Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

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

M. Akhil jabbar[1]

[1] AEC,BHONGIR ,A.P, INDIA.

Received: 12 April 2013 Accepted: 4 May 2013 Published: 15 May 2013

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.

Index terms— andhra pradesh, artificial neural network, chi-square, data mining, feature subset selection, genetic search, heart disease, principal component analy

1 Introduction

ata 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 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. Year 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.

2 II.

3 Basic Concepts

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

4 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. 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 **??** shows ANN and figure **??** 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.

5 D

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. 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.

6 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. 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. 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 (? 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 .

7 Procedure for principle component analysis is shown below

r X 2 = ï ?" (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 **??** : Chi square computation 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. 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.

8 Global Journal of Computer Science and Technology

Volume XIII Issue III Version I Coronary heart disease occurs when the arteries of the heart that normally provide blood and oxygen to the heart are narrowed or even completely

9 7)

Poorly controlled stress an danger can lead to heart attacks and strokes.

10 d) Genetic Search

11 Stress:

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

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

12 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.

13 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, isM v=E {v} (1)

The covariance matrix is **??**——————————————Total no. of samples in test data PROPOSED ALGORITHMCv=E{(V-M v){ V-M v

V.

14 Results and Discussion

We have evaluated accuracy of our approach on various data sets taken from tuned it repository [19] The rank of the attributes is done w.r.t the values of PCA, and ? 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

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 ? 2 indicates 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.

15 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 **??** shows parameters for performing multilayer preceptor. we set learning grates 0.3 and training time as 500.figure 8 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

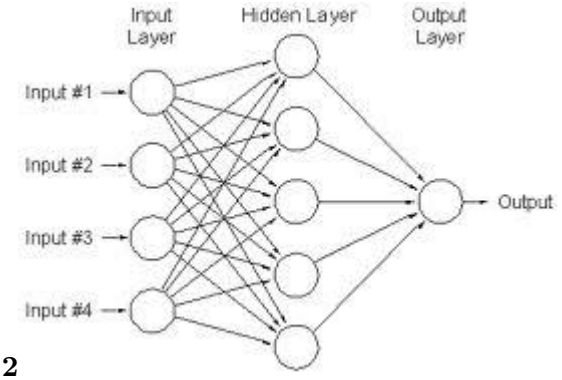
classification systems has been given in table12 and figure **??**1.The results acquired reveals that integrating GA with ANN performed well for many data sets and especially for heart disease A.P .we compared three feature selection methods in table 13.GA works well for 6 data sets .overall PCA with ANN works performed better than other classification methods.

VI.

16 Conclusion

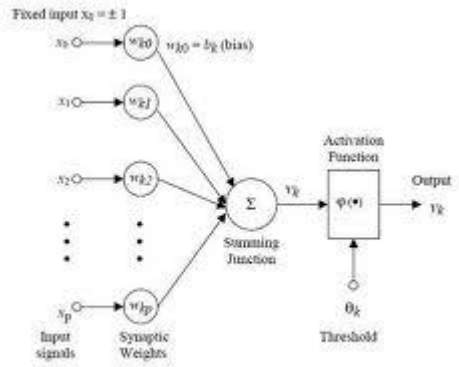
In this paper we have proposed a new feature selection method for heart disease classification using ANN and various feature selection methods for Andhra Pradesh Population. We applied different feature selection methods to rank the attributes which contribute more towards classification of heart disease, 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.

Figure 1:



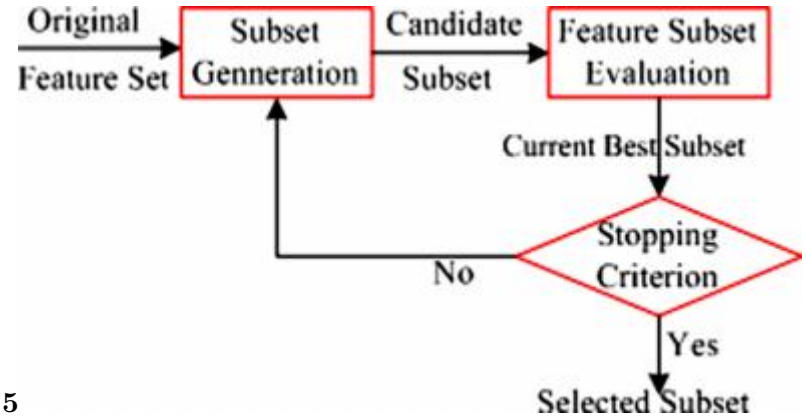
12

Figure 2: Figure 1 :Figure 2 :



4

Figure 3: Step 4 :



5

Figure 4: 5)

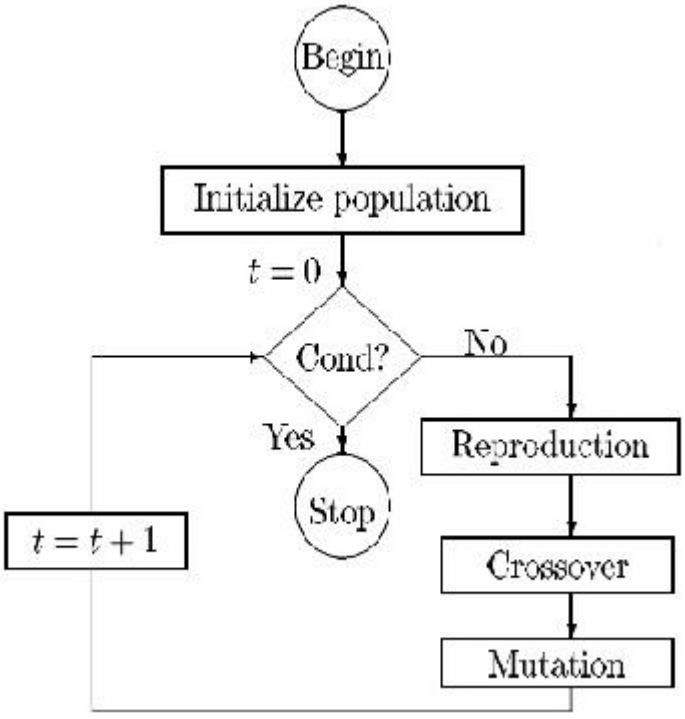
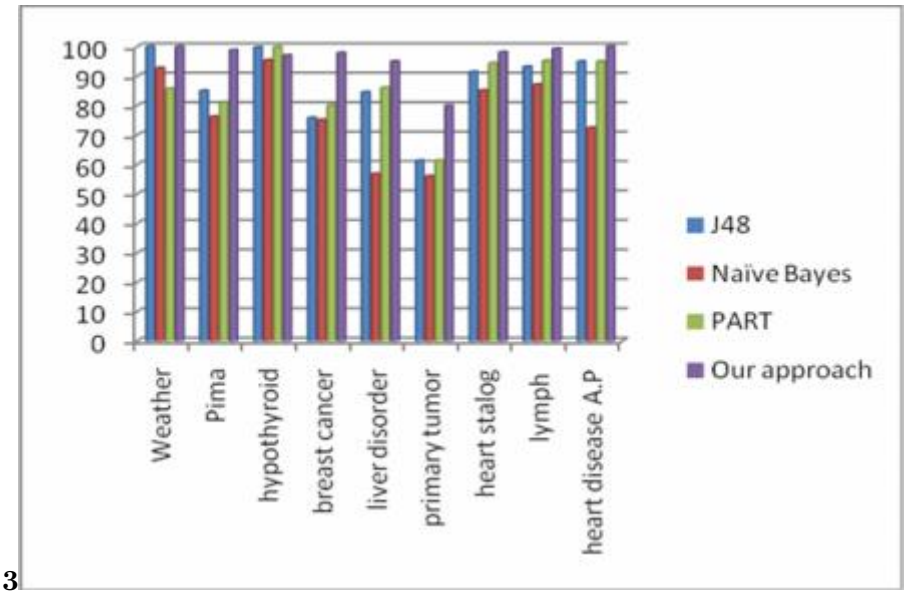


Figure 5:



3

Figure 6: Figure 3 :

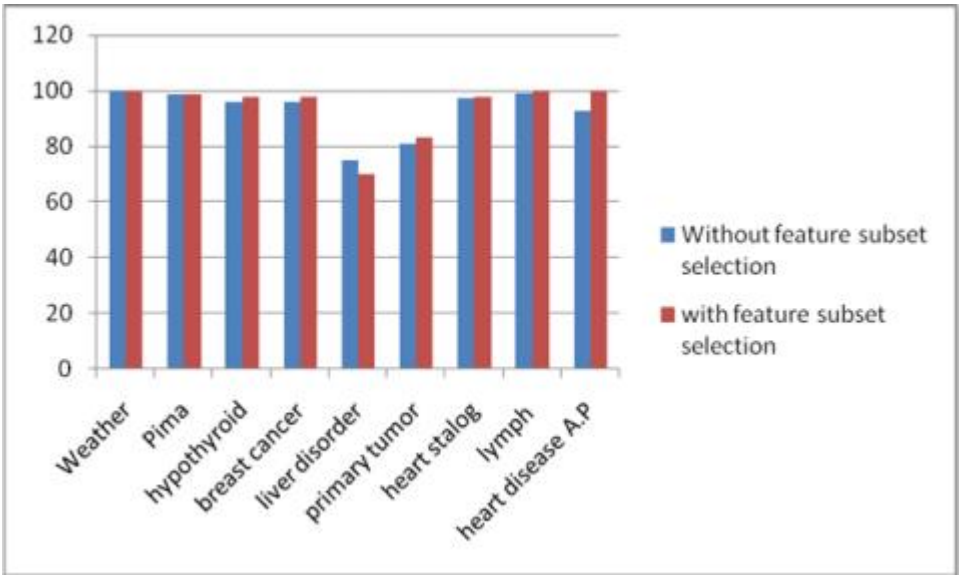
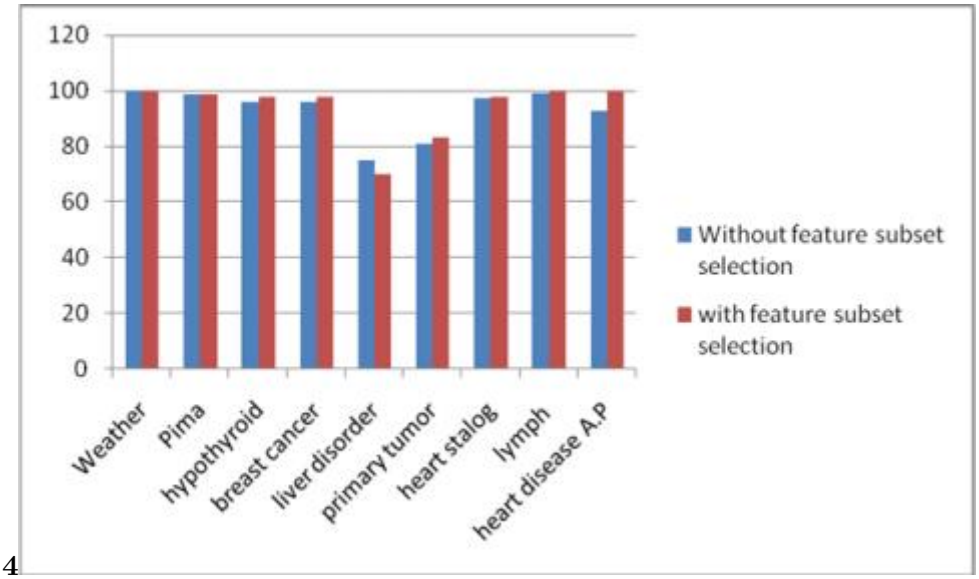


Figure 7: D



4

Figure 8: Figure 4 :

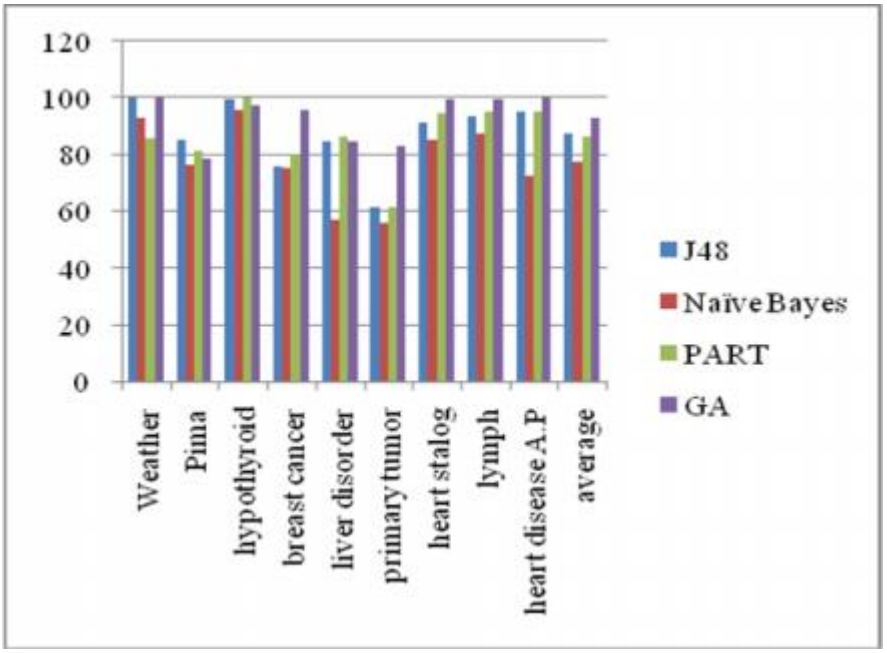
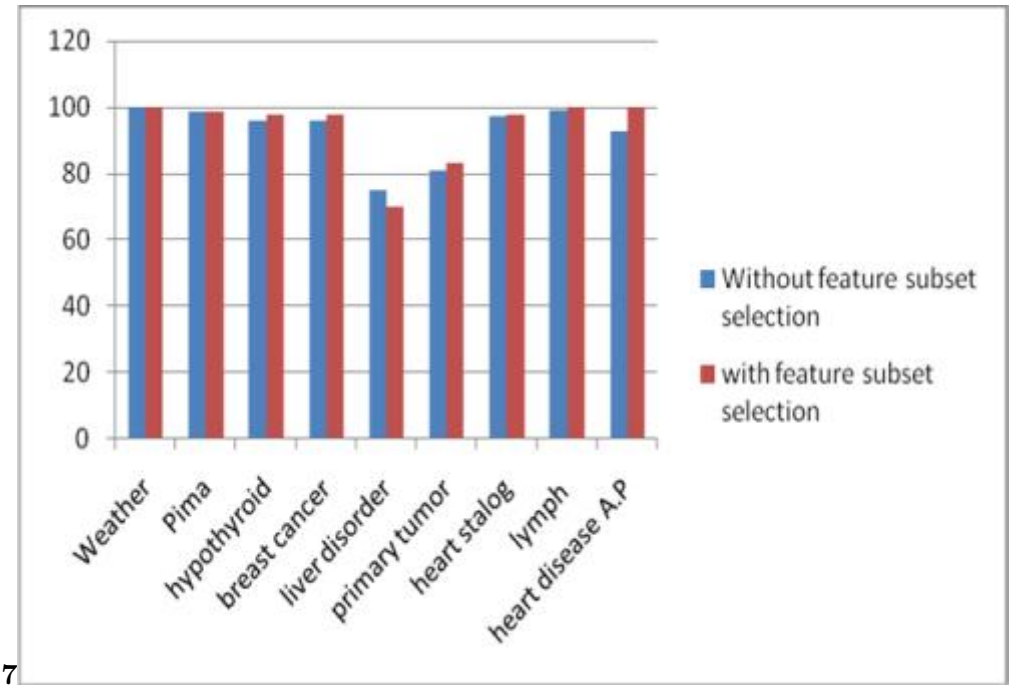


Figure 9:



7

Figure 10: Figure 7 :

2

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

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

No.	Outlook	Temperature	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

Figure 11: Table 2 :

3

Accuracy= No. of samples correctly classified in test Data set

Figure 12: Table 3 :

013
2
Year
10
Volume XIII Issue III Version I
D D D D) D
(
Global Journal of Computer Science and Technology

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

© 2013 Global Journals Inc. (US)

Figure 13:

4

and table 8 clearly

Figure 14: Table 4 :

5

Figure 15: Table 5 :

6

Figure 16: Table 6 :

7

Figure 6 : Parameters of PCA

Figure 17: Table 7 :

8

Figure 18: Table 8 :

9

Figure 19: Table 9 :

10

Figure 20: Table 10 :

11

013 2
Year

12
Volume XIII Issue III Version I (D D D D) D

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

Global Journal of Computer Science and Technology

Data set	Weather	Pima	hypothyroid	breast cancer	liver disorder	primary tumor	heart stalog	lymph

© 2013 Global Journals Inc. (US)

Figure 21: Table 11 :

13

Figure 22: Table 13 :

[Liping and Lingyun ()] 'A rough neural expert system for medical diagnosis'. A Liping , T Lingyun . *Service systems and service management*, 2005. 2 p. .

[Ma et al. ()] 'An evolutionary algorithm for heart disease prediction'. Ma , B Jabbar , Priti Deekshatulu , Chandra . *CCIS pp 378-389springer Verlag*, 2012.

[Ma et al. (2011)] 'cluster based association rule mining for heart disease prediction'. Ma , B Jabbar , Priti Deekshatulu , Chandra . *JATIT* October (2011. 32 (2) .

[Ambarasi ()] 'Enhanced Prediction of heart disease with feature subset selection using genetic algorithm'. M Ambarasi . *JEST* 2010. 2 (10) p. .

[Sarojini Balakrishnan ()] *feature subset selection using FCBF in type II data bases*, Sarojini Balakrishnan . 2009. ICIT Thailand March.

[Ma et al. ()] 'Graph based approach for heart disease prediction'. Ma , B Jabbar , Priti Deekshatulu , Chandra . *LNEE pp*, 2012. Springer Verlag. p. .

[Tsymbal ()] 'Guest editorial" introduction to the special section on mining biomedicaldata'. Tsymbal . *IEEE Transactions on information technology in Biomedicine* 2006. 10 p. .

[Ma et al.] *Heart Disease prediction system using associative classification*, Ma , B Jabbar , Priti Deekshatulu , Chandra . ICECIT 2012. Elsevier. p. .

[Sellappan ()] 'Intelligent heart disease prediction system using data mining techniques'. Sellappan . *IEEE* 2008.

[Pang-Ning Tan ()] *Introduction to data mining*, Pang-Ning Tan . 2009. Pearson.

[Ma and Jabbar ()] 'Knowledge discovery from mining association rules for heart disease prediction'. Ma , Jabbar . *JAJIT* 2012. 41 (2) p. .

[Ma and Jabbar ()] 'Knowledge discovery using associative classification for heart disease prediction'. Ma , Jabbar . *AISC* 2012. Springer-Verlag. 182 p. .

[Ma and Jabbar ()] 'Predictions of risk score for heart disease using associative classification and hybrid feature subset selection'. Ma , Jabbar . *Proceedings of 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, (12th International Conference on Intelligent Systems Design and Applications (ISDA)Cochin) 2012. p. .

[Iyengar ()] *Probability and statistics" Scand Publishers pp*, T K Iyengar . 2008. p. .

[Gupta ()] 'Recent trends in CHD epidemiology in India'. Rajeev Gupta . *Indian Heart journal* 2008. p. .

[Shah and Mathur ()] *Surveillance of cardiovascular disease risk factors in India: The need and scope*, Bela Shah , Prashant Mathur . 2010. Nov. (Review article, Indian journal of medicine pp 634-642)

[Nang ()] *The Hand book of data mining*, Y Nang . 2003. Lawrence Erlbaum associates.

[Tuned it Repository www.tunedit.com References Références Referencias] *Tuned it Repository www.tunedit.com References Références Referencias*,