective diagnosis of disease. Feature subset selection is a preprocessing step used to reduce dimensionality, remove irrelevant data. In this paper we introduce a classification approach which uses ANN and feature subset selection for the classification of heart disease. PCA is used for preprocessing and to reduce no. Of attributes which indirectly reduces the no. of diagnosis tests which are needed to be taken by a patient. We applied our approach on Andhra Pradesh heart disease data base. Our experimental results show that accuracy improved over traditional classification techniques. This system is feasible and faster and more accurate for diagnosis of heart disease.

Keywords : *andhra pradesh, artificial neural network, chi-square, data mining, feature subset selection, genetic search, heart disease, principal component analysis.*

GJCST-D Classification : *I.2.6*

Strictly as per the compliance and regulations of:

Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection

M. Akhil Jabbar[α], B.L Deekshatulu [σ] & Priti Chandra [ρ]

Abstract - Now a day's artificial neural network (ANN) has been widely used as a tool for solving many decision modeling problems. A multilayer perception is a feed forward ANN model that is used extensively for the solution of a no. of different problems. An ANN is the simulation of the human brain. It is a supervised learning technique used for non linear classification Coronary heart disease is major epidemic in India and Andhra Pradesh is in risk of Coronary Heart Disease. Clinical diagnosis is done mostly by doctor's expertise and patients were asked to take no. of diagnosis tests. But all the tests will not contribute towards effective diagnosis of disease. Feature subset selection is a preprocessing step used to reduce dimensionality, remove irrelevant data. In this paper we introduce a classification approach which uses ANN and feature subset selection for the classification of heart disease. PCA is used for preprocessing and to reduce no. Of attributes which indirectly reduces the no. of diagnosis tests which are needed to be taken by a patient. We applied our approach on Andhra Pradesh heart disease data base. Our experimental results show that accuracy improved over traditional classification techniques. This system is feasible and faster and more accurate for diagnosis of heart disease.

Keywords : andhra pradesh, artificial neural network, chi-square, data mining, feature subset selection, genetic search, heart disease, principal component analysis.

I. Introduction

Data mining is the process of extracting knowledge able information from huge amounts of data. It is an integration of multiple disciplines such as statistics, machine learning, neural networks and pattern recognition. Data mining extracts biomedical and health care knowledge for clinical decision making and generate scientific hypothesis from large medical data.

Association rule mining and classification are two major techniques of data mining. Association rule mining is an unsupervised learning method for discovering interesting patterns and their association in large data bases. Whereas classification is a supervised learning method used to find class label for unknown sample.

Classification is the task of assigning objects tone of special predefined categories. It is pervasive problem that encompasses many applications. Classification is designed as the task of learning a target function F that maps each attribute set A to one of the predefined class labels C [1]. The target function is also known as classification model. A classification model is useful for mainly two purposes. 1) descriptive modeling 2) Predictive modeling.

An artificial neural network (ANN) is the simulation of human brain and is being applied to an increasingly number of real world problems. Neural networks as a tool we can mine knowledgeable data from data ware house. ANN are trained to recognize, store and retrieve patterns to solve combinatorial optimization problems. Pattern recognition and function Estimation abilities make ANN prevalent utility in data mining. Their main advantage is that they can solve problems that are too complex for conventional technologies. Neural networks are well suited to problem like pattern recognition andforecasting.ANN are used to extract useful patterns from the data and infer rules from them. These are useful in providing information on associations, classifications and clustering.

Heart disease is a form of cardio vascular disease that affects men and women. Coronary heart disease is an epidemic in India and one of the disease burden and deaths. It is estimated that by the year 2012, India will bear 60% of the world's heart disease burden. Sixty years is the average age of heart patients in India against 63-68 in developed countries. In India Andhra Pradesh state is in risk of more deaths due to CHD. Hence there is a need to combat the heart disease. Diagnosis of the heart disease in the early phase solves many lives.

Feature subset selection is a processing step in Machine learning and is used to reduce dimensionality and removes irrelevant data. It increases accuracy thus improves resultcomprehensibility.PCA is the oldest multivariate statistical technique used by most of the scientific disciplines. The goal of PCA is to extract the important information from data bases and express this information as principle components. Chi square test evaluates the worth of an attribute by computing the values of chi squared statistics w.r.t class. The larger the value of chi square, the more likely the variable is related to class.

In this paper we introduce a classification approach which combines multi layer perception with back propagation learning algorithm and feature

Author α : Associate Professor Aurora's Engineering College Bhongir, A.P, India. E-mail : jabbar.researchscholar@gmail.com
Author σ : Distinguished Fellow IDRBT, RBI (Govt of India) Hyderabad.
Author ρ : Senior Scientist ASL, DRDO Hyderabad.

selection for classification of heart disease with reduced no of attributes. Our approach him proves classification and determines the attributes which contribute more towards the predication of heart disease which indirectly Reduces no. of diagnosis test which are needed tube taken by a patient.

In section 2 we review basic concepts of neural networks, PCA, chi square and heart disease section 3 deals with related work. Section 4explains our proposed approach .Section 5 deals with results and discussion. We conclude our final remarks in section 6.

II. Basic Concepts

In this section we will discuss basic concepts of Neural networks, PCA, chi square and heart Disease.

a) Artificial Neural Networks

An ANN also called as neural network is a mathematical model based on biological neural networks. Artificial neural network is based on observation of a human brain [2].Human brain is very complicated web of neurons. Analogically artificial neural network is an interconnected set of three simple units namely input, hidden and output unit. The attributes that are passed as input to the next form a first layer. In medical diagnosis patients risk factors are treated as input to the artificial neural network.

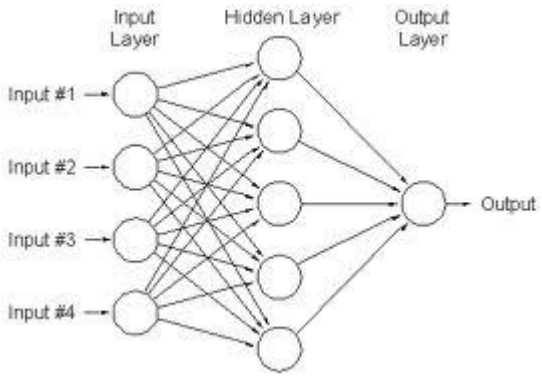

Figure 1 : Example ANN

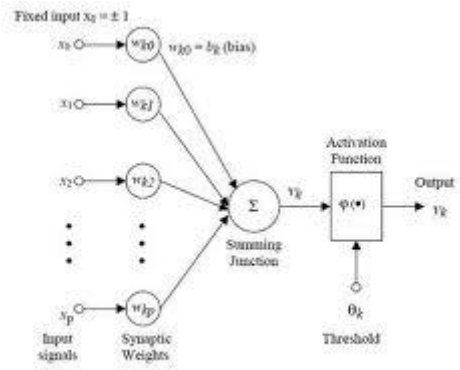

Figure 2 : Modeling a Boolean Function using Preceptor

Popular neural network algorithms are Hopfield, Multilayer perception, counter propagation networks, radial basis function and self organizing maps etc. The feed forward neural network was the first and simplest type of artificial neural network consists of 3 units input layer, hidden layer and output layer. There are no cycles or loops in this network. A neural network has to be configured to produce the desired set of outputs. Basically there are three learning situations for neural network. 1) Supervised Learning, 2) Unsupervised Learning, 3) Reinforcement learning the perception is the basic unit of a artificial neural network used for classification where patterns are linearly separable. The basic model of neuron used in perception is the Mc culluchpitts model. The perception takes an input value Vector and outputs 1 if the result is greater than predefined thresholds or -1 otherwise. The proof convergence of the algorithm is known apperception convergence theorem. Figure 1 shows ANN and figure 2 shows modeling a Boolean function using single layer perception [1]. Output node is used to represent the model output the nodes in neural network architecture are commonly known as neurons. Each input node is connected to output node via a weighted link. This is used to emulate the strength of synaptic connection between neurons. Simple perception learning algorithm is shown below.

Step 1 : Ret D= {{Xi, Yi}/i=1, 2, 3----n}} be the set of training example.
Step 2 : Initialize the weight vector with random value, W (o).
Step 3 : Repeat.
Step 4 : For each training sample (Xi, Yi) ∈ D.
Steps 5 : Compute the predicted output Yi^ (k).
Step 6 : For each weight we does.
Step 7 : Update the weight we (k+1) = Wj(k) + ª (y i - yi^ (k))xij.
Step 8 : End for.
Step 9 : End for.
Step 10 : Until stopping criteria is met.

The main function of artificial neural network is prediction. The effectiveness of artificial neural network was proven in medicine [3]. The note worthy achievement of neural network was in the application of coronary heart disease [4]. There is numerous advantages of ANN some of these include

1) High Accuracy.
2) Independent from prior assumptions about the distribution of the data.
3) Noise tolerance.
4) Ease of Maintatenance.
5) ANN can be implemented in parallel hardware.

The following are the examples of where ANN are used

1) Accounting.
2) Fraud Detection.
3) Telecommunication.

4) Medicine.
5) Marketing.
6) Insurance.
7) Human Resources.

Performance of ANN can be improved by designing ANN with evolutionary algorithms and developing neuron fuzzy systems.

b) Feature Subset Selection

Feature subset selection is a preprocessing step commonly used in machine learning. It is effective in reducing dimensionality and removes irrelevant data thus increases learning accuracy. It refers to the problem of identifying those features that are useful in predicting class. Features can be discrete, continuous or nominal. Generally features are of three types. 1) Relevant, 2) Irrelevant, 3) Redundant. Feature selection methods wrapper and embedded models. Filter model rely on analyzing the general characteristics of data and evaluating features and will not involve any learning algorithm, where as wrapper model uses après determined learning algorithm and use learning algorithms performance on the provided features in the evaluation step to identify relevant feature. Embedded models incorporate feature selection as a part of the model training process.

Data from medical sources are highly voluminous nature. Many important factors affect the success of data mining on medical data. If the data is irrelevant, redundant then knowledge discovery during training phase is more difficult. Figure 3shows flow of FSS.

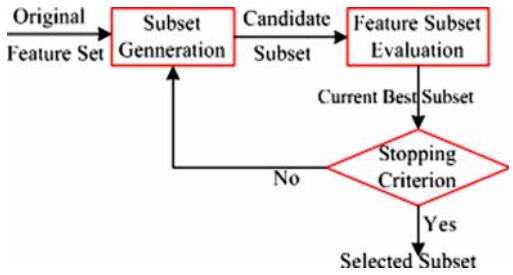

Figure 3 : Feature subset selection

Medical diagnosis is an important yet complicated task that needs to be executed accurately and efficiently [5].feature subset selection if applied on medical data mining will leads to accurate results. The detection of disease from several factors is a multi layered problem and sometimes leads to false assumptions frequently associated with erratic effects. Therefore it appears reasonable if we apply feature subset selection in medical data mining towards assisting the diagnosis process. In this paper we apply feature subset selection. Using neural network for heart disease prediction For Andhra Pradesh population.

i. *Principle Component Analysis*

Principle component analysis (PCA) is a statically technique used in many applications like face recognition, pattern recognition, image compression and data mining. PCA is used to reduce dimensionality of the data consisting of a large no. of attributes. PCA can be generalized as multiple factor analysis and as correspondence analysis to handle heterogeneous sets of variables and quantitative variables respectively. Mathematically PCA depends on SVD of rectangular matrices and Eigen decomposition of positive semi definite matrices. The goals principle component analyses are

1) To extract the most important information·
2) from the data base Compress the size of the data and keeping only important information.
3) Simplify the data set description and to analyze the variables and structure of the observations.
4) Simplification.
5) Modeling.
6) Outlier Detection.

Procedure for principle component analysis is shown below

a. *Procedure*

Step 1 : Obtain the input matrix.
Step 2 : Subtract the mean from the data set in all dimensions.
Step 3 : Calculate covariance matrix of this mean subtracted data set.
Step 4 : Calculate the Eigen values and Eigen vector from covariance matrix.
Step 5 : Form a feature vector.
Steps 6 : Derive the new data set.

ii. *Chi-Squared Test*

One of the first steps in data mining and knowledge discovery is the process of eliminating redundant and irrelevant variables. There are various reasons for taking this step. The obvious reason is that going from a few hundred variables to few variables will make the result interpretation in an easy way. The second reason is due to curse of dimensionality. If the dimensionality is large. It is necessary to have a large training set. This is also known as peaking phenomenon. If the dimensionality increases up to a certain point accuracy increases, but after there is a reduction in classification accuracy [6].

Simplest way of determining relevant variables is to use chi square technique (χ 2). Chi square technique is used if all the variables are continuous. Assume that a target variable is selected; every parameter is checked to see if the use chi square technique detects the existence of a relationship between the parameter and the target.

Karl Pearson proved that the statistic

$$X_r^2 = \frac{\Sigma (Oi - Ei)^2}{E_i}$$

Where O observed frequency and E expected frequency.

If the data is given in a series of n numbers then degrees of freedom = n-1.

In case of binomial distribution degrees of freedom = n-1.

In case of poison distribution degrees of freedom = n-2.

In case of normal distribution degrees of freedom= n-3 [7].

The following example illustrates chi square hypothesis. The total no of automobiles accidents per week in a certain community are as follows12, 8, 20,2,14,10,15,6,9 and 4.We have to verify to see whether accident conditions were same during 10week period using chi square test.

Expected frequency of accidents each week=100/10=10.

Null hypothesis H: The accident conditions were the same during the 10 week period.

Table 1 : Chi square computation

Observed frequency	Expected frequency	O-E	$(O-E)^2/E$
12	10	2	0.4
8	10	-2	0.4
20	10	10	10.0
2	10	-8	6.4
14	10	4	1.6
10	10	0	1.0
15	10	5	2.5
6	10	-4	1.6
9	10	-1	0.1
4	10	-6	3.6
Total	100		26.6

Chi square=26.6 and degree of freedom=10-1=9
Tabulated chi square=16.9
Calculated chi square> Tabulated chi square

So the null hypothesis is rejected i.e. accident condition were not same during the 10 week period.

Following are the requirements for chi square test

1) One or more categories of quantitative data
2) Independent observations
3) Adequate sample size
4) Simple random sample
5) Data in frequency form
6) And all observations must be used

c) Heart Disease

Coronary heart disease occurs when the arteries of the heart that normally provide blood and oxygen to the heart are narrowed or even completely blocked. Cardiovascular diseases account for high mortality and morbidity all around the world. In India mortality due to Chewier 1.6 million in the year 2000.by the year 2015,61 million cases will be due to CHD[8].Studies to determine the precise cause of death in rural areas of Andhra Pradesh have revealed that CVD Cause about 30% deaths in rural areas [9].

i. *Risk factors for heart disease*

Some of the risk factors for heart disease are

1) *Smoking:* Smokers risk a heart attack twice as much as non smokers.
2) *Cholesterol:* A diet low in cholesterol and saturated Tran's fat will help lower cholesterol levels and reduce the risk of heart disease.
3) *Blood pressure:* High BP leads to heart Attack
4) *Diabetes:* Diabetes if not controlled can lead to significant heart damage including heart attack and death
5) *Sedentary life style:* Simple leisure time activities like gardening and walking can lower our risk of heart disease.
6) *Eating Habits:* Heart healthy diet, low in salt, saturated fat, Trans fat, cholesterol and refined sugars will lower our chances of getting heart disease.
7) *Stress:* Poorly controlled stress an danger can lead to heart attacks and strokes.

This epidemic may be halted through the promotion of healthier life styles, physical activity; traditional food consumption would help to mitigate this burden.

d) Genetic Search

Generally for feature subset selection, search spaces will be large. There are 2 Power 204 possible features combinations for cloud classification problem. Search strategies such as genetic search are used to find feature subsets with high accuracy. Genetic algorithms are used for optimization problems and are well known for robust search techniques.GA searches globally and obtain global competitive solutions. Genetic Algorithms are biologically motivated optimization methods which evolve a population of individuals where each individual who is more fit have a higher probability of surviving into subsequent generation. GA uses a set of evolutionary metaphors named selection of individuals, mutation, and crossover. Many of the algorithms uses classifier as a fitness function. Figure 4shows working principle of genetic algorithm.

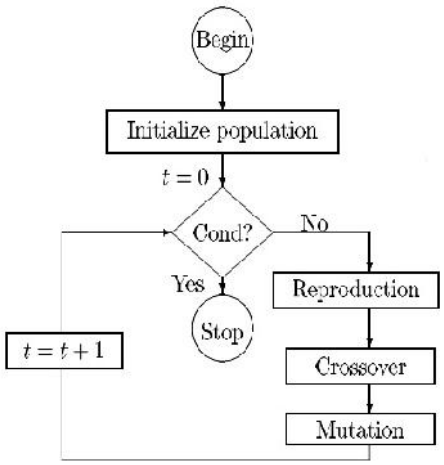

Figure 4 : Working principle of genetic algorithm

III. Related Work

Few research works has been carried out for diagnosis of various diseases using data mining. Our approach is to apply feature subset selection and artificial neural networks for prediction of heart disease.M.A.Jabbaret.al proposed a new algorithm combining associative classification and feature subset selection for heart disease prediction [5]. They applied symmetrical uncertainty of attributes and genetic algorithm to remove redundant attributes. Enhanced prediction of heart disease using genetic algorithm and feature subset selection was proposed by Anbarasiet.al[10]. Heart disease prediction using associative classification was proposed by M.A. Jabbar et.al[11].matrix based association rule mining for heart disease prediction was proposed by M.A. Jabbaret.al[12]. association rule mining and genetic algorithm based heart disease prediction was proposed in [13].cluster based association rule mining for disease prediction was proposed in [14]. sellappan palaniappan et al proposed intelligent heart disease prediction system using naïve bayes, decision tree and neural network in [15].graph based approach for heart disease prediction for Andhra Pradesh population was proposed by M.A.Jabbar et.al[16]. They combined maximum clique concept in graph with weighted association rule mining for disease prediction. Feature subset selection using FCBF in type II Diabetes patient's data was proposed by sarojinibala Krishnan et.al. [17]. Heart disease prediction using associative classification and genetic algorithm was proposed by M.A. Jabbaret.al [18].in their paper they used Z-Statistics a measure to filter out the rules generated by the system.

This paper proposes a new approach which combines Feature subset selection with artificial neural network to predict the heart disease.

IV. Proposed Method

In this paper we used PCA and chi square as a feature subset selection measure. This measure used to rank the attributes and to prune irrelevant, redundant attributes. After applying feature subset selection classification using ANN will be applied on the data sets.PCA is a mathematical procedure that transforms a no. of correlated attributes into a smaller no. of correlated variables called principle components. Assume If V is a set of N column vector of dimensions D, the mean of the of the data set, is

$$M\,v = E\{v\} \quad (1)$$

The covariance matrix is

$$Cv = E\{(V - M\,v)\{V - M\,v\}T\}$$

The components of Cox denoted by Cij represent the co variances between the random variable components Vi and Vj. the component Cii is the variance of Vi. The Eigen vectors ei and their corresponding values ---- are

$$C\,v\,ei = \square\, i\,e\,I \text{ where } i = 1,2,3\text{---}n$$

If X is a matrix of Eigen vectors of the covariance matrix, row vectors formed by transforming

$$V.r = A\,(V - mv)$$

The original data vector v can be reconstructed from r

$$V = XT\,r + mv \quad (5)$$

Consider the weather data set. Attributes are ranked by applying the principal component analysis.

Table 2 : Weather data set

No.	Outlook	Tempe rature	Humidity	Windy	Play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

a) Correlation Matrix

1 -0.47 -0.56 0.19 -0.04 -0.14 -0.15 0.04-0.47 1 -0.47 0.3 -0.23 -0.05 0 -0.09-0.56 -0.47 1 -0.47 0.26 0.19 0.15 0.04 0.19 0.3 -0.47 1 -0.55 -0.4 -0.32 0.23-0.04 -0.23 0.26 -0.55 1 -0.55 -0.29 -0.13-0.14 -0.05 0.19 -0.4 -0.55 1 0.63 -0.09-0.15 0 0.15 -0.32 -0.29 0.63 1 00.04 -0.09 0.04 0.23 -0.13 -0.09 0 1

b) Eigenvectors

V1 V2 V3 V4 V5
0.2935 0.1035 -0.6921 -0.2573 0.1079
outlook=sunny
0.2218 -0.3252 0.6278 -0.1582 0.3164
outlook=overcast
-0.5026 0.2031 0.1002 0.4064 -0.4061
outlook=rainy
0.5393 -0.2059 0.0342 0.2702 -0.356
temperature=hot
-0.1324 0.6291 0.1542 -0.1518 0.3959
temperature=mild
-0.3943 -0.4833 -0.2031 -0.1038 -0.0777
temperature=cool
-0.3694 -0.4111 -0.1475 -0.0599 0.4021
humidity
0.1081 -0.039 -0.1699 0.7957 0.5217 windy

Ranked attributes are arranged in descending order

1) outlook=rainy
2) temperature=hot
3) outlook=sunny
4) outlook=overcast
5) humidity
6) windy
7) temperature=mild
8) temperature=cool

Ranking of the attributes based on chi square is shown in table 3 and proposed method is shown in figure 5.

Table 3 : Ranking of the attributes base on chi square

Rank	Name of the attribute	Chi square value
1	outlook	3.547
2	humidity	2.8
3	windy	0.933
4	temperature	0.57

The rank of the attributes is done w.r.t the values of PCA, and χ 2 in a descending order. High values of PCA the more information the corresponding attribute has related to class. We trained the classifier to classify the heart disease data set as either healthy or sick. The accuracy of a classifier can be computed by

$$\text{Accuracy} = \frac{\text{No. of samples correctly classified in test Data set}}{\text{Total no. of samples in test data}}$$

c) Proposed algorithm

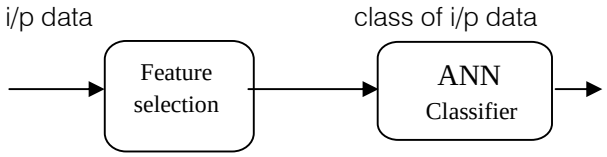

Figure 5 : Block diagram of proposed method

PROPOSED ALGORITHM

Step 1) Load the data set
Step 2) Apply feature subset on the data set.
Step 3) Rank the attributes in descending order based on their value. A high value of PCA and χ^2indicates attribute is more related to class. Least rank attributes will be pruned. Select the subsets with highest value.
Step 4) Apply multi layer perceptron on the remaining features of the data set that maximizes the classification accuracy.
Steps 5) find the accuracy of the classifier. Accuracy measures the ability of the classifier to correctly classify unlabeled data.

V. Results and Discussion

We have evaluated accuracy of our approach on various data sets taken from tuned it repository [19]. out of 8 data sets one is non medical dataset A brief description about various data sets is shown in table 4.Acuracy of various data sets is presented in table 5.comparision of our accuracy with various classification algorithms.

The heart disease data set is collected from various corporate hospitals and opinion from expert doctors. Attributes selected for A.P Heart disease is shown in is shown in Table 7.Applyingfeature subset selection helps increase computational efficiency while improving accuracy. Figure 9 shows parameters for performing multilayer preceptor. we set learning grates 0.3 and training time as 500.figure 8 and table 8 clearly depicts that classification accuracy is improved with feature subset selection with PCA. Classification accuracy has been improved for various data sets with χ 2 and ANN which is shown in table 9 and figure 10.comparison of classification accuracy for various data sets with three algorithms using ANN and chi square is shown in table 10.our approach improved 6.5%against classification accuracy against J48 and16.4% over naïve bayes algorithm and 7.3% than PART algorithm. Average accuracy for data sets using feature subset selection is 92.8%. Our proposed method was not much successful for hypothyroid and liver disorder this may be due to our method could not account for redundant attributes.

Table 4 : Description of various data sets

Data set	Instances	Attributes
Weather	14	5
Pima	768	9
hypothyroid	3770	30
breast cancer	286	10
liver disorder	345	7
primary tumor	339	18
heart stalog	270	14
lymph	148	19
heart disease A.P	23	12

Table 5 : Accuracy of various data sets using PCA

Data Set	Accuracy
Weather	100
Pima	98.82
hypothyroid	97.06
breast cancer	97.9
liver disorder	95.07
primary tumor	80
heart stalog	98.14
lymph	99.32
heart disease A.P	100

Table 6 : Comparison of classification accuracy for various data sets using PCA

Data set	J48	Naïve Bayes	PART	Our approach
Weather	100	92.8	85.7	100
Pima	85.1	76.3	81.2	98.82
hypothyroid	99.8	95.44	99.86	97.06
breast cancer	75.87	75.17	80.06	97.9
liver disorder	84.6	56.8	86.08	95.07
primary tumor	61.35	56.04	61.35	80
heart stalog	91.48	85.18	94.4	98.14
lymph	93.23	87.16	95.27	99.32
heart disease A.P	95	72.5	95	100
average	**87.3**	**77.4**	**86.5**	**96.2**

Table 7 : Attributes of heart disease A.P

Sl.no	Attribute	Data Type
1	Age	Numeric
2	Gender	Nominal
3	Diabetic	Nominal
4	BP Systolic	Numeric
5	BP Dialic	Numeric
6	Height	Numeric
7	Weight	Numeric
8	BMI	Numeric
9	Hypertension	Nominal
10	Rural	Nominal
11	Urban	Nominal
12	Disease	Nominal

Figure 6 : Parameters of PCA

Parameters of PCA

Center data: false
Max.attributes:5
Transform back to original data: true
Variance covered: 0.95

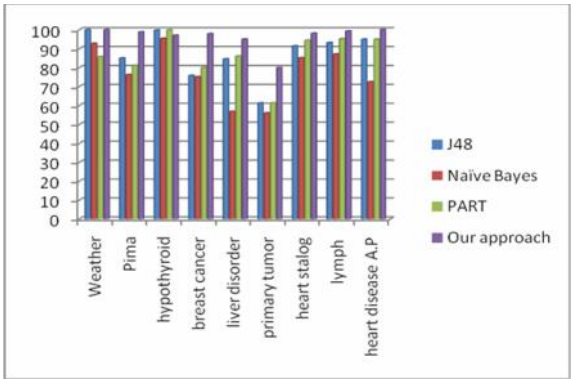

Figure 7 : Comparison of classification accuracy

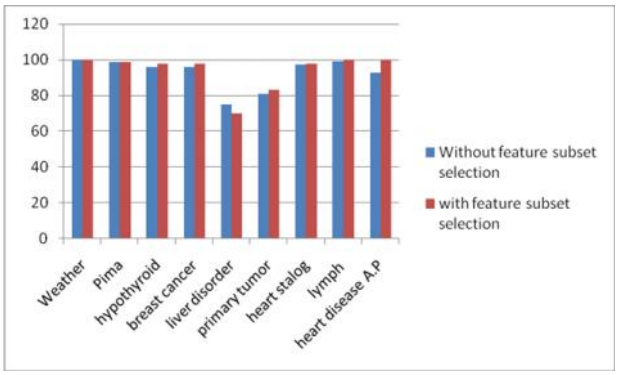

Figure 8 : Parameters of multilayer perceptron

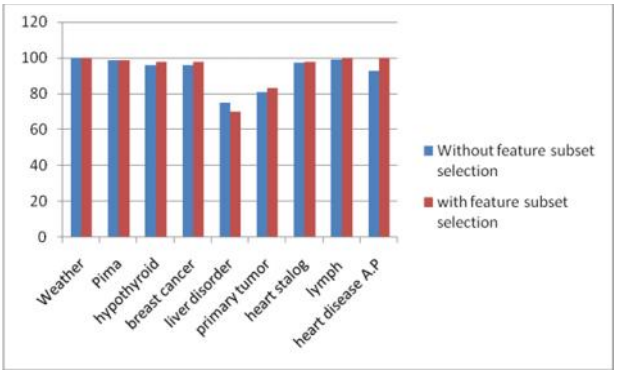

Figure 9 : Accuarcy with and with out feature subset selection using PCA

Table 8 : Accuracy comaprison with and with out feature subset selection using PCA

Data Set	Without feature subset selection	With feature subset selection
Weather	100	100
Pima	98.69	98.82
hypothyroid	95.94	97.08
breast cancer	96.5	97.9
liver disorder	95.07	85
primary tumor	80.83	80
heart stalog	97.4	98.14
lymph	99.3	99.3
heart disease A.P	100	100

Table 9 : Accuracy of various data sets with and without Feature subset selection using chi square

Data set	Without feature subset selection	with feature subset selection
Weather	100	100
Pima	98.69	98.82
hypothyroidyroid	95.84	97.64
breast cancer	95.84	97.64
liver disorder	74.78	70
primary tumor	80.82	83.18
heart stalog	97.4	97.7
lymph	99.3	100
heart disease A.P	92.5	100
AVERAGE	**92.7**	**93.8**

Parameters of multilayer perceptron during training

1) GM-False
2) Decay-False
3) Auto build-False
4) Debug-False
5) Hidden layer-a
6) Training time -500
7) Learning rate-0.3
8) Momentum-0.2
9) Nominal to Binary-True
10) Normalize-True
11) Reset-false
12) Seed -0
13) Validation set-0
14) Validation threshold-20

Table 10 : Accuracy comparison for various data sets using Chi square method

Data set	J48	Naïve Bayes	PART	Our method
Weather	100	92.8	85.7	100
Pima	85.1	76.3	81.2	98.82
hypothyroid	99.8	95.44	99.86	97.64
breast cancer	75.87	75.17	80.06	97.64
liver disorder	84.6	56.8	86.08	70
primary tumor	61.35	56.04	61.35	83.18
heart stalog	91.48	85.18	94.4	97.7
lymph	93.23	87.16	95.27	100
heart disease A.P	95	72.5	95	100
Average	**87.3**	**77.4**	**86.5**	**93.8**

Table 11 : Parameters of GA

Sl.no	Parameter	value
1	Cross over rate	0.6
2	Mutation rate	0.5%-1%
3	Population size	20-30
4	Selection	Basic roulette wheel selection

Table 12 : Accuracy comparisons GA+ANN, J48, NB, PART

Data set	J48	Naïve Bayes	PART	GA
Weather	100	92.8	85.7	100
Pima	85.1	76.3	81.2	78.3
hypothyroid	99.8	95.44	99.86	97.37
breast cancer	75.87	75.17	80.06	95.45
liver disorder	84.6	56.8	86.08	84.63
primary tumor	61.35	56.04	61.35	83.18
heart stalog	91.48	85.18	94.4	99.62
lymph	93.23	87.16	95.27	99.32
heart disease A.P	95	72.5	95	100
average	**87.3**	**77.4**	**86.5**	**93.09**

Figure 10 : Accuracy with feature subset selection for chi square

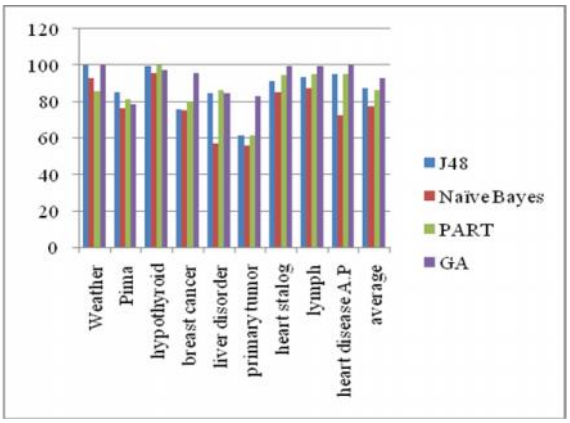

Figure 11 : Accuracy comparision for GA,j48,NB,PART

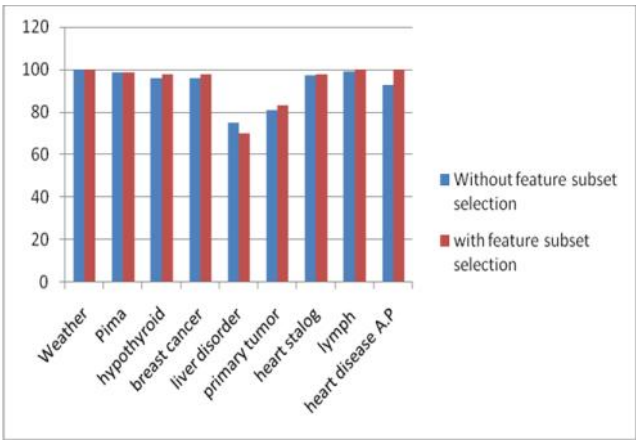

Table 13 : Accuracy comparison of PCA, χ 2,GA

Data set	PCA	Chi square	GA
Weather	100	100	100
Pima	98.82	98.82	78.3
hypothyroid	97.06	97.64	97.37
breast cancer	97.9	97.64	95.45
liver disorder	95.07	70	84.63
primary tumor	80	83.18	83.18
heart stalog	98.14	97.7	99.62
lymph	99.32	100	99.32
heart disease A.P	100	100	100
average	**96.2**	**93.8**	**93.09**

Parameters for genetic search have been given in table 11.crossover rate should be high so we set the cross over rate as 60%.Mutation rate should be very low. Best rates reported are about 0.5%-1%. Big population size usually does not improve the performance of Genetic algorithm. So good population size is about 20-30.In our method we used roulette wheel selection method. The comparison of GA+ANN system with other classification systems has been given in table12 and figure 11.The results acquired reveals that integrating GA with ANN performed well for many data sets and especially for heart disease A.P .we compared three feature selection methods in table 13.GA works well for 6 data sets .overall PCA with ANN works performed better than other classification methods.

VI. Conclusion

In this paper we have proposed a new feature selection method for heart disease classification using ANN and various feature selection methods for Andhra Pradesh Population. We applied different feature selection methods to rank the attributes which contribute more towards classification of heart disease, which indirectly reduces the no. of diagnosis tests to be taken by a patient. Our experimental results indicate that on an average with ANN and feature subset selection provides on the average better classification accuracy and dimensionality reduction. Our proposed method eliminates useless and distortive data. This research will contribute reliable and faster automatic heart disease diagnosis system, where easy diagnosis of heart disease will saves lives. Coronary heart disease can be handled successfully if more research is encourage din this area.

References Références Referencias

1. Pang-Ning Tan et.al, "Introduction to data mining" Pearson (2009).
2. Nang Y."The Hand book of data mining", Lawrence Erlbaum associates (2003).
3. Tsymbal et .al, Guest editorial" introduction to the special section on mining biomedicaldata". IEEE Transactions on information technology in Biomedicine, Vol 10, no 3, pp 425-428(2006).
4. Liping A and Lingyun T. "A rough neural expert system for medical diagnosis". Service systems and service management Vol 2, pp 1130-1135(2005).
5. MA. Jabbar et.al., "Predictions of risk score for heart disease using associative classification and hybrid feature subset selection". In Proceedings of 12th International Conference on Intelligent Systems Design and Applications (ISDA), Cochin 2012 pp 628-634.
6. http://nptel.iitm.ac.in/course/106108657/10
7. T.K.V Iyengar et.al, "Probability and statistics" Scand Publishers pp 301-302(2008).
8. Bela shah and Prashant mathur, "Surveillance of cardiovascular disease risk factors in India: The need and scope", Review article, Indian journal of medicine pp 634-642 Nov (2010).
9. Rajeev gupta, "Recent trends in CHD epidemiology in India", Indian Heart journal B4-B18 (2008).
10. M. Ambarasi et.al, "Enhanced Prediction of heart disease with feature subset selection using genetic algorithm", JEST Vol2 (10) pp 5370-5376(2010).
11. MA. Jabbar et.al," Knowledge discovery using associative classification for heart disease prediction", AISC 182 pp29-39, Springer-Verlag (2012).
12. MA. Jabbar et.al, "Knowledge discovery from mining association rules for heart disease prediction", JAJIT, Vol 41(2) pp 45-51 (2012).
13. MA.Jabbar, B.L Deekshatulu, Priti Chandra, "An evolutionary algorithm for heart disease prediction", CCIS pp 378-389springer Verlag(2012)
14. MA. Jabbar, B.L Deekshatulu, Priti Chandra, "cluster based association rule mining for heart disease prediction", JATIT vol. 32 no. 2 October (2011).
15. Sellappan et al., "Intelligent heart disease prediction system using data mining techniques", IEEE (2008).
16. MA. Jabbar, B.L Deekshatulu, Priti Chandra, "Graph based approach for heart disease prediction", LNEE pp 361-369 Springer Verlag 2012.
17. Sarojini balakrishnan et al, "feature subset selection using FCBF in type II data bases", ICIT Thailand March (2009).
18. MA. Jabbar, B.L Deekshatulu, Priti Chandra, "Heart Disease prediction system using associative classification", ICECIT 2012, Elsevier vol. no. 1 pp 183-192.
19. Tuned it Repository www.tunedit.com