Heart Disease Prediction using Lazy Associative Classification

M.AKHIL JABBAR
Associate professor,
Aurora's Engineering College, Bhongir
Jabbar.meerja@gmail.com

Dr B.L Deekshatulu
Distinguished Fellow, IDRBT
RBI (Govt of INDIA)
deekshatulu@hotmail.com

Dr Priti Chandra
Senior Scientist,ASL
DRDO, Hyderabad
priti_murali@yahoo.com

ABSTRACT

Medical data mining is used to extract knowledgeable information from a huge amount of medical data. Associative classification is a rule based new approach which integrates association rule mining and classification, if applied on medical data sets, lends them to an easier interpretation. It selects a small set of high quality rules and uses these rules for prediction. Heart disease rates among the major cause of mortality in developing countries and is rapidly becoming so in developing countries like India. India is the second most populous country in the world with an estimated population of over 1 billion. Rapid industrialization and urbanization have resulted in tremendous growth in the economy over the last deacde.Concurrently India has also seen an exponential rise in prevalence of Heart disease. It has predicted that CVD will be the most important cause of mortality in India by the year 2015, and A.P is in risk of CVD.Hence a decision support system should be proposed to predict the risk score of a patient, which will help in taking precautionary steps like balanced diet and medication which will in turn increase life time of a patient. Through this paper we propose a lazy associative classification for prediction of heart disease in Andhra Pradesh and present some experimental results which will help physicians to take accurate decisions.

Keywords: Andhra Pradesh, Lazy associative classification, Heart disease, Priniciple component analysis

1. INTRODUCTION

The medical data is highly voluminous in nature. With increasing size of the amount of data stored in medical data bases, there is a need for effective data mining techniques. Medical data are processed and analyzed using different data mining techniques to extract useful information. This extracted information is valuable for decision making and for diagnosis, risk analysis and predictions. Classification is the most important technique in data mining used to solve problems related to areas like medical data mining, finance and others. Building precise and efficient classifier is the key challenge in data mining. In recent years a new approach called associative classification is proposed which integrates association and classification [1].It adopts association rule mining algorithms like Apriori and FP Growth to generate class association rules. It selects small set of high quality rules for prediction.Merschmann and plastino [2]classified association rule mining into two categories.1) Eager classification 2) Lazy classification. First method constructs the generalized model to predict the class label whereas lazy associative classification delays the processing of data until a new test instance is needs to be classified and does not build the model to classify the test instance.

Coronary heart disease is the leading in India and world wide.CHD affects people at younger ages in low and middle income countries, compared with high income countries, thereby having a greater economic impact on low and middle income countries. India is the second most populous country in the world with an estimated population of over 1 billion. Rapid industrialization and urbanization have resulted in tremendous growth in the economy over the last decade. Concurrently; India has also seen an exponential rise in prevalence of Heart disease[3].WHO has declared India the CAD capital of the world[4].Data from the registrar general of India shows that heart disease are major cause of dealth in india,studies to determine the precise cause of death in A.P have revealed that CVD cause about 30%of deaths in rural areas[5].Hence there is a need for decision support system for predicting heart disease. The objective of this work is to build a model for heart disease prediction which helps the physicians to make accurate diagnosis.

The paper is organized as follows: In the next section, some related concepts associative classification, principle component analysis and heart disease are presented. In section 3 we provide overview of work related to associative classification and heart disease. In section 4 proposed algorithm is presented. Performance results are presented in section 5 and conclusions are given in section 6.

978-1-4673-5090-7/13/$31.00 ©2013 IEEE

2. BASIC CONCEPTS

In this section we will discuss associative classification, principle component analysis and heart disease epidemioly in India and in Andhra Pradesh.

2.1 *Associative Classification*

Association rule mining and classification are two important functionalities in data mining. It is possible to build more accurate classifier if we focus on limited set of association rules where the consequent of the rule is restricted to class. There is growing evidence that merging classification and association rule mining together can produce more efficient and accurate classification system than traditional classification techniques. Normally association rule mining performs global search for strong rules. The richness of the rules gives the true classification. Associative classification was first proposed by Liu et al[1] and named CBA.This method implements the famous apriori algorithm. In recent years , a number of different classification algorithms have been proposed based on association rules[6][7].An actual class association rule is represented in the form$(A_{i1},a_{i1})^(A_{i2},a_{i2})^(A_{1m},a_{im})->C_j$ where antecedent of the rule is an itemset and consequent is a class. The main task of associative classification is to build a model that is able to predict the class label of unknown sample, known as the test data set, as accurately as possible. Associative classification techniques are generally categorized in two types: Eager and lazy approach. Eager associative classification constructs a generalization model from the training data set before any unknown sample is received for classification. They classify new instances by directly using the learned model. Whereas lazy approach do not previously build a generalization model from the training data set to classify new instances. For each instances to be classified, they process the stored training samples.

2.2 *Principal Component Analysis*

In medical data mining we often encounters situations where there are large no.of features in the data base. In such situations it is very likely that subsets of features are highly correlated with each other. The accuracy of the classifier will suffer if we include highly correlated with each other. One of the key steps in the KDD is to find the ways to reduce dimensionality without compromising the accuracy.

Principal component analysis is a well known scheme for dimensionality reduction. The reduced dimensions are chosen in a way that captures essential features of the data with very little loss of information.

Method for computing Principal component analysis:

1) Get some data and obtain input matrix table.
2) Subtract the mean from the data set in all the n-dimensions
3) Calculate the covariance matrix
4) Calculate Eigen vectors and Eigen values from the covariance matrix
5) Choosing components and forming feature vector
6) Deriving new data set and rank the attributes.

Example:

Consider the weather data set. By applying the principal component analysis attributes are ranked.

Table 1. Weather data set

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

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 0
0.04 -0.09 0.04 0.23 -0.13 -0.09 0 1

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:

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

Based on the ranking we can remove the redundant attributes.

2.3 Heart disease

Before defining heart disease it is better to know functions performed by our heart. The heart is one of the most important organs in our body. It is a four chambered mechanical pump made of complex muscles. These four chambers are separated by valves and divided into two halves. Each contains one chamber called an atrium and one called a ventricle. The atria collect blood and the ventricles contract to push blood out of the heart. The right half of the heart pumps oxygen poor blood to the lungs where blood cells can obtain more oxygen. Newly oxygenated blood travels from the lungs into the left atrium and the left ventricle. This left ventricle pumps the newly oxygen rich blood to the organs and tissues of the body. This oxygen provides our body with energy and essential to keep our body health [8].According to world health statistics report, compiled by WHO, mortality due to cardiac causes has overtaken, mortality due to all cancers put together. In India alone, we have about 4,280 sudden cardiac deaths per lakh deaths annually. By the year 2030, India will rank among the highest risk for heart disease [9].

Risk factors of Coronary Heart Disease:

Over 300 risk factors have been associated with CHD.The major established risk factors are categorized into three.

1) **Major modifiable risk factors**
   a) high BP b) abnormal blood lipids c)use of tobacco d) obesity e) physical inactivity f) unhealthy diets g)diabetes mellitus
2) **Other modifiable risk factors**
   a) Low economic status b) mental ill health c) alcohol used) psychosocial stress e)use of certain medication f)lipoprotein g)LVH
3) **Non modifiable risk factors**
   a) Age b) gender c)family history)ethnicity

Symptoms of Coronary Heart Disease

1) Chest pain 2) syncope 3) dyspnea 4) heart palpitations 5) fatigue or day time sleepiness

Heart disease in Andhra Pradesh:

Several studies show a high prevalence of diabetes and other risk factors for heart disease in Andhra Pradesh. The prevalence of risk factors was shown in table 2.sudden cardiac death contributed to 10% of overall mortality in Andhra Pradesh. Study in rural Andhra Pradesh showed the CVD was the leading cause of mortality accounting for 32%of all deaths, a rate as high as Canada (35%) and U.S.A

Table 2. Prevalence of risk factors in A.P

Risk Factor	Prevalence %
Diabetes	24%
High BP	28%
Cholesterol problem	58%
Smoking	24%
Obesity	36%

Heart disease in India:

CHD prevalence appears to be worsening in India. Prevalence of CHD will rise in India compared to china and established market economies from the year 1990-2020[10].

Leeder et al [11] estimated total years of life lost due to CVD among Indian men and women aged 35-64 to be higher than the countries china and Brazil as shown in Table 3

Table 3.Etimates of total years of life lost due CVD in 2000 and 2030

country	2000		2030	
	Total years of life lost	Rate per 1 lakh	Total years of life lost	Rate per 1 lakh
India	9221165	3572	17937070	3707
Brazil	1060840	2121	1741620	1957
China	6666990	1595	10460030	1863

Primary prevention strategies to control risk factors impact the prevalence of heart disease.

3. LITERATURE REVIEW

Several research techniques have been proposed in the literature on heart disease using data mining have motivated our work. Carlos Ordonez et al[12]introduced an improved algorithm to discover constrained association rules to predict heart disease. Data mining by soft computing methods for the coronary heart disease data base was proposed by Akira hara et al in[13].The accumulation of their research results using the CHD data base will become profitable data for the data mining domain. Shantakumar patil et al proposed extraction of significant patterns from heart disease ware house for heart attack prediction [14]. They used MAFIA algorithm to extract patterns relevant to heart disease. Positive and negative association rule analysis in health care database was proposed by E.Ramaraj et al[15].Their paper focuses a new algorithm called bit array negative pos that mines positive and negative rules from the real time medical data base.Sellappan palaniaapn et al developed a prototype intelligent heart disease prediction system using data mining techniques[16]. They implemented the model using .net frame platform. Intelligent and effective heart attack prediction system using data mining and artificial neural network was proposed by shantakumar patil et al [17].They employed neural network with back propagation as the training algorithm. Genetic algorithm based heart disease prediction was proposed by akhil jabbar et al [18].They used G-Mean measure to effectively reduce no. of rules generated by association rule mining. Matrix based approach for heart disease prediction was proposed by akhil jabbar et al [19].They applied transaction reduction while generating association rule mining. Graph based approach for heart disease prediction was proposed by akhil jabbar et al[20].They employed maximum clique based weighted association rule mining to predict heart disease. Feature selection using FCBF in type 2 diabetes was proposed by sarojini balakrishnan[21].M.A jabbar et al proposed knowledge discovery using associative classification for heart disease prediction in [22].Cluster based association rule mining for heart disease prediction was proposed in[23].Their method combines the clustering and bit sequence. Prediction of risk score for heart disease using associative classification was proposed by jabbar et al[24].They applied feature subset selection measures like SU,IG,and Genetic search using associative classification. Heart disease prediction system using associative classification, hypothesis testing and genetic algorithm was proposed in [25] .They employed Z-Statistic measure for hypothesis testing and genetic algorithm for heart disease prediction for Andhra Pradesh population.

In this research paper we propose lazy associative classification to predict heart disease for Andhra Pradesh population.

4. PROPOSED METHOD

In this section we propose a method to predict the heart disease using classification. Lazy associative classification method induces class association rules specific to test instance. The lazy learning approach projects the training data D, only on those features in the test data. Generally Apriori based rule generation algorithm generates 2^k -1 rules for the data set with k items [26].so it leads to high computation cost. To reduce the no. of rules generated and to improve accuracy we used information centric attribute PCA in lazy associative classification.PCA is dimensionality reduction technique used to find a new set of attributes that better captures the variability of the data .the goal of PCA is to find a transformation of the data that satisfies the following four properties.

1) Each pair of new attributes has zero covariance
2) The attributes are ordered w.r.t how much the variance of the data each attribute captures.
3) The first ranked attribute captures as much of the variance of the data as possible.
4) Subject to the orthoganolity requirement, each successive attributes captures as much of the remaining variance as possible.[27]

4.1 Proposed algorithm

Step 1) let D be the set of all n training instances and T be the m test instances

Step 2) for each t_i € T do

Step 3) let D_t be the projection Of D on features only From t_i

Step 4) In the process of generating the class association rules, instead of considering all the attributes, apply PCA and rank all attributes. The attribute with highest ranking is used to generate the class association rules. This approach generates n*m rules for a single test instance with n non class attributes and m classes in the entire data set. If t Test cases are to be predicted the no. Of rules generated will be t*n*m.After identifying the principle

component attribute, the subsets are generated. For each generated subset, probability values are calculated.

Step 5) the decision of which class will be assigned to test instance X is based on the analysis of the subsets of attributes values with the highest posterior probabilities.

Steps 6) find the accuracy of the data set.

$$\text{Accuracy} = \frac{\text{no. of correctly predicted test data}}{\text{Total no. of test data}}$$

5. EXPERIMENTAL RESULTS

The proposed lazy associative classification was tested on 7 data sets from UCI repository [28] and one real life data set Andhra Pradesh heart disease data set. A brief description about the data set is presented in table 4.Table 5and Table 6 shows the accuracy for various non medical and medical data sets. Table 5 shows that our approach has improved10.8%against J4.8 and 19.8%improvement over naïve bayes for non medical data sets .Our approach reached10.26% improvement over J4.8 and8.6% improvement against naïve bayes respectively for heart disease data set. Figure 1 shows accuracy comparison for medical data set.

Table 4 Data set description

Data set name	Attributes	Instances
Weather data	5	14
Contact lenses	5	24
Students data	6	498
Liver disorder	7	345
Diabetic	9	768
Heart disease A.P	12	40

Table 5 Accuracy comparison for non medical data sets

Data set	**J 48**	**Naïvebayes**	**Our proposed**
Weather data	85.7	57.14	87.5
Contact lenses	70.83	70.83	83.3
Students data	81.7	84.7	100
Average	79.41	70.89	90.2

Table 6 Accuracy comparison for medical data sets

Data set	J 48	Naïvebayes	Our Proposed
Liver disorder	63.23	55.36	80
Diabetic	77.06	77.59	75
Heart disease A.P	73.91	86	90
Average	71.4	72.98	81.66

Table7 A. P Heart disease attributes

Data set
Age
Gender
Diabetic
BP Systolic
BP Dialic
Height
Weight
BMI
Hypertension
Rural
Urban
Disease Status

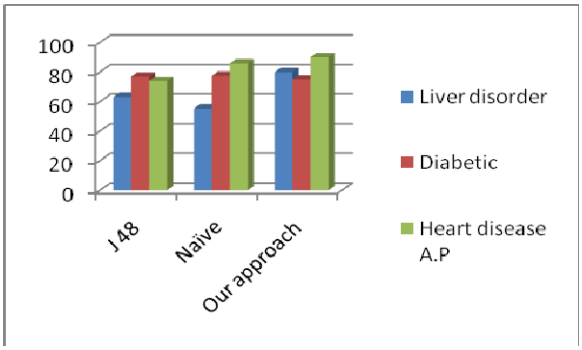

Fig 1.comparision of accuracy for medical data sets

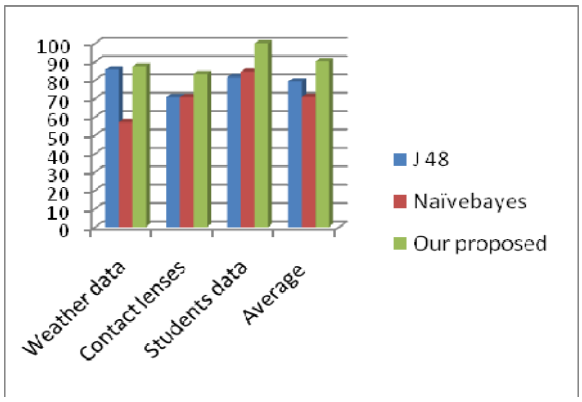

Fig 2. comparision of accuracy for non medical data sets

Fig 2 shows the comparison of classification accuracy for various non medical data sets. We compared our proposed approach with J48 and Naïve bayes .As shown in table 6 for Pima Indian diabetes accuracy is slightly reduced by our approach. This is due to the fact in Pima Indian diabetes data sets are highly distinct and values for some attributes are missing. Table 7 shows attributes selected for A.P Heart disease data set. The attributes are collected from various corporate hospitals and opinion from expert doctors. Classification rules generated from our approach are

1) gender=male BP Dialic>80 ==> Disease=yes
2) gender=male BP Systolic=>120 hypertension=yes urban=yes ==> Disease=yes
3) Bp Systolic Systolic=>120 BP Dialic >80 urban=yes 5 ==> Disease=yes
4) age=>4 5 ==> Disease=yes 5
5) height=>152 weight=>52 gender=male hypertension=yes urban=yes 4 ==> Disease=yes

The patterns of heart disease in Andhra Pradesh population has been reported as

1) Males are affected more than females.
2) Hypertension accounts for about 47.8% of all cases.
3) 53% of Males whose systolic blood pressure above 120 are associated with CHD.
4) 87%of females whose systolic blood pressure above 120 are associated with CHD.
5) Among the participants of the study people who are residing in urban and having BMI>25 accounts for 75%of all the cases.
6) Above 90%of the cases are from urban areas.
7) 43%of the people who are having diabetic are associated with heart disease.
8) 52%of the population whose BMI>25 are associated with heart disease.
9) Among the participants of the study 63%of the males are obeys and 37%of females are obeys.
10) Males who attain at the age above 45 are more associated with heart disease (73%)

6. CONCLUSION

In this research paper we presented a lazy data mining approach for heart disease classification. We applied information centric arrtibute measure PCA to generate class association rules. This class association rules will be used to predict the occurrence of heart disease. The system is designed for Andhra Pradesh population. Andhra Pradesh is in risk of more death due to heart disease. Heart disease can be handled successfully if more research is encouraged to develop prediction system in this area.

REFERENCES

[1] Liu, B Hsu et al., Integrating classification and association rule mining. In ACM international conference on SIGKDD (1998)
[2] Merschmann L plastino.Lazy data mining approach for protein classification.IEEE transactions on nano science vol 16 no 1 pp 36-42(2007)
[3] Madhavan et al .Epidemiology of sudden cardiac death in rural south India, IPEJ 11(4)93-102(2011)
[4] Reddy KS, Yusuf. Emerging epidemic of CVD in developing countries. Circulation 1998:596-601
[5]Rajeev gupta.Recent trends in CHD epidemiology in India, Indian heart journal pp B4-B18 (2008)
[6] Yin and Han CPAR: Classification based on association rules .in proceedings of SIAM KDM pp 369-376(2003)
[7]S.P Syed Ibrahim,K.R Chandran.,Efficient associative classification using genetic network programming,IJCA Vol 29 no 6 sep (2011).
[8] Benjamin M .Introduction to heart disease http://www.mentalhelp.net
[9] The weak. Mental health, august 30, pp 18-19(2009)
[10] Mark D Huttman.Coronary heart disease in India, centre for chronic disease control.
[11]Leader s et al .Race against time: The challenge of CVD in developing Economies. Columbia University NY 2004
[12]Carlos Ordonez et al, Mining constrained association rules to predict heart disease.
[13]Akira hara et al.Data mining by soft computing methods for the coronary heart disease data bases. In 4th IWCIA Dec (2008)
[14]Shantakumar B patil et al.Extraction of significant patterns from heart disease warehouses for heart attack prediction,IJCSNS vol 9 no 2 Feb(2009)
[15]E Ramaraj N venkateshan .Positive and negative associative rule analysis in health care data bases, IJCSNS vol 8 no 11 (2008)
[16] Sellappan palaniapan et al .Intelligent heart disease prediction system using data mining techniques IJCSNS vol 8 no 8 Aug (2008)
[17] Shantakumar B patil et al.Intelligent heart disease prediction using data mining and artificial neural network,EJSR vol 31,no 4 pp 642-656 (2009)

[18] MA.Jabbar, B.L.Deekshatulu and Priti Chandra.: An evolutionary algorithm for heart disease prediction, ICIP, CCIS 292 PP 378-389, Springer-Verlag (2012)

[19] M.A.Jabbar, B.L.Deekshatulu,Priti Chandra, Knowledge discovery from mining associative rules for heart disease prediction JATIT Vol 41,2 pp 45-51(2012)
[20] M.A.Jabbar, B.L.Deekshatulu, Priti Chandra, Graph based approach for heart prediction, LNEE pp 361-369Springer verlag 2012
[21] Sarojini balakrishnan et a Feature selection using FCBF in type II Diabetes data bases, ICIT Thailand march (2009)

[22] M.A.Jabbar, B.L.Deekshatulu, Priti Chandra, knowledge discovery using associative classification for heart disease prediction" AISC 182 pp 29-39, Springer Verlag 2012
[23] M.A.Jabbar, B.L.Deekshatulu, Priti Chandra, Cluster based association rule mining for heart disease prediction, JATIT, Vol 32 no 2 October (2011)
[24] M.A.Jabbar, B.L.Deekshatulu, Priti Chandra**.** Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection.In International conference on Intelligent Systems Design and Applications .IEEE ISDA 2012 Cochin INDIA pp 628-634
[25] M.A.Jabbar, B.L.Deekshatulu, Priti Chandra Heart Disease Prediction System using Associative Classification and Genetic Algorithm .In ICECIT 2012 INDIA.Vol (1)pp183-192
[26] S.P Syed Ibrahim et al .An evolutionary approach for rule sets selection in a class based associative classifiers,EJSR Vol 50 no 3 pp 422-429(2011)
[27]Pang ning Tan et al.Introduction to Data mining Pearson 2006
[28] www.ics.**uci**.edu/~mlearn.