Heart Disease Severity Prediction

Mrs. Geetanjali Nilesh Sawant
Dept of Computer Engineering
Rajendra Mane College of Engineering and Technology,Ambav

Abstract- *Todays heart disease death rate compels to predict the severity that person is prone to heart disease through the identification and evaluation of different controllable and uncontrollable risk factors which are causing heart disease. Age like uncontrollable risk factor can not be treated in curing or controlling heart disease whereas cholesterol like factors when exceed their normal range; contributes to disease and required treatment to bring it to their normal level. Controllable factors might dependent or independent as well as their affection towards disease may also different. If such correlation among them are found and analyzed over time, then definitely it will help in early and accurate diagnosis of disease. This may lead to time and cost effective treatments as well as assurance of speedy recovery of patients. Many data mining techniques serve this purpose. This paper proposes hybrid approach of predicting heart disease severity by using sequential combination of association rule mining and decision tree. Hidden relevance among factors is drawn by applying association rule mining and keeping relevant factors at root level further levels of tree are constructed. Leaf node labels 'High', 'Moderate' and 'Low' of resulting classifier-tree imply severity of heart disease.*

Keywords- Attribute_tagging, n-factor decision tree, confusion matrix, accuracy.

I. INTRODUCTION

Most of the people are ignorant of their health. It is observed that these people approach clinic when they find known symptoms of certain disease or when they get caught by disease due to which they lose their work efficiency. In minor diseases, treatments work out to cure it; but in severe diseases like cancer, heart attack, etc., such ignorance may cost the life of a patient. General worldwide survey tells Heart diseases death rate is 31% of all global deaths. Decision support system is needed to track all parameters (which are if abnormal or if reach to incurable level may attack heart) and analyze these to predict the severity in early stage or extent by which these are to be cared to avoid heart disease. This motto can be achieved with the proposed system. This predictor first identifies association among different heart disease causing risk factors, once association is evaluated for its trueness, it guides decision tree building. Leaf nodes of tree indicate class labels, "High','Moderate'or 'Low' are the ones used in this study.

II. LITERATURE SURVEY

Clinical decisions are often made based on doctor's intuition and experience[1]. Mostly different patients undergo the same treatment on diagnosis of same disease even though they are different in their physical statistics which may cause any side disease. Decisions are expected to be based on patterns hidden in the dataset storing risk factors. Data mining techniques are rich to drag out patterns, the health experts are interested in.

In order to apply any data mining technique training dataset is needed. Cleveland dataset from Cleveland Clinic Foundation is open for research work. This dataset contains many factors which can be grouped as Controllable, Uncontrollable[1].

Uncontrollable factors are age, gender, family medical background of patient. Controllable factors are smoking, blood pressure, etc[2]. These can be further classified as bad habits representing factors and physical statistics representing factors. Smoking, drinking addictions, poor diet, etc. are bad habits representing factors whereas physical statistics expressing factors are resting blood pressure, cholesterol, fasting blood sugar, maximum heart rate achieved, etc.

"Intelligent Heart Disease Prediction System Using Data Mining Techniques" [1]paper informs evaluation of Decision trees, Naive Bayes and Neural network independently towards making correct prediction and found Neural network is efficient among all. "Early Prediction of Heart Diseases Using Data Mining Techniques"[3] paper presented analysis through different decision tree algorithms-CART(Classification and Regression Trees), ID3(Iterative Dichotomiser), decision tree and concluded with CART as a more accurate in comparison of other whereas ID3 is time efficient. "Prediction of Heart Disease using Data Mining Techniques"[2] paper proposed the hybrid approach to attain more accuracy with which Decision tree and Bayesian classification are used to intimate about the possibility of heart disease.

III. PROPOSED SYSTEM

Even though decision tree takes into account the relevance among attributes, there is no any way to check the reliability of obtained order of attribute relevance. This issue is addressed by initiating the prediction process with association rule mining technique as shown in fig.1. Factors like age, gender, chest pain type, resting blood pressure, cholesterol, resting electrographic results, fasting blood sugar, maximum heart rate achieved, exercise induced agina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, no. of major vessels colored by floursopy, defect type, obesity, smoking with their discrete values are used to build Association Rule Mining model.

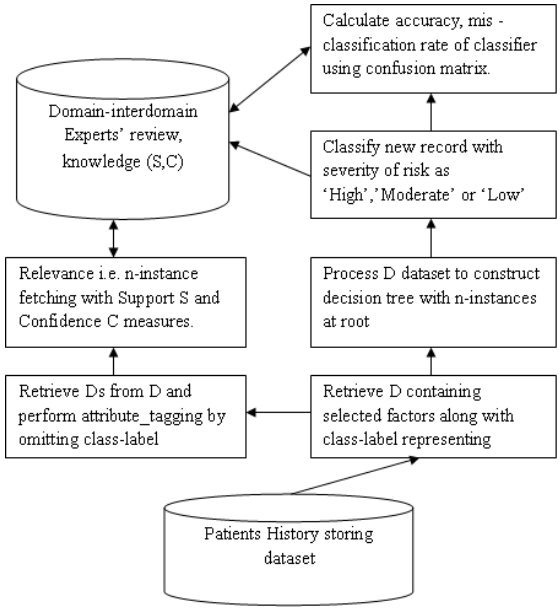

Fig.1 Architecture of Heart Disease Severity Predictor

Association rule mining is a technique to identify the relevance among the factors cause to heart disease. Currently available association rule mining is capable to work on the presence and absence of an instance in the tuple of given training dataset. In case of heart disease prediction the factors are to be used are representing physical statistics of a patient. So these factors' sure existence is recorded in dataset to be worked on. Taking absence and presence of instance into consideration is not suitable for determining relevance among the different factors responsible for heart disease and due to which existed factors are need to be expressed in innovative way. To achieve the same, we have adopted initiating two-step attribute_tagging stage. As all the factors involved in prediction are continuous valued, these are discretized with the help of domain experts. e.g. Age is categorized as 'youth' involving 0 to 25 aged, 'middle_aged' involving 26 to 45 aged and 'senior' for 46 above encompassing. Blood pressure is categorized as 'high' if valued above 120/80 mm, 'low' if valued below 90/60 mm and 'normal' otherwise. Likely all the factors are discretized to accomplish first step of attribute_tagging. With second step, attribute name and its value are concatenated, e.g. Age with three different values Youth, Middle_aged and Senior give rise to three different subfactors on concatenating it with its values which are Age_Youth, Age_Middle_aged and Age_Senior. All factors are processed in same manner to generate the final instances as shown in fig.2. These instances' presence and absence mentioned from original dataset to new table, that the association rule mining can be applied on.

Record from Ds	Age	Blood Pressure	Weight
	45	123/85	70
1.Discretization	Middle_aged	Normal	Normal
2.Concatenation of attribute and its value	Age_middleaged	Blood_pressure_normal	Weight_normal

Fig.2 Attribute Tagging

Association finding emerge candidate set for every possible combination of instances, its support can be determined by counting the occurrence/s of earlier obtained instances. By scanning dataset, Candidate-k for k=1,2,3,... instance sets can be generated. Frequent-k instance sets can be drawn out satisfying support threshold s. Support threshold, s and Confidence threshold,c satisfying instances are considered as reliable. For actual purpose of predicting severity, relevant data get fed to decision tree. Attribute relevance is taken into consideration by decision tree with which highest risk factor/attribute forms root of tree whereas all other attributes form intermediate nodes following root node according to their risk. With root node and every intermediate node, data get splited. This procedure is repeated till every record from dataset,D get classified as "prone to heart disease with risk 'High', 'Moderate' or 'Low' ". Leaf node denotes class label. Root node and internal nodes represent attributes whereas branches emerging out from node represent attribute values. 'n'-instances obtained from association rule technique application act as a root and forms 0^{th} level of decision tree. Tree with n-instances root is addressed as n-factor tree. This 'n'-instance is combination of attribute and its value which being a root splits dataset on basis of their values. For the obtained subset of dataset, as per attribute relevance decision tree is constructed. Decision tree will help in predicting severity of heart disease as high, moderate and low. Tree is optimized by pruning it.

Confusion matrix is a table in matrix form used as a performance describer of a classification model. Mentioned application resulting in three classes 'High','Moderate' and 'no'. Built decision tree is referred as classifier classifies the

tuple from training dataset itself either 'High', 'Moderate' or 'Low' .

		Predicted Class			
		High	Moderate	Low	Total Records
Actual Class	High	HH	HM	HL	P
	Moderate	MH	MM	ML	Q
	Low	LH	LM	LL	R

Fig.3 Confusion matrix

HH, MM, LL are no.of records correctly predicted as that of actual High, Moderate and Low class respectively.

HM,HL,MH,ML,LH,LM are no of records misclassified by classifier.

Total no.of records are (P+Q+R).

Accuracy can be decided as follows:

Accuracy(n-factor tree) = (HH+MM+LL) / (P+Q+R) We can obtain misclassification rate i.e. how probable classifier misclassify, as follows:

Misclassification rate (1-factor DT)= (HM+HL+MH+ML+LH+LM)/(P+Q+R).

Accuracy and misclassification rate can be calculated for every n-factor tree and highest accuracy offering tree is selected as classifier for further practical use.

Algorithm :Heart disease severity prediction.
Input:

- D, a database containing heart disease responsible factors.
- S, the minimum support threshold.
- C, the minimum confidence threshold.

Output:

Heart disease severity prediction as High, Moderate and Low.
Method:

(1) Retrieve Ds by excluding heart-disease-severity denoting attribute from D.
(2) Generate instance by concatenating attribute_name separately with its every possible attribute_value.
(3) Apply Apriori algorithm using 'S' to get frequent k-instance set/s.
(4) Apply 'C' to frequent n-instance set/s resulted in step (2), to obtain relevant instances.
(5) Retrieve Df, n-instances belonging dataset along with severity denoting attribute is extracted from D.
(6) Apply decision tree algorithm to construct tree with level-1 onwards using factors not in n-instance by keeping n-instances at 0^{th} level.
(7) Use confusion matrix to calculate accuracy of the hybrid model.

Once it is identified, heart disease is probability of its severity can be calculated by feeding record to Bayesian model. Bayesian technique assumes every attribute is independent and predicts class label along with its probability

Challenges

Accuracy- the sensitive fragment of datamining's practical use in medical field, may get hamperedS due to certain issues:

1. All attributes are not discrete in nature. E.g. Age : it need to categorized in ranges –say 0-25, 26-45, 46 and above.
2. Training data set used for building classifier must be complete. If it contains measurable incompleteness then even after processing such data affects the based classifier.
1. 3.Prediction system indulges reasonable human interference

i.e. domain expert's view from discretization of attribute values to tree optimization. Relevance finding may be affected by changed support and confidence thresholds. This is again depending upon his/her intellectual level as well as experience.

IV. CONCLUSION

Heart disease contributing factors are so many. If their relevance is determined, then with allowable specificity, heart disease severity can be predicted and accordingly effective treatment can be offered to patient. This work is capable to inform severity only, with certain updation it may be applicable to non-patients for predicting infuture chances of disease based on current physical statistics. Introducing temporal support and confidence threshold, periodically these factors may be reviewed to obtain more accurate results.

REFERENCES

[1] Sellappan Palaniappan, Rafiah Awang-Intelligent Heart Disease Prediction System using Data Mining

Techniques, International Journal of Computer Science and Network Security, Vol.8,August 2008

[2] S.Kruthika Devi, S.Krishnapriya, Dristipona Kalita- Prediction of Heart Disease using Data Mining Techniques, Indian Journal of Science and Technology. Vol.9(39), October2016

[3] Vikas Chaurasia, Saurabh Pal-Early prediction of Heart Diseases Using Data Mining Techniques, Caribbean Journal of Science and Technology, vol.1, 2013

[4] Keerthana T.K. -Heart Disease Prediction System using Data Mining Method, International Journal of Engineering Trends and Technology, Vol.47(6), May 2007

[5] Chaitrali S. Dangare, Sulabha S. Apte - Improved study of Heart Disease Prediction System using Data Mining Classification Techniques, International Journal of Computer Applications, Vol.47, No.10,June 2012.

[6] Jiawei Han, Micheline Kamber,Jian Pei -Data Mining Concepts and Techniques, MK publication, 3rd ed.