International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Diagnosis of Heart Disease Using Data Mining Technique

Shinde S. B.[1], Amrit Priyadarshi[2]

[1]PG Student, Department of Information Technology, DKGOI-FOE, Daund, Pune, Maharashtra, India

[2]Assistant Professor, Department of Computer Engineering, DKGOI-FOE, Daund, Pune, Maharashtra, India

Abstract: ***Data mining techniques have been widely used in clinical decision support systems for prediction or diagnosis of various diseases with accuracy. These techniques are used to discover hidden patterns and relationships from the hospital data. One important applications of data mining technique is to diagnose the heart diseases because it is one of the reasons for deaths over the world. Almost all systems which predict heart diseases, use medical dataset as inputs like age, sex, cholesterol, blood sugar etc. There is no system which predicts heart diseases based on the attributes such as diabetes, family history, tobacco smoking, intake of alcohol, obesity, hypertension or any other physical inactivity etc. Heart disease patients have lot of these visible risk factors in common which can be used very effectively for detecting. System based on the risk factors would not only help medical professionals but also it would give patients a warning about the probable presence of heart disease even before he visits a hospital or goes for costly medical Checkups. Hence this system presents a technique for prediction of heart disease. These techniques involve one successful data mining technique named Naïve Bayesian algorithm. It also provides the training tool for nurses and medical students to predict patient having heart disease.***

Keywords: Heart Disease, Data Mining, KDD, Decision Support and Naïve Bayes.

1. Introduction

Data Mining is the non trivial process which discover useful and understandable patterns from historical databases. Data mining is used for extracting or "mining" knowledge from the large amount of data. It extracts the relationships and patterns from the database.It includes various types of areas like machine learning, statistics, pattern recognition, artificial intelligence and data visualization [1].Traditional database queries access database using a well defined query stated in a language such as SQL [2]. Output of the SQL query consist data from the database that satisfies the query. The output is just a subset of the database but it may contain aggregations. Data mining access the database which differs from the traditional access.The process of data mining or Knowledge Discovery (KDD) is the conversion of inputted data into knowledge for decision making. KDD process consists an iterative sequence of data cleaning, data selection, data integration, data transformation, data mining, result interpretation and validation, incorporation of the discovered knowledge.The output of the data mining query probably is not a subset of the database, instead it is the output of some analysis of the contents of database.

Many hospital information systems are designed to support inventory management, billing of patient etc. Some hospitals use different systems, but are largely limited. They can answer simple queries like 'What is the average age of patients who has diagnosed a heart disease?'. But can not answer complex queries like 'Find the probability of patients who diagnosed a heart disease.' [3]. Diagnosis of disease is depend on doctor's decision and experience of doctor instead of knowledge data which is hidden in the database. This practice leads to unwanted errors and more medical costs and it affects the quality of service provided to patients. This proposed system uses data mining technique such as naive bayesian algorithm which reduce medical errors, improves patient safety, decrease practice variation which are unwanted and also improves patient result [4] . So, data mining are used to generate a knowledge rich data which improves the quality of clinical decisions.

The prediction of heart disease with classification algorithm is described in this paper. The knowledge data is classified by using different classification algorithms such as Naive Bayes, K-Nearest Neighbor, Decision Tree and the accuracy of each classification algorithm is noted. From all these algorithm, NB performs better than other methods for heart disease classification. Medical decision support systems are designed to support clinicians in their diagnosis for heart disease. They typically work through an analysis of medical data and a knowledge base of clinical experts. The quality of medical decisions for heart disease can be increased by using Bayesian algorithm.

2. Data Mining Techniques

Data mining techniques are used to explore, analyze and extract medical data using complex algorithms in order to discover unknown patterns. Researchers are using data mining techniques for the diagnosis of many diseases such as heart disease, diabetes, cancer etc with good accuracy. Researchers have been applying different data mining techniques such as naïve bayes, KNN algorithm, neural network, decision tree and support vector machine (SVM) for prediction of heart diseases. Decision Support System on Heart Disease Prediction was built using data mining technique such as Naïve Bayesian Algorithm and it proposed extracting significant patterns for heart disease prediction. Polatet al., developed system using hybrid fuzzy and k nearest neighbor approach for the prediction of heart disease, which had 75.18 % accuracy in diagnosis. In another system neural network ensemble was used in the diagnosis of heart

disease with an accuracy of 78.148 %. and Decision Tree with an accuracy 76.6%.

3. Implementation of Naïve Bayesian Algorithm

The Naïve Bayesian Classification Algorithm represents a statistical method as well as supervised learning method for classification. Assumes a probabilistic model which allows us to solve the diagnostic and predictive problems. Bayes classification has been proposed which is based on Bayes rule of conditional probability. Naïve Bayesian rule is a technique used to estimate the likelihood of a property from the given data set. The approach is called "naïve" because it assumes the independence between the various attribute values. Bayesian classification can be seen as both a descriptive and a predictive type of algorithm. The probabilities are descriptive and used to predict the class membership for a target tuple.

3.1 Bayes Rule

Conditional probability is likelihood of some conclusion, *C*, given some evidence/ observation, *E*, where a dependence relationship exists between *C* and *E*. This probability is denoted as P*(C |E)* where

P (C | E) = P (E | C) P (C) / P (E)

3.2 Data Source

A total of 1000 records with 13 attributes were obtained from the database. These records are divided into two database i.e training daabase (700) and testing database (300). Records for each set are selected randomly to avoid bias. The 'Diagnosis' attribute is used to predict the heart disease with value "2" for patient having heart disease and "1" for patient having no heart disease. The 'PatientID' attribute is used as a key and others are input attributes.

Predictable attributes
1. Diagnosis (value 2 – Patient having heart disease and value 1- Patient having no heart disease)

Key Attribute
1. PatientID – Patient's identification number

Input Attributes
1. Age (value 1:<=40, value 2:<=60 and >40, value 3: >60)
2. Sex (value 0: female, value 1: male)
3. Chest Pain Type (value 1:Low, value 2: Medium, Value 3: High, Value 4: Very High)
4. Blood Pressure (value 1:<=80, value 2:<=120 and >80, value 3: >120)
5. Blood Sugar (value 0: Low, value 1: High)
6. Serum Cholesterol(value 1:<=180, value 2:<=400 and >180, value 3: >400)
7. Resting ECG (value 0: normal, value 1: wave abnormality, value 2: showing probable or definite left, ventricular hypertrophy)
8. Heart Rate (value 1:<=120, value 2:<=180 and >120, value 3: >180)
9. Exercise Induced Angina (value 0: Low, value 1: High)
10. Oldpeak (ST depression value 1:<=1, value 2:<=2.5 and >1, value 3: >2.5)
11. Slope of the peak Exercise (value 1: unslopping, value 2: flat, value 3: downslopping)
12. No.of major vessels(value 1:Low, value 2: Medium, Value 3: High, Value 4: Very High)
13. Thal (value 3: normal; value 6: fixed defect; value 7:reversible defect)

Figure 1: Description of Attributes

3.3 Algorithm

Given the Hospital data set
1. Estimate the prior probability P (cj) for each class by counting how often each class occurs in the training data.
2. For each attribute Xi find P (xi) by counting the number of occurrences of each attribute value.
3. Find probability P (xi/cj) by counting how often each value occurs in the class in the training data.
4. Do this for all attributes and all values of these attributes. To classify a target tuple estimate P (ti/cj) = ∏Pk=1 P (xik/cj).
5. Calculate P (ti). This can be done by finding the likelihood that this tuple is in each class and then adding all this values.
6. Find posterior probability P (cj/ti) for each class. It is the product of conditional probabilities for each attribute value.
7. Select class with highest probability value of P (cj/ti) value for test tuple

Mathematical Formulae

P (HeartDis Yes)= No.of Records with Result Yes / Total no. of Records

P (HeartDis No)= No.of Records with Result No / Total no. of Records

P (t/yes) = P (Age (low) yes) * P (Sex (Male) yes) * P (BP (High) yes) * P (Chol (High) yes) *P (Heart_Rate (High) yes)*P(Vessels(High)yes)*P(Chest_Pain(High)yes)*P(ECG(High)yes)*P(Exer_angina(High)yes)*P(old_peak(High)yes) *P(Thal(High)yes)*P(Blood_sugar(High)yes)*P(Slope_peak (High)yes)

P (t/no) = P (Age (low) no) * P (Sex (Male) no) * P (BP (High) no)* P (Chol (High) no) * P (Heart_Rate (High) no) * P(Vessels(High)no)*P (Chest_Pain(High) no)*P(ECG(High) no)* P(Exer_angina(High) no)* P (old_peak(High) no)* P (Thal(High)no)*P(Blood_sugar(High)no)*P(Slope_peak (High) no)

P (Likelihood of yes) = P (t/yes) * P (Heart_Disease yes)
P(Likelihood of no)= P (t/no) * P (Heart_Disease no)

Now we find the total probability,

P(yes/t) = P (t/yes) * P (Heart_Disease yes) / P (T)
P(no/t) = P (t/no) * P (Heart_Disease no) / P (T)

P (yes/t) >= P (no/t) then input query is classified as Heart Disease category

Else No Heart Disease category

Accuracy Calculation

Accuracy refers to the percentage of correct predictions made by the model compared with actual classifications in the test data.

Accuracy = Total no. of Correctly Predicted Record / Total no. of training Record

4. Proposed System

The main goal of this system is to predict heart disease using data mining technique such as Naive Bayesian Algorithm. Raw hospital data set is used and then preprocessed and transformed the data set. Then apply the data mining technique such as Naïve Bayes algorithm on the transformed data set. After applying the data mining algorithm , heart disease is predicted and then accuracy is calculated.

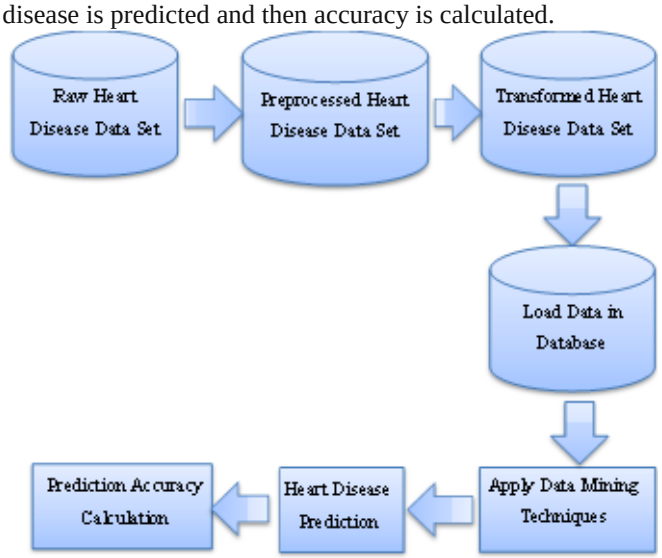

Figure 2: System Architecture

5. Result and Analysis

Result and analysis is done on the hospital data set. Table 1 shows the accuracy obtained by changing the number of records in training data set.

Table 1: Accuracy (%)

No. of Records in Training Data set	No. of Records in Testing Data set	No. of correctly claasified instances	No. of correctly claasified instances	Accuracy (%)
700	300	265	35	88.33
700	270	234	36	86.66

6. Conclusion

This paper presents automated and effective heart attack prediction using data mining technique such as Naïve Bayesian Classification Algorithm. Data Mining is used to extract the different patterns i.e Hidden Knowledge from historical Heart related database. This system answers the complex queries like “Find the probability of patients who cause a heart disease .” This system can easily predict heart disease with good accuracy. It can be further enhanced and expanded. For Predicting heart disease 13 attributes are used. Besides this list, other attributes which will effect on results such as stress, pollution and previous medical history can be used. It also provides the training tool for nurses and medical students to predict patient having heart disease.

Other data mining techniques such as Clustering and Association Rule can be used to analyze patients behavior. Continuous data can also be used instead of just categorical data. Text Mining also can be used to mine the vast amount of unstructured data available in healthcare databases.

References

[1] Han and kamber, “Data mining concepts and techniques”, 2nd edition(2010)
[2] Margaret H. Dunham, Southern Methodist University ,’Data Mining- Introductory and Advanced Topics, ISBN: 0130888923 published by Pearson Education, Inc.,Sixh Impression (2009).
[3] 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
[4] Wu, R., Peters, W., Morgan, M.W.: “The Next Generation Clinical Decision Support: Linking
[5] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi ,Business Review, Vol. 8, No. I (January - June 2007)
[6] Milan Kumari, Sunila Godara, Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, IJCST Vol. Iss ue 2, June 2011.
[7] Shantakumar B.Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using DataMining and Artificial Neural Network, European Journal of Scientific Research, ISSN 1450-216X, Vol.31 No.4 (2009), pp.642-656.
[8] Tang, Z. H., MacLennan, J.: “Data Mining with SQL Server 2005”, Indianapolis: Wiley, 2005. Evidence to Best Practice”, Journal Healthcare Information Management.