

A Novel Frequent Features Prediction Model for Heart Disease Diagnosis

Atul Kumar Pandey*

Assistant Professor
Department of Physics
Govt. PG Science College
Rewa (M.P.)-India*

Prabhat Pandey**

OSD
Additional Directorate
Higher Education, Division
Rewa (M.P.)-India**

K.L. Jaiswal***

Assistant Professor
Department of Physics
Govt. PG Science College
Rewa (M.P.)-India***

Ashish Kumar Sen****

Research Scholar
Department of Math/Computer Science
Govt. PG Science College
Rewa (M.P.)-India*

Abstract- Heart disease is the leading cause of death in the world over the past 10 years. Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease patients. Decision Tree is one of the data mining techniques used in the diagnosis of heart disease showing considerable success. It is essential to find the best fit classification algorithm that has greater accuracy on classification in the case of heart disease prediction. Since the data is huge attribute selection method used for reducing the dataset. Then the reduced data is given to the classification. We also propose a novel feature selection method algorithm which is the AttributeSelectedClassifier method combining CFS subset evaluator and BestFirst method followed by J48 Decision tree then integrating the repetitive Maximal Frequent Pattern. The proposed algorithm provides better accuracy compared to the traditional algorithm and the hybrid Algorithm CFS. This research paper proposed a Novel frequent feature selection method for Heart Disease Prediction. Good performance of this method comes from the use of the Repetitive Maximal Frequent Pattern Method and the nonadditivity of the method against different target nominal attributes measure reflects the importance of the feature attributes as well as their interactions.

Keywords: Data mining, Maximal Frequent Pattern and Decision Tree.

1. INTRODUCTION

The term –cardiovascular disease includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular disease (CVD) [2, 3, 4] results in severe illness, disability, and death. A sudden blockage of a coronary artery, generally due to a blood clot results in a heart attack [3]. Chest pains arise when the blood received by the heart muscles is inadequate. High blood pressure, coronary artery disease, valvular heart disease, stroke, or rheumatic fever/rheumatic heart disease are the various forms of cardiovascular disease. Life itself is completely dependent on the efficient operation of the heart.

Knowledge discovery in databases is well-defined process consisting of several distinct steps [5]. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of Knowledge discovery

in databases is given as follows: “Data mining is the nontrivial extraction of implicit previously unknown and potentially useful information about data”. Data mining technology [5] provides a user-oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. Several data mining techniques are used in the diagnosis of heart disease such as naïve bayes, decision tree, neural network, kernel density, bagging algorithm, and support vector machine showing different levels of accuracies [10-16].

Decision Tree is one of the successful data mining techniques used in the diagnosis of heart disease patients [11, 14, 17]. Researchers have been applying different data mining techniques over different heart disease datasets to help health care professionals in the diagnosis of heart disease [10-11, 14-17]. The results of the different data mining research cannot be compared because they have used different datasets.

Decision tree is one of the data mining techniques showing considerable success compared to other data mining techniques over different heart disease datasets [11, 13-14, 17]. Applying decision tree in diagnosing heart disease patients showed different accuracies on different datasets that ranged between 60.4% and 94.93% [14, 18]. Tu et al. applied decision tree classifier on the Cleveland heart disease dataset showing accuracy of 78.9% [17].

Recently researchers are investigating enhancing decision tree performance in classification problems. Anbarasi et al. investigated enhancing decision tree performance through integrating genetic algorithm as a feature subset selection method in the diagnosis of heart disease patients [19]. This paper investigates enhancing decision tree performance in the diagnosis of heart disease patients through the integration of Repetitive Maximal Frequent Pattern Method.

2. DECISION TREE

The decision tree type used in this research is the gain ratio decision tree. The gain ratio decision tree is based on the entropy (information gain) approach, which selects the splitting attribute that minimizes the value of entropy, thus maximizing the information gain [20]. To identify the splitting attribute of the decision tree, one must calculate the information gain for each attribute and then select the attribute that maximizes the information gain. The information gain for each attribute is calculated using the following formula [9, 20]:

$$E = \sum_{i=1}^k P_i \log_2 P_i$$

Where k is the number of classes of the target attributes P_i is the number of occurrences of class i divided by the total number of instances (i.e. the probability of i occurring). To reduce the effect of bias resulting from the use of information gain, a variant known as gain ratio was introduced by the Australian academic Ross Quinlan [20]. The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values [8]. Gain Ratio adjusts the information gain for each attribute to allow for the breadth and uniformity of the attribute values.

Gain Ratio = Information Gain/Split Information

Where the split information is a value based on the column sums of the frequency table [20]. After extracting the decision tree rules, reduced error pruning was used to prune the extracted decision rules. Reduced error pruning is one of the fastest pruning methods and known to produce both accurate and small decision rules [21]. Applying reduced error pruning

provides more compact decision rules and reduces the number of extracted rules.

To evaluate the performance of the proposed model the sensitivity, specificity, and accuracy are calculated. The sensitivity is the proportion of positive instances that are correctly classified as positive (e.g. the proportion of sick people that are classified as sick). The specificity is the proportion of negative instances that are correctly classified as negative (e.g. the proportion of healthy people that are classified as healthy). The accuracy is the proportion of instances that are correctly classified [20].

$$\text{Sensitivity} = \text{True Positive} / \text{Positive}$$

$$\text{Specificity} = \text{True Negative} / \text{Negative}$$

$$\text{Accuracy} = (\text{True Positive} + \text{True Negative}) / (\text{Positive} + \text{Negative})$$

3. FEATURE SELECTION

The main purpose of feature selection [6] is to reduce the number of features used in classification while maintaining acceptable classification accuracy. For example, the Sequential Forward Floating Selection (SFFS) algorithm [7] proposed by Pudil et al. was one of the commonly used algorithms. The main advantage of this method is that it produces a hierarchy of feature subsets with the best selection for each dimension. In our previous work, information gain is used to find the relevant features. Information gain [1] is the difference between the original information content and the amount of information needed. The features are ranked by the information gains, and then the top ranked features are chosen as the potential attributes used in the classifier.

Frequent Item set Mining (FIM) [7] is considered to be one of the elemental data mining problems that intends to discover groups of items or values or patterns that co-occur frequently in a dataset. It is of vital significance in a variety of Data Mining tasks that aim to mine interesting patterns from databases, like association rules, correlations, sequences, episodes, classifiers, clusters and the like. Numerous algorithms like the Apriori and FP-Tree have been proposed to support the discovery of interesting patterns. The proposed approach utilizes an efficient algorithm called MAFIA [8] (Maximal Frequent Itemset Algorithm) which combines diverse old and new algorithmic ideas to form a practical algorithm. The cluster that contains data most relevant to heart attack is fed as input to MAFIA algorithm [7] to mine the frequent patterns present in it. After mining the frequent patterns using MAFIA algorithm, the significance weightage of each pattern is calculated. It is calculated based on the weightage of each attribute

present in the pattern and the frequency of each pattern [8].

4. PROPOSED METHOD

A. ASC and Maximal Frequent Pattern

We proposed a new novel Attribute selection method by combining ASC (AttributeSelectedClassifier) and MFP (Maximal Frequent Pattern) techniques. The ASC algorithm reduces the number of attributes based on the SU measure, In ASC each attributes are compared pair wise to find the Similarity and the Attributes are compared to class attribute to find the amount of contribution it provides to the class value , based on these the attributes are removed. The selected attributes from the ASC algorithm has been applied in the Maximal Frequent Pattern technique for further reduction with the desired minimum support value.

B. Dataset used in the Experiment

The following is the sample of the Heart Disease Data.arff @relation heart-statlog

```
@attribute age real
@attribute sex real
@attribute chest real
@attribute resting_blood_pressure real
@attribute serum_cholesterol real
@attribute fasting_blood_sugar real
@attribute resting_electrocardiographic_results real
@attribute maximum_heart_rate_achieved real
@attribute exercise_induced_angina real
@attribute oldpeak real
@attribute slope real
@attribute number_of_major_vessels real
@attribute thal real
@attribute class {absent, present}
```

@data

```
70,1,4,130,322,0,2,109,0,2,4,2,3,3,present
67,0,3,115,564,0,2,160,0,1,6,2,0,7,absent
57,1,2,124,261,0,0,141,0,0,3,1,0,7,present
```

The Heart Disease data after applying traditional method in Weka, The number high number of attributes reduced is 7 against the nominal class attributes and then these attributes can be fed to various classifiers. The ASC+ MFP algorithm is coded, where the attribute after **ASC** is 7 and the selected attributes after removing the numeric attribute is only 6. **ASC** Feature selection method which selects the attributes based on the symmetrical uncertainty reduces the number of attributes from 14 to 6. The reduced attributes is fed to ASC again followed by Maximal Frequent pattern method for further reduction.

The heart ARFF will contain large quantity of data and applying classification algorithms to this dataset is time consuming and also gives result with less accuracy. Hence we have to reduce the data set by using attribute selection method. Likewise, all other attribute selection and classification algorithms are applied for heart disease dataset. From that we identified that ASC algorithm gives better accuracy after applying the Maximal Frequent Pattern method.

C. Feature Selector

The best Feature Selection methods ASC are applied in sequence with different nominal attribute. (i.e), in this method the reduced number of attributes using frequent pattern mining method is 6. We proposed a novel algorithm that is ASC+ Maximal Frequent Pattern. When applying this feature selection algorithm, the attributes are reduced as 3.

5. RESULT AND DISCUSSION

A. Attribute Selected Classifier

The feature selector method is automated, where the number of reduced attributes by

AttributeSelectedClassifier against nominal target attributes.

Table 5.1: Reduced attributes by ASC against 8 Nominal Target Attributes

S.No.	Class Attribute	Reduced Attributes	Correctly Classified Instances	Incorrectly Classified Instances	Size of the Tree	Number of Leaves	Time(Seconds)	Accuracy
1	age	Numeric						
2	sex	5,13,14	232	71	17	11	0.00	76.5677
3	Cp	6,8,9,10,14	214	89	56	30	0.02	70.6271
4	trestbps	Numeric						
5	chol							
6	fbs	3,7,11,13	258	45	1	1	0.02	85.1485*
7	restecg	3,6,11,14	175	128	6	5	0.00	57.7558
8	thalach	Numeric						

9	exang	3,8,14	246	57	12	9	0.00	81.1881*
10	oldpeak	Numeric						
11	slope	6,8,10,14	236	67	30	17	0.02	77.8878
12	ca	Numeric						
13	thal	2,6,8,14	222	79	10	7	0.02	73.7542
14	num	3,7,8,9,10,12,13	268	35	36	22	0.03	88.4488*

Table 5.2: Reduced Attributes with their Frequency Count value 2

Reduced Attribute	After Removing Numeric Attribute
6,7,8,9,11,13,14	6,7,9,11,13,14

Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data. If the label is categorical (classification), accuracy is commonly

reported as the rate which a case will be labeled with the right category. If the label is continuous, accuracy is commonly reported as the average distance between the predicted label and the correct value.

Table 5.3: Detailed Accuracy By Class Attribute FBS ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0	0	0	0	0	0.5	t
	1	1	0.851	1	0.92	0.5	f
Weighted Avg.	0.851	0.851	0.725	0.851	0.783	0.5	

A confusion matrix displays the number of correct and incorrect Predictions made by the model compared with the actual classifications in the test data. The matrix is n -by- n , where n is the number of classes. From that we calculated the accuracy of each classification algorithms.

Table 5.4: == Confusion Matrix ==

a	b	<- classified as
0	45	a = t
0	258	b = f

Table 5.5: Again applying the AttributeSelectedClassifier with the Selected Nominal Attributes.

S.No.	Class Attribute	Reduced Attributes	Correctly Classified Instances	Incorrectly Classified Instances	Size of the Tree	Number of Leaves
6	fbs	7,11,13	258	45	1	1
7	restecg	6,11,14	175	128	6	5
9	exang	14	218	85	6	5
11	slope	7,9,13,14	203	100	12	9
13	thal	6,14	219	82	6	5
14	num	9,11,13	239	64	9	6

B. AttributeSelectedClassifier against the Reduced Selected Nominal Attributes.

We proposed a new Novel Feature selector combining ASC and Maximal Frequent Pattern.

Table 5.6: Novel feature selection

Attribute Selection methods	Selected attributes(Support=3)
ASC+ Maximal Frequent Pattern	3(11,13,14)

Then this reduced data is given to the classification algorithms and calculate the accuracy for identifying the best algorithm.

6. CONCLUSION

Heart disease is the leading cause of death all over the world. Researchers have been investigating applying different data mining techniques to help health care professionals in the diagnosis of heart disease patients. Decision Tree is one of the successful data mining techniques used in the diagnosis of heart disease patients. This paper proposed a Novel frequent feature selection method for Heart Disease Prediction. The novel feature selection method algorithm which is the AttributeSelectedClassifier method including CFS subset

evaluator and BestFirst search method followed by J48 decision tree then integrating the Repetitive Maximal Frequent Pattern Method was proposed. The proposed algorithm gives better accuracy comes from the use of the Repetitive Maximal Frequent Pattern Method.

Good performance of this method comes from the use of the repetitive Maximal Frequent Pattern. The nonadditivity of the method against different target nominal attributes measure reflects the importance of the feature attributes as well as their interactions. The proposed work can be further enhanced and expanded for the automation of Heart disease prediction. We intend to extend our work applying various classification methods to predict the heart disease more efficiently.

7. REFERENCES

- [1] Kwong-Sak Leung,kin hong Lee,Jin-Feng Wang,Eddie Y.T.Ng,Henry L.Y.Chan,Stephen K.W.Tsui,Tony S.K. Mok,Pete Chi-Hang Tse, Joseph Jao-yui Sung, Data Mining on DNA Sequences of Hepatitis B virus. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 8,No 2,March/April 2011
- [2] Sunita Soni, Jyoti Soni, Ujma Ansari, Dipesh Sharma, Predictive Data Mining for Medical Diagnosis:An Overview of Heart Disease Prediction. International Journal of Computer Application (IJCA, 0975 – 8887) Volume 17– No.8, March 2011.
- [3] Minas A. Karaolis, Member, IEEE, Joseph A. Moutiris, Demetra Hadjipanayi, and Constantinos S. Pattichis, Senior Member, IEEE, Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees. IEEE Transactions On Information Technology In Biomedicine, Vol. 14, No. 3, May 2010.
- [4] Milan Kumari and Sunila Godara, Comparative study of Data Mining Classification Methods in Cardiovascular Disease Prediction. IJCST Vol 2, Issue 2, June 2011.
- [10] Das, R., I. Turkoglu, and A. Sengur, Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications. Elsevier, 2009. 36 (2009): p. 7675–7680.
- [11] Andreeva, P., Data Modelling and Specific Rule Generation via Data Mining Techniques. International Conference on Computer Systems and Technologies - CompSysTech, 2006.
- [12] Hara, A. and T. Ichimura, Data Mining by Soft Computing Methods for the Coronary Heart Disease Database. Fourth International Workshop on Computational Intelligence & Applications, IEEE, 2008.
- [5] 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.
- [6] 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
- [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] 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.
- [9] Han, j. and M. Kamber, Data Mining Concepts and Techniques. 2006: Morgan Kaufmann Publishers. Lee, I.-N., S.-C. Liao, and M. Embrechts, Data.
- [13] Rajkumar, A. and G.S. Reena, Diagnosis of Heart Disease Using Data mining Algorithm. Global Journal of Computer Science and Technology, 2010. Vol. 10 (Issue 10).
- [14] Sitar-Taut, V.A., et al., Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics, 2009.
- [15] Srinivas, K., B.K. Rani, and A. Govrdhan, Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering (IJCSE), 2010. Vol. 02, No. 02: p. 250-255.

- [16] Yan, H., et al., Development of a decision support system for heart disease diagnosis using multilayer perceptron. Proceedings of the 2003 International Symposium on, 2003. vol.5: p. pp. V-709- V-712.
- [17] Tu, M.C., D. Shin, and D. Shin, Effective Diagnosis of Heart Disease through Bagging Approach. Biomedical Engineering and Informatics, IEEE, 2009.
- [18] Palaniappan, S. and R. Awang, Web-Based Heart Disease Decision Support System using Data Mining Classification Modeling Techniques. Proceedings of iiWAS, 2007.
- [19] Anbarasi, M., E. Anupriya, and N.C.S.N. Iyengar, Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology, 2010. Vol. 2(10).
- [20] Bramer, M., Principles of data mining. 2007: Springer.
- [21] Esposito, F., D. Malerba and G. Semeraro, A Comparative Analysis of method for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 997. Vol. 19, No. 5.