RESEARCH ARTICLE OPEN ACCESS

An Efficient Heart Disease Prediction using Various Data Mining Techniques

A Krishna[1], Ch Narsimha Chary[2], M Rakesh Chowdary[3], Dr R P Singh [4]

[1,2,3](Research scholar, Sri Satya Sai University of Technology & Medical Sciences, Bhopal,Mp,India)
[4](Research supervisor, Sri Satya Sai University of Technology & Medical Sciences, Bhopal,Mp,India)

Abstract:

The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The Healthcare industry is generally "information rich", but unfortunately not all the data are mined which is required for discovering hidden patterns & effective decision making .Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining modeling techniques can help remedy this situation.

I INTRODUCTION

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases.

The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure – a condition that causes around half of all deaths from stroke and heart disease. Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. 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. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor's experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial. Our work attempts to present the detailed study about the different data mining techniques which can be deployed in these automated systems.

II RELATED WORK

This paper exhibits the analysis of various data mining techniques which can be helpful for medical analysts or practitioners for accurate heart disease diagnosis. Due to resource constraints and the nature of the paper itself, the main methodology used for this paper was through the survey of journals and publications in the fields of medicine, computer science and engineering.

RESEARCH FINDINGS

2.1 Data Mining in the Heart Disease Prediction.

Different supervised machine learning algorithms i.e. Naïve Bayes, Neural Network, along with weighted association Apriori algorithm, Decision algorithm have been used for analyzing the dataset in [1]. The data mining tool Weka 3.6.6 is used for experiment. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Decision Tree is a popular classifier which is simple and easy to implement. There is no requirement of domain knowledge or parameter setting and can high dimensional data can be handled. It produces results which are easier to read and interpret. The drill through feature to access detailed patients' profiles is only available in Decision Trees.

Naïve Bayes is a statistical classifier which assumes no dependency between attributes. This classifier algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes. The advantage of using naïve bayes is that one can work with the Naïve Bayes model without using any Bayesian methods.

2.2 Neural Networks

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural network. In other words, it is an emulation of biological neural system [9]. In feed-forward neural networks the neurons of the first layer forward their output to the neurons of the second layer, in a unidirectional fashion, which explains that the neurons are not received from the reverse direction. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input xi into neurons in hidden layer. Neuron of hidden layer adds input signal xi with weights wji of respective connections from input layer. The output Yj is function of

$$Yj = f\ (\Sigma\ wji\ xi)$$

where f is a simple threshold function such as sigmoid or hyperbolic tangent function.

III ASSOCIATIVE CLASSIFICATION

Associative classification mining is a promising approach in data mining that utilizes the association rule discovery techniques to construct classification systems, also known as associative classifiers. . Association rule mining is used to find associations or correlations among the item sets. It is an unsupervised learning where no class attribute is involved in finding the association rule. On the other hand, classification is a supervised learning where class attribute is involved in the construction of the classifier and is used to classify or predict the data unknown sample. Associative classification is a recent and rewarding technique which integrates association rule mining and classification to a model for prediction and achieves maximum accuracy. Associative classifiers are especially fit to applications where maximum accuracy is desired to a model for prediction. Various Techniques which can be used are Apriori algorithm, éclat algorithm, FP-growth algorithm. Here we will be using Apriori Algorithm for discovering interesting relations in heart based diseases.

Apriori Algorithm:

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. A frequent item set can be a defined as a subset of frequent item set i.e., if {PQ} is a frequent item set, both {P} and {Q} should be a frequent item set.

1. Iteratively discover frequent item sets with cardinality from 1 to k (k-item set).

2. Use the frequent item sets produce association rules. Join Step: Ck is generated by joining Lk-1with itself Prune Step: Any (k-1)-item set that is not frequent cannot be a subset of a frequent k-item set Initialize: K: = 1, C1 = all the 1- item sets; read the database to count the support of C1 to determine L1. L1:= {frequent 1- item sets};

K: =2;

//represents the pass number// While (Lk-1 ≠) do

Begin
Ck: = gen_candidate_itemsets with the given Lk-1 Prune (Ck) for all candidates in Ck do

count the number of transactions of at least k length that are common in each item Ck Lk := All candidates in Ck with minimum support;

```
k := k + 1; end
```

IV PROPOSAL SYSTEM

Frequent Pattern mining using MAFIA

Mining frequent item sets is an active area in data mining that aims at searching interesting relationships between items in databases[11]. It can be used to address to a wide variety of problems such as discovering association rules, sequential patterns, correlations and much more. The proposed approach utilizes an efficient algorithm called MAFIA (Maximal Frequent Item set Algorithm) which combines diverse old and new algorithmic ideas to form a practical algorithm. The proposed algorithm is employed for the extraction of association rules from the clustered dataset besides performing efficiently when the database consists of very long item sets specifically.

Pseudo code for MAFIA :

```
MAFIA(C, MFI, Boolean IsHUT)
{
name HUT = C.head C.tail;
if HUT is in MFI
stop generation of children and return
Count all children, use PEP to trim the tail, and recorder by increasing support,

For each item i in C, trimmed_tail
{

IsHUT = whether i is the first item in the tail newNode = C I MAFIA (newNode, MFI, IsHUT) }

if (IsHUT and all extensions are frequent)
Stop search and go back up subtree
If (C is a leaf and C.head is not in MFI)
Add C.head to MFI
}
```

The cluster that contains data most relevant to heart attack is fed as input to MAFIA algorithm to mine the frequent patterns present in it. Then the significance weightage of each pattern is calculated using the approach described in the following subsection.

Significance Weightage Calculation

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.

$$S_{wi} = \sum_{i=1}^{n} W_i f_i$$

Where Wi represents the weightage of each attribute and fi denotes the frequency of each rule. Subsequently the patterns having significant weightage greater than a predefined threshold are chosen to aid the prediction of heart attack

$$SFP = \{x : Sw\,(x) = \Phi\}$$

Where SFP represents significant frequent patterns and Φ represents the significant weightage.
This SFP can be used in the design of heart attack prediction system.

CONCLUSION

In this paper the focus is on using different algorithms and combinations of several target attributes for effective heart attack prediction using data mining. Decision Tree has outperformed with 99.62% accuracy by using 15 attributes. Also the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.Association classification technique apriori algorithm, was along with a new algorithm MAFIA was used .Straight Apriori-based algorithms count all of the 2k subsets of each k-item set they discover, and thus do not scale for long item sets. They use "look a heads" to reduce the number of item sets to be counted. MAFIA is an improvement when the item sets in the database are very long.

REFERENCES

[1] P .K. Anooj, Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules ; Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40.
[2] Nidhi Bhatla, Kiran Jyoti"An Analysis of Heart Disease Prediction using Different Data Mining Techniques".International Journal of Engineering Research & Technology

[3] Jyoti Soni Ujma Ansari Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction".

[4] Chaitrali S. Dangare Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888)

[5] Dane Bertram, Amy Voida, Saul Greenberg, Robert Walker, "Communication, Collaboration, and Bugs: The Social Nature of Issue Tracking in Small, Collocated Teams".

[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.

[7] Erickson, T., Smith, D., Kellogg, W., Laff, M., Richards, J and Bradner, E. Socially translucent conversations: Social proxies, persistent conversation, and the design of "Babble."Proc. ACM CHI (1999), 72–79.

[8] Hollan, J., Hutchins, E. and Kirsh, D. Distributed cognition: Toward a new foundation for human computer interaction research. ACM TOCHI, 7(2),(2000), 174–

[9] 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.

[10] Statlog database: http://archive.ics.uci.edu/ml/machinelearning- databases/statlog/heart

[11] Shantakumar B.Patil,Dr.Y.S.Kumaraswamy "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.2, February 2009