Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers

Jyoti Soni, Uzma Ansari, Dipesh Sharma
Computer Science
Raipur Institute of Technology, Raipur
C.G., India

Sunita Soni
Computer Applications
Bhilai Institute of technology, Bhilai
C.G., India

Abstract--**The healthcare environment is still 'information rich' But 'knowledge poor'. There is a wealth of data available within the health care systems. However, there is a lack of effective analysis tools to discover hidden relationships in data. The aim of this work is to design a GUI based Interface to enter the patient record and predict whether the patient is having Heart disease or not using Weighted Association rule based Classifier. The prediction is performed from mining the patient's historical data or data repository. In Weighted Associative Classifier (WAC), different weights are assigned to different attributes according to their predicting capability. It has already been proved that the Associative Classifiers are performing well than traditional classifiers approaches such as decision tree and rule induction. Further from experimental results it has been found that WAC is providing improved accuracy as compare to other already existing Associative Classifiers. Hence the system is using WAC as a Data mining technique to generate rule base. The system has been implemented in java Platform and trained using benchmark data from UCI machine learning repository. The system is expandable for the new dataset.**

Keywords- **Weighted Associative Classifier; prediction; UCI machine learning repository; accuracy.**

I. INTRODUCTION

Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of such systems would be extremely advantageous. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource persons at certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together. The World Health Organization has estimated that 12 million deaths occur worldwide, every year due to the Heart diseases. Half the deaths in the United States and other developed countries occur due to cardio vascular diseases. It is also the chief reason of deaths in numerous developing countries. On the whole, it is regarded as the primary reason behind deaths in adults. The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease was the major cause of casualties in the different countries including India. It kills one person every 34 seconds in the United States.

In this paper we have evaluated the performance of new classification approach that uses the experienced Doctor's knowledge to assign the weight to each attribute. More weight is assigned to the attribute having high impact on disease prediction. Experiments have been performed using UCI machine learning dataset and outcome from the experiment it has been found that Weighted Associative Classifier (WAC) is performing well as compare to other already existing Associative Classifiers. Further we have designed a GUI to accept the patient's test results and predict the presence of Heart Disease using CAR rules generated by WAC.

II. RELATED WORK

In [1] three different supervised machine learning algorithms i.e. Naive Bayes, K-NN, Decision List algorithm have been used for analyzing the dataset . Tanagra tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared. **Tanagra** is a data mining suite build around graphical user interface algorithms. The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, and allowing to analyze either real or synthetic data. Tanagra is powerful system that contains clustering, supervised learning, meta supervised learning, feature selection, data visualization supervised learning assessment, statistics, feature selection and construction algorithms. The experiment is performed using training data set consists of 3000 instances with 14 different attributes. The dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing. Based on the experimental results, it is clear that the classification accuracy of Naive Bayes algorithm is better compared to other algorithms.

Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network is implemented in [9] using .NET platform. IHDPS is Web-based, user-friendly, scalable, reliable and expandable system. It can also answer complex “what if” queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease. As a **Data source** a total of 909 records with 15 medical attributes (factors) were obtained from the Cleveland Heart Disease database. Naïve Bayes appears to be most effective as it has the highest percentage of correct predictions for patients with heart disease, followed by Neural Network (with a difference of less than 1%) and Decision Trees. Decision Trees, however, appears to be most effective for predicting patients with no heart disease compared to the other two models.

Genetic algorithm have been used in [4], to reduce the actual data size to get the optimal subset of attributed sufficient for heart disease prediction. Classification is a supervised learning method to extract models describing important data classes or to predict future trends. Three classifiers e.g. Decision Tree, Naïve Bayes and Classification via clustering have been used to diagnose the presence of heart disease in patients. Experiments were conducted with Weka 3.6.0 tool. Data set of 909 records with 13 attributes. All attributes are made categorical and inconsistencies are resolved for simplicity. To enhance the prediction of classifiers, genetic search is incorporated. Observations exhibit that the Decision Tree data mining technique outperforms other two data mining techniques after incorporating feature subset selection but with high model construction time. Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time. Classification via clustering performs poor compared to other two methods.

Association rules represent a promising technique to improve heart disease prediction. Unfortunately, when association rules are applied on a medical data set, they produce an extremely large number of rules. Most of such rules are medically irrelevant and the time required to find them can be impractical. In [11], four constraints were proposed to reduce the number of rules: item filtering, attribute grouping, maximum item set size, and antecedent/consequent rule filtering. When association rules are applied on a medical data set, they produce an extremely large number of rules. Most of such rules are medically irrelevant and the time required to find them can be impractical. A more important issue is that, in general, association rules are mined on the entire data set without validation on an independent sample. To solve these limitations, the author has introduced an algorithm that uses search constraints to reduce the number of rules, searches for association rules on a training set, and finally validates them on an independent test set. Instead of using only support and confidence, one more parameter i.e. lift have been used as the metrics to evaluate the medical significance and reliability of association rules.

III. RESEARCH OBJECTIVES

The main objective of this research is to develop an Intelligent Heart Disease Prediction System using Weighted Associative Classifiers that can be used in making expert decision with maximum accuracy. The system can be implemented in remote areas like rural regions or country sides, to imitate like human diagnostic expertise for treatment of heart ailment. The system can be easily updated as and when the new training data set will be available. Also the system is user friendly and includes less number of steps, as the prediction rules are already generated and stored in Rule Base. The system is helpful for practice nor to confirm his/her decision during heart Disease prediction.

IV. PREDICTIVE DATA MINING REVIEW.

A. Classifiers

Classification rule mining is an important data mining techniques and has been studied extensively over the years. It aims to discover a small set of rules in the database to form an accurate classifier (e.g., Quinlan 1992; Breimanet al 1984). Given a set of cases with class labels as a *training set*, *classification* is to build a model (called *classifier*) to predict future data objects for which the class label is unknown. Several classification models have been proposed, e.g. Bayesian classification, neural networks, and regression and decision trees. Decision tree classification is probably the most popular model, because it is simple and easy to understand. A number of algorithms for constructing decision trees are available including ID3, C4.5, SPRINT, SLIQ, and PUBLIC.

B. Classification based on Association rule mining.

Associative Classification is an integrated framework of Association Rule Mining (ARM) and Classification. A special subset of association rules whose right-hand-side is restricted to the classification class attribute is used for classification. This subset of rules is referred as the Class Association Rules (CARs). Recent studies propose the extraction of a set of high quality association rules from the training data set, which satisfy certain user-specified support and confidence thresholds. Such a method takes the most effective rule(s) from among all the rules mined for classification. Since association rules explore highly confident associations among multiple variables, it may overcome some constraints introduced by a decision-tree induction method, which examines one variable at a time. Extensive performance studies [11] show that association based classification may have better accuracy in general [9].

V. METHODOLOGY.

Weighted Associative Classifier(WAC) is a new concept that uses Weighted Association Rule for classification. Weighted ARM uses Weighted Support and Confidence Framework to extract Association rule from data repository. The WAC has been proposed as a new Technique to get the significant rule instead of flooded with insignificant relation.
The major steps are as follows.

1) Initially, the heart disease data warehouse is pre -processed in order to make it suitable for the mining process.
2) Each attribute is assigned a **weight** ranging from o to 1 to reflect their importance in prediction model. Attributes that have more impact will be assigned a high weight(nearly 0.9)and attributes having less impact are assigned low weight(nearly 0.1)
3) Once the preprocessing gets over, Weighted Association Rule Mining (WARM) algorithm is applied to generate interesting pattern. This algorithm uses the concept of Weighted Support and Confidence framework instead of tradition support and confidence. Rules generated in this step are known as CAR(Classification Association Rule) and is represented as X → Class label where X is set of symptoms for the disease. Example of such rules are (Hypertension, “yes”) →Heart_Disease=”yes” and {(Age,” >62”), (Smoking_habits, “yes”), (Hypertension, “yes”)}→ Heart_Disease=”yes” .
4) These rules will be stored in **Rule Base.**
5) Whenever a new patient’s record is provide, the CAR rule from the rule base is used to predict the class label.

VI. HEART ATTACK PREDICTION SYSTEM USING WEIGHTED ASSOCIATIVE CLASSIFIERS.

A. Weighted Associative Classifiers.

A weighted associative classifiers consists of training dataset T={r_1, r_2, r_3…. r_i…} with set of weight associated with each {attribute, attribute value} pair. Each i^{th} record r_i is a set of attribute value and a weight w_i attached to each attribute of r_i tuple / record. In a weighted framework each record is set of triple {a_i, vi, w_i} where attribute a_i is having value vi and weight w_i, $0<w_j<=1$. Weight is used to show the importance of the item.

1) Attribute Weight: Attribute weight is assigned depending upon the domain. For example item in supermarket can be assigned weight based on the profit on per unit sale of an item. In web mining visitor page dwelling time can be used to assign weight In medical domain symptoms can be assigned weight by expert doctor.

2) Attribute set weight: Weight of attribute set X is denoted by W(X) and is calculated as the average of weights of enclosing attribute. And is given by

$$W(X) = \frac{\sum_{i=1}^{|X|} \text{weight}(a_i)}{\text{Number of attributes in X}}$$

3) Record weight/Tuple Weight: Consider the data in relational table, the tuple weight or record weight can be defined as type of attribute weight. It is average weight of attributes in the tuple. If the relational table is having n number of attribute then Record weight is denoted by $W(r_k)$ and given by

$$W(r_k) = \frac{\sum_{i=1}^{|r_k|} weight(a_i)}{\text{Number of attributes in a record}}$$

4) Weighted Support: In associative classification rule mining, the association rules are not of the form X →Y rather they are subset of these rules where Y is the class label. Weighted support WSP of rule X→Class_label, where X is set of non empty subsets of attribute-value set, is fraction of weight of the record that contain above attribute-value set relative to the weight of all transactions. This can be given as

$$WSP(X \rightarrow Class\_label\ ) = \frac{\sum_{i=1}^{|X|} weight(r_i)}{\sum_{k=1}^{|n|} weight(r_i)}$$

Here n is the total number of records.

5) Weighted Confidence: Weighted Confidence of a rule X→Y where Y represents the Class label can be defined as the ratio of Weighted Support of (X∪Y) and the Weighted Support of (X).

$$\text{Weighted Confidence} = \frac{\text{Weighted Support } (X \cup Y)}{\text{Weighted Support } (X)}$$

B. Data source

The For training the system the Cleveland Heart Disease database consisting of 303 records with 14 medical attributes(factors) have been used. Normalized dataset is available in http://csc.liv.ac.uk/~frans/KDD/Software/LUCS-KDD-DN/DataSets/dataSets.html in .num format. For our experiment the data have been converted in access database. Table I shows the different weight assigned to the attributes.

VII. Experimental Results.

To implement WAC, Java has been used as front end and MS Access as backend tool. Three benchmark Medical data set (UCI Machine learning dataset) i.e. heart.D53.N303.C5.num, breast.D20.N699.C2.num and hepatitis.D56.N155.C2.num have been converted in Ms. Access databases. For the training, entire record has been used and testing has been performed again on entire dataset. The screenshots of heart attack prediction with 2 cases(Heart Diesese and No Heart Disease) are shown in Figure1 and Figure 2. The initial experimental result of Weighted Associative Classifier (WAC) yields following observations.

(1). From Table II it is clear that WAC outperforms as compare to three other associative Classifiers i.e. CBA, CMAR and CPAR in terms of average accuracy. The code for these three Associative Classifiers has been downloaded from following side. http://www.csc.liv.ac.uk/~frans/KDD/Software/CBA/cba.html

(2). Increasing the high weight doesn’t necessarily increase the amount of significant item sets; rather it always makes those item sets containing high weight items more likely to have a higher weighted support, hence holding more chances to become significant item sets containing no high weight items become relatively less likely to be significant.

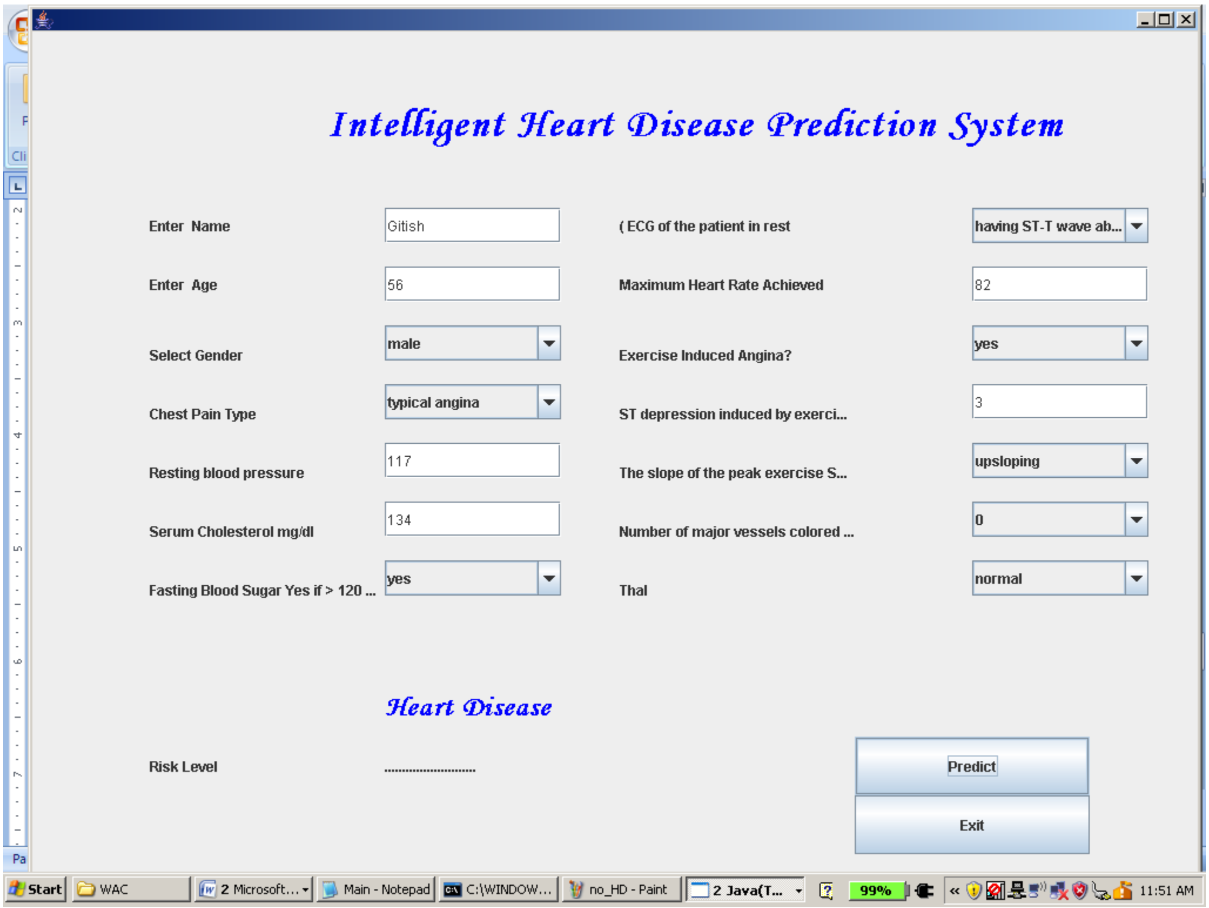
(3). We noticed that the association rule classifier is sensitive to the unbalanced data. The heart.D53.N303.C5.num dataset is having almost 40% of cases with no heart disease and remaining 60% is further divided in to 4 types of heart disease hence the data is found to be suitable for predicting “No Heart disease”. When the data set has been modified to incorporate only two class label one for “No Heart Disease” and others “Heart Disease” the accuracy is found to be **81.51%** for WAC. The maximum accuracy have been achieved using support value 25% and confidence to be 80%.

S. No.	Attribute name [Range of attribute value]		
1	Age [29.0, 77.0]	1	0.2
		2	0.3
		3	0.5
		4	0.7
		5	0.8
2	Sex (value 1: Male; value 0 : Female)	6	0.5
		7	0.4
3	Chest Pain Type [1.0, 4.0]	8	0.7
		9	0.7
		10	0.7
		11	0.7
4	Trestbps(Blood Pressure)[0.0, 1.0]	12	0.2
		13	0.3
		14	0.5
		15	0.7
		16	0.9
5	Chol (Serum Cholesterol) [126.0, 564.0]	17	0.3
		18	0.5
		19	0.7
		20	0.8
		21	0.9
6	Trestbps (Fasting Blood Sugar) [94.0, 200.0] Yes if > 120 mg/dl	22	0.8
		23	0.4
7	Restecg (resting electrographic results) [0.0, 2.0]	24	0.2
		25	0.9
		26	0.9
8	Thalach(maximum heart rate achieved) [71.0, 202.0]	27	0.3
		28	0.5
		29	0.7
		30	0.8
		31	0.9
9	Exang(exercise induced angina) real [0.0, 1.0]	32	0.5
		33	0.5
10	Oldpeak (ST depression induced by exercise relative to rest) [0.0, 6.2]	34	0.5
		35	0.5
		36	0.5
		37	0.5
		38	0.5
11	Slope(the slope of the peak exercise ST segment) [1.0, 3.0]	39	0.5
		40	0.5
		41	0.5
12	Ca (number of major vessels colored by floursopy) [0.0, 3.0]	42	0.5
		43	0.5
		44	0.5
		45	0.5
13	Thal real [3.0, 7.0]	46	0.3
		47	0.7
		48	0.8
14	Output(Class label representing four type of Heart Disease) Num {0,1,2,3,4}	49, 50, 51, 52, 53	

TABLE I. NORMALIZED HEART.D53.N303.C5 DATA SET WITH ATTRIBUTE WEIGHT.

S. No.	Data Set	WAC	CBA	CMAR	CPAR
1	heart	57.75	58.28	53.64	52.32
2	hepatitis	79.35	40.26	77.92	59.74
3	Cancer	90.41	93.7	88.82	92.84
Average Accuracy		**75.84**	64.08	73.46	68.3

TABLE II. RFORMANCE OF WAC, CBA, CMAR AND CPAR.



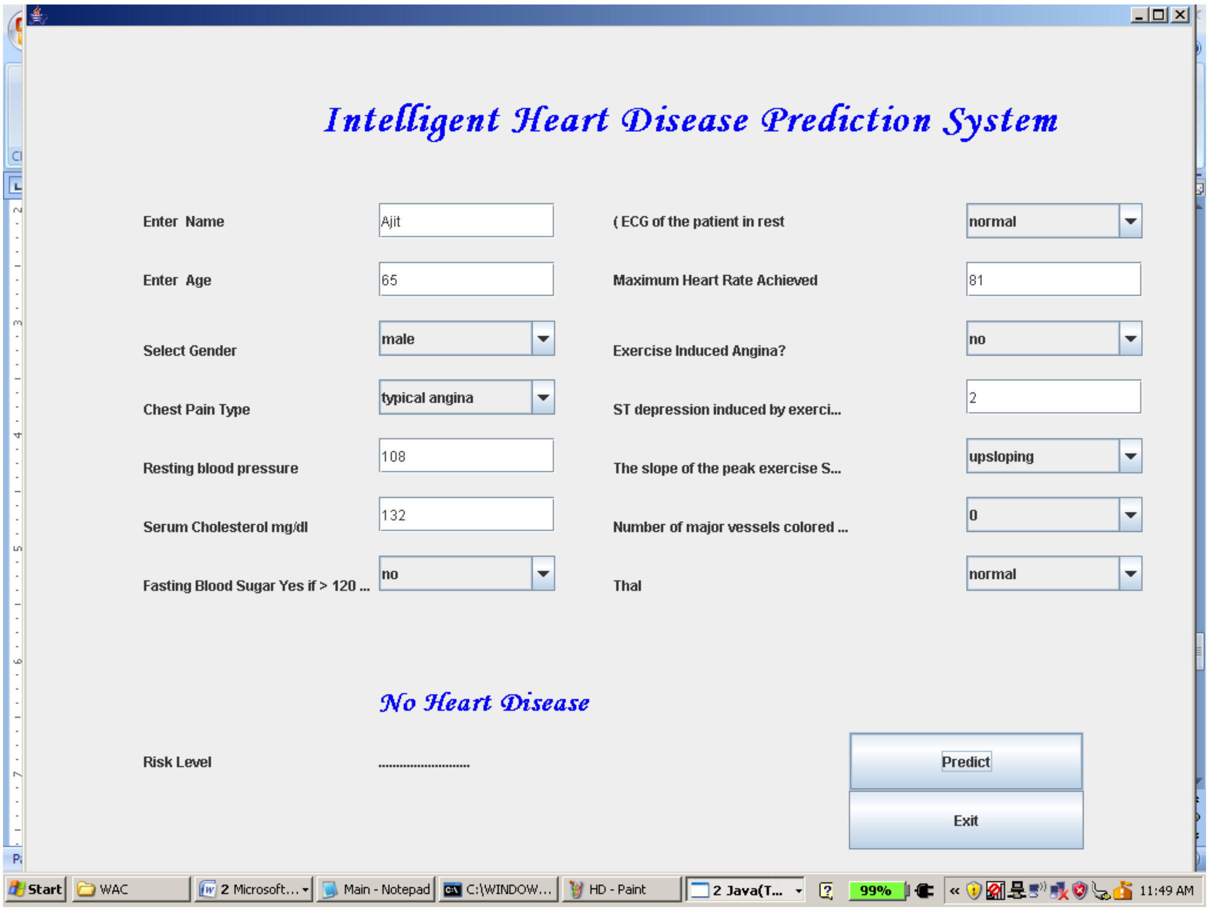
Intelligent Heart Disease Prediction System

Enter Name	Gitish	(ECG of the patient in rest	having ST-T wave ab...
Enter Age	56	Maximum Heart Rate Achieved	82
Select Gender	male	Exercise Induced Angina?	yes
Chest Pain Type	typical angina	ST depression induced by exerci...	3
Resting blood pressure	117	The slope of the peak exercise S...	upsloping
Serum Cholesterol mg/dl	134	Number of major vessels colored ...	0
Fasting Blood Sugar Yes if > 120 ...	yes	Thal	normal

Heart Disease

Risk Level ....

Figure 1. GUI for Heart Disease Prediction System.



Intelligent Heart Disease Prediction System

Enter Name	Ajit	(ECG of the patient in rest	normal
Enter Age	65	Maximum Heart Rate Achieved	81
Select Gender	male	Exercise Induced Angina?	no
Chest Pain Type	typical angina	ST depression induced by exerci...	2
Resting blood pressure	108	The slope of the peak exercise S...	upsloping
Serum Cholesterol mg/dl	132	Number of major vessels colored ...	0
Fasting Blood Sugar Yes if > 120 ...	no	Thal	normal

No Heart Disease

Risk Level ....

Figure 2. GUI for No Heart Disease Prediction System.

VIII. BENEFITS AND LIMITATIONS.

The proposed system is a Web-based, user-friendly, scalable and reliable that can be implemented in remote areas like rural regions or country sides, to imitate like human diagnostic expertise for treatment of heart ailment. The system is expandable in the sense that more number of records or attributed can be incorporated and new significant rules can be generated using underlying Data Mining technique. Presently the system has been using 13 attributes and 303 records only and the data is from UCI machine learning dataset which is mainly used for research purpose. As the symptoms that cause a particular disease may vary from region to region, the system should be trained using local dataset collected from the clinic.

IX. CONCLUSIONS

In this paper, we have presented an intelligent and effective heart attack prediction system using Weighted Associative Classifier. Firstly, we have evaluated the performance of WAC in terms of accuracy using benchmark (UCI machine learning repository) dataset available in http://csc.liv.ac.uk/~frans/KDD/Software/LUCS-KDD-DN/DataSets/dataSets.html. Different weights have been assigned to the attributes after consulting with expert doctor. Experimental results reveal that WAC is an efficient approach for the extraction of significant patterns from the heart disease dataset. These patterns are stored in rule base in the form of Prediction rules. A little modification has been incorporated in the database and instead of considering 5 class label (4 for four types of Heart Disease and 1 for no heart Disease) we have considered only 2 class labels 1 for “Heart Disease” and another for “No Heart Disease” as the data set is having less number of records for different types of Heart Disease. The maximum accuracy (81.51%) have been achieved using support value 25% and confidence to be 80% . A GUI has been designed to enter the patient’s records and the presence of Heart disease for the patient is predicted using the rules stored in the rule base.

REFERENCES

[1] Asha Rajkumar, G.Sophia Reena, *Diagnosis Of Heart Disease Using Datamining Algorithm,* Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver. 1.0 September 2010.

[2] Sunita Soni, O.P.Vyas, *Using Associative Classifiers for Predictive Analysis in Health Care Data Mining,* International Journal of Computer Application (IJCA, 0975 – 8887) Volume 4– No.5, July 2010, pages 33-34.

[3] K.SRINIVAS, B.KAVIHTA RANI , A.GOVRDHAN , *APPLICATIONS OF DATA MINING TECHNIQUES IN HEALTHCARE AND PREDICTION OF HEART ATTACKS*, (IJCSE) INTERNATIONAL JOURNAL ON COMPUTER SCIENCE AND ENGINEERING VOL. 02, NO. 02, 2010, 250-255.

[4] M. ANBARASI, E. ANUPRIYA, N.CH.S.N.IYENGAR, *Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm,* International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370-5376

[5] N.A. Setiawan, P.A. Venkatachalam, and Ahmad Fadzil M.H. , *Rule Selection for Coronary Artery Disease Diagnosis Based on Rough Set,* International Journal of Recent Trends in Engineering, Vol 2, No. 5, November 2009.

[6] Sunita Soni , Jyothi Pillai, O.P.Vyas, *An Associative Classifier Using Weighted Association Rule* , IEEE proceedings of the World Congress on Nature and Biologically Inspired Computing (NaBIC'09), December 09-11, 2009, 1492-1496.

[7] Shantakumar B.Patil, Y.S.Kumaraswamy, *Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network*, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656

[8] Ruben D. Canlas Jr.,DATA MINING IN HEALTHCARE: CURRENT APPLICATIONS AND ISSUES, August 2009.

[9] Sellappan Palaniappan Rafiah Awang, *Intelligent Heart Disease Prediction System Using Data Mining Techniques*, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008

[10] Fadi Thabtah, *A review of associative classification mining,* **The Knowledge Engineering Review,** Volume 22 , Issue 1 (March 2007),Pages 37-65, 2007.

[11] Carloz Ordonez, *Association Rule Discovery with Train and Test approach for heart disease prediction*, IEEE Transactions on Information Technology in Biomedicine, Volume 10, No. 2, April 2006.pp 334-343.

[12] Harleen Kaur , Siri Krishan Wasan and Vasudha Bhatnagar, *THE IMPACT OF DATA MINING TECHNIQUES ON MEDICAL DIAGNOSTICS*, Data Science Journal, Volume 5, 19 October 2006 pp119-126.

[13] Yin, X., Han, J. *CPAR: Classification based on predictive association rule.* In Proceedings of the SIAM International Conference on Data Mining. San Francisco, CA: SIAM Press, 2003, pp. 369-376.

[14] Feng Tao, Fionn Murtagh, Mohsen Farid. *Weighted Association Rule Mining using Weighted Support and Significance Framework,* Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining 2003, Pages:661-666 Year of Publication: 2003

[15] W.Li, J. Han, J.Pei , *CMAR- Classification based on Multiple Association Rules*, ICDM'01, , San Jose, CA, Nov. 2001. pp. 369-376

[16] W. Wang, J. Yang and P. Yu. *Efficient mining of weighted association rules (WAR)*, Proc. of the ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 270-274, 2000.

[17] Liu,B, Hsu. W. Ma, *Integrating Classification and association rule mining* . Proceeding of the KDD, 1998(CBA) pp 80-86.

[18] Magnus Stensmo, Terrence J. Sejnowski *Automated Medical Diagnosis based on Decision Theory and Learning from Cases,* World Congress on Neural Networks 1996 International Neural Network society pp. 1227-1 231.

[19] R. Agrawal and R. Srikant. *Fast algorithms for mining association rules*. In VLDB'94, , Santiago, Chile, Sept. 1994. pp. 487-49