ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st&22ndMarch, Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Heart Disease Diagnosis Using Predictive Data mining

B.Venkatalakshmi, M.V Shivsankar

TIFAC-CORE, Pervasive Computing Technologies, Velammal Engineering College, Chennai, India

TIFAC-CORE, Pervasive Computing Technologies, Velammal Engineering College, Chennai, India

Abstract— Heart disease is a major health problem and it affects a large number of people. Cardiovascular Disease (CVD) is one such threat. Unless detected and treated at an early stage it will lead to illness and causes death. There is no adequate research focus on effective analysis tools to discover relationships and trends in data especially in the medical sector. Health care industry today generates large amounts of complex clinical data about patients and other hospital resources. Data mining techniques are used to analyze this rich collection of data from different perspectives and deriving useful information. This project intends to design and develop diagnosis and prediction system for heart diseases based on predictive mining. Number of experiments has been conducted to compare the performance of various predictive data mining techniques including Decision tree and Naïve Bayes algorithms. In this proposed work, a 13 attribute structured clinical database from UCI Machine Learning Repository has been used as a source data. Decision tree and Naive Bayes have been applied and their performance on diagnosis has been compared. Naive Bayes outperforms when compared to Decision tree.

Keywords—Predictive data mining,Naïve Bayes,Decision Tree.

I. INTRODUCTION

Medical Informatics is the applied science at the junction of the disciplines of medicine and information technology, which provides measurable improvements in both quality of care and effectiveness. Information technologies are playing a crucial role in advancing the science of quality measurement but more can be done to apply it to quality improvement. The Health care provides various services which are used to: (1) improve quality and efficiency; (2) engage patients and families; improve care coordination, and population and public health; and (3) Maintain privacy and security of patient health information. The most predominant health issue is heart failure which occurs especially in old patients because of diet, non-steroidal anti-inflammatory drugs and will leads even towards death. One of the commonly occurred heart diseases is Cardio vascular disease. Thus it is highly essential to predict such diseases through suitable symptoms.There are various types of algorithms which are present for the prediction of heart diseases which are Decision Trees, Naïve Bayes etc.

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 for bringing the efficient and accurate result. Appropriate computer-based information and decision support systems can aid in achieving clinical tests at a reduced cost.

In this work a performance comparison of heart disease diagnosis is executed with the help of Decision tree and Naïve Bayes. The rest of the paper has been organized as follows. Section 2 reviews some of the related works of the proposed solution. Section 3 elaborates various algorithms which are used for diagnosing the heart disease. Section 4 defines the simulation technique called Weka 3.7.9.

II. RELATED WORKS

Many experiments are being carried out for evaluating the performance of Naïve Bayes and Decision Tree algorithm. The results observed so far indicate that Naïve Bayes outperforms and sometimes Decision Tree. In addition to that an optimization process using genetic algorithm is also being planned in order to reduce the number of attributes without sacrificing accuracy and efficiency for diagnosing the heart disease.

There are many possible algorithms for the diagnosis of heart disease which are:

A. Naïve Bayes

A Naive Bayes classifier predicts that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [8].

This classifier is very simple, efficient and is having a good performance. Sometimes it often outperforms more sophisticated classifiers even when the assumption of independent predictors is far. This advantage is especially pronounced when the number of predictors is very large. One of the most important disadvantages of Naive Bayes is that it has strong feature independence assumptions.

B. Decision Trees

Decision Trees (DTs) are a non-parametric supervised learning method used for classification. The main aim is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The structure of decision tree is in the form of a tree. Decision trees classify instances by starting at the root of the tree and moving through it until a leaf node. Decision trees are commonly used in operations research, mainly in decision analysis. Some of the advantages are they can be easily understand and interpret, robust, perform well with large datasets, able to handle both numerical and categorical data. Decision-tree learners can create over-complex trees that do not generalise well from the training data is one the limitation.

C. Clustering

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a data set. Clustering is an unsupervised classification and has no predefined classes.

They are used either as a stand-alone tool to get insight into data distribution or as a pre-processing step for other algorithms. Moreover, they are used for data compression, outlier detection, understand human concept formation. Some of the applications are Image processing, spatial data analysis and pattern recognition. Classification via Clustering is not performing well when compared to other two algorithms.

All these algorithms are implemented with the help of WEKA tool for the diagnosis of heart diseases. Data set of 294 records with 13 attributes. These algorithms have been used for analyzing the heart disease dataset. The Classification Accuracy should be compared for this algorithm. After the comparison attributes are to be reduced for further purpose.

III. PRINCIPLES OF PREDICTIVE DATA MINING

There are many principles which are used for predicting the heart disease.

A. Bayes theorem

Bayes rule is used in naive bayes algorithm for the manipulation of conditional probabilities. Bayes' theorem gives the relationship between the probabilities of *A* and *B*, *P*(*A*) and *P*(*B*), and the conditional probabilities of *A* given *B* and *B* given *A*, *P*(*A*|*B*) and *P*(*B*|*A*).

$$P(A|B)= P(A\cap B)/P(B) \quad (1)$$

B. Entropy

Entropy is one of the principles which is used in decision tree and is to measure the amount of information in an attribute and also the impurity.

The general formula is:

$$\text{Entropy}(S) = \text{Entropy}(S)=\sum_{i=1}^{c} (-p(I)\log_2 p(I)) \quad (2)$$

IV. PARAMETERS OF PDM

Some of the parameters [4] which are used for Predictive data mining are

A. Sensitivity

It is also known as True Positive Rate. It is used for measuring the percentage of sick people from the dataset.

Sensitivity = Number of true positives/Number of true positives + Number of false negatives (3)

B. Specificity

It is also known as True Negative Rate. It is used for measuring the percentage of healthy people who are correctly identified from the dataset.

Specificity = Number of true negatives/Number of true negatives + Number of false positives (4)

C. Precision and recall

It is also known as positive predictive value. It is defined as the average probability of relevant retrieval.

Precision = Number of true positives/Number of true positives + False positives (5)

Recall

It is defined as the average probability of complete retrieval.

Recall= True positives/True positives + False negative (6)

D. Accuracy

A measure of a predictive model that reflects the proportionate number of times that the model is correct when applied to data [11].

The formula for calculating the Accuracy,

Accuracy=Number of correctly classified samples/Total number of samples (7)

E. Confusion Matrix

It is used for displaying the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is represented in the form of n-by-n, where n is the number of classes. The accuracy of each classification algorithms can be calculated from that.

V. IMPLEMENTATION

The implementation method of the predictive data mining has been described in this paper.

A. Architecture of PDM: Proposed Approach

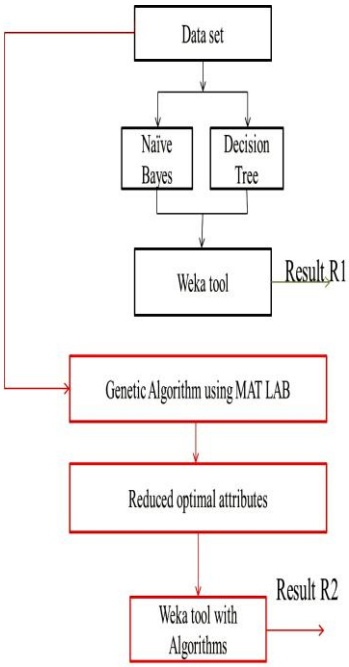

Fig 5.1: Proposed System

The working of the architecture is as follows: the data's of the patients who are having heart disease has been collected from the hospital. For the diagnosis of heart disease here two algorithms are being used which are Naïve Bayes and Decision Tree. The prediction of heart disease will be executed with the help of a tool known as Weka. Here the dataset is being used as the input for the prediction. The dataset consists of attributes and values. This tool will results the accuracy that how many patients are having the heart disease with in a particular time. In order to improve the efficiency and accuracy an optimizations process is carried out using genetic algorithm.

Summary of working of genetic algorithm:

1. Create a random initial population.
2. To create new population fitness value of current population has to be found.
3. Scales the raw fitness scores to convert them into a more usable range of values.
4. Selects members, called parents, based on their fitness.
5. Some of the individuals in the current population that have lower fitness are chosen as elite. These elite individuals are passed to the next population.
6. Produces children from the parents and the operation is known as crossover. Children are produced either by making random changes to a single parent called mutation

The genetic algorithm is being implemented with the help of Matlab. The optimized attributes are fed into Weka tool for the prediction purpose. Hence we will get a conclusion that optimization technique is the best method for improving the prediction of heart disease.

The Implementation has been done for finding the accuracy of decision tree and naïve bayes. The optimization part is the future work which is colored in red box.

B. DATA SET

The data set used in this work is collected from UCI machine learning repository which is a repository of databases, domain theories and data generators.

These are the attribute names which is the input given for patients record.

The data set attributes which are used in the paper and description as shown in Table 1

TABLE 1
UCI MACHINE LEARNING REPOSITORY

Predictable Attribute Diagnosis (value 0: <50% diameter narrowing (no heart disease); value 1: >50% diameter narrowing (has heart disease))
Key Attribute Patient ID – Patient's identification number
Input Attributes 1. Age in Year 2. Sex (value 1: Male; value 0: Female) 3. Chest Pain Type (value 1:typical type 1 angina, value 2: typical type angina, value3:non-angina pain; value 4: asymptomatic) 4. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl) 5. Restecg – resting electrographic results (value 0:normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy) 6. Exang - exercise induced angina (value 1: yes; value 0: no) 7. Slope – the slope of the peak exercise ST segment (value 1:unsloping; value 2: flat; value 3: downsloping) 8. CA – number of major vessels colored by floursopy (value 0-3) 9. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect) 10. Trest Blood Pressure (mm Hg on admission to the hospital) 11. Serum Cholestrol (mg/dl) 12. Thalach – maximum heart rate achieved 13. Oldpeak – ST depression induced by exercise

C. TOOLS USED

For the diagnosis of heart diseases there are many tools which are present and out of that the most efficient tool is Weka tool. Weka is an open source tool which supports graphical user interface. System is developed at the University of Waikato in New Zealand. It is a collection of machine learning algorithms for data mining tasks. This system is written using object oriented language Java. In this tool the algorithms can either be applied directly to a dataset. Weka contains tools for data pre-processing, classification, regression clustering, association rules, and visualization. Weka is an open source tool which supports graphical user interface. The input data file format should be in ARFF format. This file contains the complete information regarding the set of all attributes and also the values for that attributes.

D. COMPARISION RESULT

DM Technique	Accuracy
Naïve Bayes	85.03
Decision Tree	84.01

E.SIMULATION RESULTS

The collected attributes of different patient and values given for each and every attribute.

```
@relation hungarian-14-heart-disease
@attribute 'age' real
@attribute 'sex' { female, male}
@attribute 'chest_pain' { typ_angina, asympt, non_anginal, atyp_angina}
@attribute 'trestbps' real
@attribute 'chol' real
@attribute 'fbs' { t, f}
@attribute 'restecg' { left_vent_hyper, normal, st_t_wave_abnormality}
@attribute 'thalach' real
@attribute 'exang' { no, yes}
@attribute 'oldpeak' real
@attribute 'slope' { down, flat, up}
@attribute 'ca' real
@attribute 'thal' { fixed_defect, normal, reversable_defect}
@attribute 'num' { '<50', '>50_1', '>50_2', '>50_3', '>50_4'}
@data
28,male,atyp_angina,130,132,f,left_vent_hyper,185,no,0,?,?,?,'<50'
29,male,atyp_angina,120,243,f,normal,160,no,0,?,?,?,'<50'
29,male,atyp_angina,140,?,f,normal,170,no,0,?,?,?,'<50'
30,female,typ_angina,170,237,f,st_t_wave_abnormality,170,no,0,?,?,fixed_defect,'<50'
31,female,atyp_angina,100,219,f,st_t_wave_abnormality,150,no,0,?,?,?,'<50'
32,female,atyp_angina,105,198,f,normal,165,no,0,?,?,?,'<50'
32,male,atyp_angina,110,225,f,normal,184,no,0,?,?,?,'<50'
32,male,atyp_angina,125,254,f,normal,155,no,0,?,?,?,'<50'
33,male,non_anginal,120,298,f,normal,185,no,0,?,?,?,'<50'
34,female,atyp_angina,130,161,f,normal,190,no,0,?,?,?,'<50'
34,male,atyp_angina,150,214,f,st_t_wave_abnormality,168,no,0,?,?,?,'<50'
34,male,atyp_angina,98,220,f,normal,150,no,0,?,?,?,'<50'
35,female,typ_angina,120,160,f,st_t_wave_abnormality,185,no,0,?,?,?,'<50'
35,female,asympt,140,167,f,normal,150,no,0,?,?,?,'<50'
35,male,atyp_angina,120,308,f,left_vent_hyper,180,no,0,?,?,?,'<50'
```

Fig 5.2: Datasets of different patients

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.11 seconds

=== Summary ===

Correctly Classified Instances         250               85.034  %
Incorrectly Classified Instances        44               14.966  %
```

Fig 5.3: Naïve Bayes Implementation In Weka

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances         247               84.0136 %
Incorrectly Classified Instances        47               15.9864 %
```

Fig 5.4: Decision Tree Implementation In Weka

F.LINKS AND BOOKMARKS

1. Weka Data Mining Software
   http://www.cs.waikato.ac.nz/ml/weka

VI. CONCLUSIONS

Many sessions of experiments were conducted with the same datasets in Weka 3.6.0 tool. Data set of 294 records with 13 attributes is used and the outcome reveals that the Naïve Bayes outperforms and sometime Decision Tree. In Future Genetic algorithm will be used in order to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction. Prediction of the heart disease will be evaluated according to the result produced from it. Improvement is done to increase its consistency and efficiency. Benefit of using genetic algorithm is the prediction of heart disease can be done in a short time with the help of reduced dataset. Genetic algorithm will be implemented with the MATLAB.

REFERENCES

1. Asha Rajkumar and G.Sophia Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm," Global Journal of Computer Science and Technology,Vol.10, Issue 10 Ver. 1.0,2010.
2. Hai H.Dam., Hussain A.Abbass and Xin Yao, "Neural – Based Learning Classifier Systems", IEEE Transactions on Knowledge and Data Engineering, Vol.20, No.1, 2008.
3. Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006
4. Jiawei Han and Micheline kamber,"Data Mining Concepts and Techniques", Second Edition, Elsevier Inc, San Francisco, 2006.
5. M. Anbarasi and E.Anupriya, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology, Vol. 2(10), pp.5370-5376,2010.
6. M. Ilayaraja,"Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm", IEEE-International Conference on Pattern Recognition, Informatics and Mobile Engineering,2013
7. Nidhi Bhatla, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 8,2012.
8. Sunita Soni and Ujma Ansari ," Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887), Volume 17– No.8, pp. 43-48,2011.
9. TanG. and Cbye H, "Data mining applications in healthcare," Journal of Healthcare Information Management. Vol. 19, No.2,2004.
10. T.John Peter, "An empirical study on prediction of heart disease using classification data mining techniques", IEEE-International Conference On Advances In Engineering, Science And Management, pp. 514-518, 2012.

11. Wasan, K. and Kaur, H, "Empirical study on applications of data mining techniques in healthcare," Journal of Computer Science, Vol. 2, No.2.,2006