International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

A Survey on Heart Disease Diagnosis and Prediction using Naive Bayes in Data Mining

B.V. Baiju[†*] and R.J. Remy Janet[‡]

[†]Department of Information Technology, Hindustan University, Chennai, India
[‡]Department of CSE, M.N.M.Jain Engineering College, Chennai, India

Accepted 05 April 2015, Available online 10 April 2015, **Vol.5, No.2 (April 2015)**

Abstract

Data Mining is non trivial extraction of implicit data, previously not known, and imaginably useful information from data. Data mining is an essential process where intelligent methods are applied in order to extract data patterns. Using data mining we can evaluate patterns which we can use in future to take intelligent decisions and we can present the knowledge we extracted in better way. Data Mining refers to using a variety of techniques to identify information or decision making knowledge in the database and extracting these in a way that they can put to use in areas such as decision making, predictions, for valuable forecasting and computation. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information, to take decisions effectively, to discover the relations that connect parameters in a database is the subject of data mining. As large amount of data is generated in medical organisations (hospitals, medical centers) but as this data is not properly used. There is a wealth of hidden information present in the datasets. This unused data can be converted into useful data. For this purpose we can use different data mining techniques. In this study, we are applying Naïve Bayes data mining classifier technique which produces an optimal prediction model using minimum training set. Data mining is the analysis step of the Knowledge Discovery in Databases process (KDD). Data mining involves use of techniques to find underlying structures and relationships in a large database. Using medical profile such as age, sex, blood pressure and blood sugar we can easily predict the likelihood of patients getting heart disease. In this paper we have evaluated the performance of new classification approach that uses the experienced Doctor's knowledge to assign the weight to each attribute. More weight is assigned to the attribute having high impact on disease prediction.

Keywords: *Data mining, Naive bayes, heart disease, KDD, disease prediction*

1. Introduction

Data Mining is the nontrivial process of identifying true, novel, potentially useful and finally understandable pattern in data with the wide use of databases and the explosive growth in their sizes. Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining is the search for the relationships, association and global patterns that exist in large databases but are hidden among large amounts of data. The essential process of Knowledge Discovery is the conversion of data into knowledge in order to aid in decision making, referred to as data mining. Knowledge Discovery process consists of an iterative sequence of data pre-processing like cleaning, data integration, correct data selection, data mining pattern identification and knowledge presentation. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data.

In this fast moving world people want to live a very luxurious life so they work like a machine in order to earn lot of money and live a comfortable life therefore in this race they forget to take care of themselves, because of this there food habits change their entire lifestyle change, in this type of lifestyle they are more tensed they have blood pressure, sugar at a very young age and they don't give enough rest for themselves and eat what they get and they even don't bother about the quality of the food if sick the go for their own medication as a result of all these small negligence it leads to a major threat that is the heart disease. It is a world known fact that heart is the most essential organ in human body if that organ gets affected then it also affects the other vital parts of the body. Therefore it is very important for people to go for a heart disease diagnosis.

As a result of this people go to healthcare practitioners but the prediction made by them is not 100% accurate. Quality service implies diagnosing patients correctly and administering treatments that

*Corresponding author: **B.V. Baiju**

are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This research has developed a prototype Heart Disease Prediction System (HDPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals . HDPS can answer complex what if queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. HDPS is Web-based, user-friendly, scalable, reliable and expandable.

2. Related Work

The main objective of this research is to develop a prototype Health Care Prediction System using, Naive Bayes .The System can discover and extract hidden knowledge associated with heart diseases from a historical heart disease database. Most hospitals today employ sort of hospital information systems to manage their healthcare or patient data. These systems typically generate huge amounts of data. There is a wealth of hidden information in these data that is largely untapped. How data is turned into useful information that can enable healthcare practitioners to make intelligent clinical decisions. The main objective of this research is to develop a Decision Support in Heart Disease Prediction System (DSHDPS) using one data mining modeling technique, namely, Naïve Bayes. DSHDPS is implemented as web based questionnaire application. Based on user answers, it can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database. We provide the report of the patient in two ways using chart and pdf which indicates whether that particular patient having the heart disease or not. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

The diagnosis of diseases is a significant and tedious task in medicine. The detection of heart disease from various factors or symptoms is a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects. Thus the effort to utilize knowledge and experience of numerous specialists and clinical screening data of patients collected in databases to facilitate the diagnosis process is considered a valuable option. Providing precious services at affordable costs is a major constraint encountered by the healthcare organizations (hospitals, medical centers). Valuable quality service denotes the accurate diagnosis of patients and providing efficient treatment. Poor clinical decisions may lead to disasters and hence are seldom entertained. Besides, it is essential that the hospitals decrease the cost of clinical test. Appropriate computer-based information and/or decision support systems can aid in achieving clinical tests at a reduced cost. Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables.

To develop the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability) a novel technique. To achieve this, they have used several classifiers e.g. Bayesian Classifiers, CMAR (Classification based on Multiple Association Rules), C4.5 (Decision Tree) and SVM (Support Vector Machine). An Intelligent Heart Disease Prediction System (IHDPS) is developed by using data mining techniques Naive Bayes, Neural Network, and Decision Trees. Each method has its own strength to get appropriate results. To build this system hidden patterns and relationship between them is used. It is web-based, user friendly & expandable. The prediction of Heart disease, Blood Pressure and Sugar is done with the aid of neural networks. The dataset contains records with 13 attributes in each record. The supervised networks i.e. Neural Network with back propagation algorithm is used for training and testing of data.

3. Data Mining Concepts In Health Care

Data Mining aims at discovering knowledge out of data and presenting it in a form that is easily compressible to humans. It is a process that is developed to examine large amounts of data routinely collected. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. In practice, the two primary goals of data mining tend to be prediction and description. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Description, on the

other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data-mining activities into one of two categories.

3.1 Predictive data mining

Predictive models can be used to forecast explicit values, based on patterns determined from known results. For example, from a database of customers who have already responded to a particular offer, a model can be built that predicts which prospects are likeliest to respond to the same offer.

3.2 Descriptive data mining

Descriptive models describe patterns in existing data, and are generally used to create meaningful subgroups such as demographic clusters.

On the predictive end of the spectrum, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, or other similar tasks. On the other, descriptive, end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets.

4. Basic Terms Related To Data Mining

4.1 Classification

Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be “sunny”, “rainy” or “cloudy”.

4.2 Supervised learning

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier. The inferred function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

4.3 Unsupervised learning

In machine learning, unsupervised learning refers to the problem of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. This distinguishes unsupervised learning from supervised learning.

5. Heart Disease

The heart is important organ or part of our body. Life is itself dependent on efficient working of heart. If operation of heart is not proper, it will affect the other body parts of human such as brain, kidney etc. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart and blood vessel system within it.

There are number of factors which increase the risk of Heart disease:

- Family history of heart disease
- Smoking
- Cholesterol
- Poor diet
- High blood pressure
- High blood cholesterol
- Obesity
- Physical inactivity
- Hyper tension

The following are the symptoms of a heart attack

- Discomfort, pressure, heaviness, or pain in the chest, arm or below the breastbone.
- Discomfort radiating to the back, jaw, throat, or arm.
- Fullness, indigestion, or choking feeling (may feel like heartburn).
- Sweating, nausea, vomiting, or dizziness.
- Extreme weakness, anxiety, or shortness of breath.
- Rapid or irregular heartbeats

5.1 Types of Heart diseases

Heart disease is a broad term that includes all types of diseases affecting different components of the heart. Heart means 'cardio.' Therefore, all heart diseases belong to the category of cardiovascular diseases. Some types of Heart diseases are

Coronary heart disease

It also known as coronary artery disease (CAD), it is the most common type of heart disease across the world. It is a condition in which plaque deposits block the coronary blood vessels leading to a reduced supply of blood and oxygen to the heart.

Angina pectoris

It is a medical term for chest pain that occurs due to insufficient supply of blood to the heart. Also known as angina, it is a warning signal for heart attack. The chest pain is at intervals ranging for few seconds or minutes.

Congestive heart failure

It is a condition where the heart cannot pump enough blood to the rest of the body. It is commonly known as heart failure.

Cardiomyopathy

It is the weakening of the heart muscle or a change in the structure of the muscle due to inadequate heart pumping. Some of the common causes of cardiomyopathy are hypertension, alcohol consumption, viral infections, and genetic defects.

Congenital heart disease

It also known as congenital heart defect, it refers to the formation of an abnormal heart due to a defect in the structure of the heart or its functioning. It is also a type of congenital disease.

Arrhythmias

It is associated with a disorder in the rhythmic movement of the heartbeat. The heartbeat can be slow, fast, or irregular. These abnormal heartbeats are caused by a short circuit in the heart's electrical system.

Myocarditis

It is an inflammation of the heart muscle usually caused by viral, fungal, and bacterial infections affecting the heart. It is an uncommon disease with few symptoms like joins pain, leg swelling or fever that cannot be directly related to the heart.

6. Data Source

Clinical databases have accumulated giant quantities of information about patients and their medical conditions. The term cardiovascular disease encompasses the diverse diseases that affect the heart. Heart disease is the major cause of casualties in the world. Cardiovascular disease kills one person each thirty four seconds within the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular diseases 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. Cardiovascular disease (CVD) results in severe illness, disability, and death.
Record set with medical attributes was obtained from the Cleveland Heart Disease database. With the assistance of the dataset, the patterns vital to the heart attack prediction are extracted. The records were split equally into two datasets: training dataset and testing dataset. To avoid bias, for every set were hand-picked haphazardly.

The attribute "Diagnosis" is known as the predictable attribute with value "1" for patients with heart disease and value "0" for patients with no heart disease. "Patient's test" is employed as a record, last attribute as output and, the remaining are input attributes. It is assumed that issues like missing, inconsistent, and redundant data have all been resolved.

7. Naive Bayes Algorithm

Naïve Bayes Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. A naive Bayes classifier is a term dealing with a simple probabilistic classification based on applying Bayes theorem. In simple terms, a naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting Naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Here independent variables are considered for the purpose of prediction or occurrence of the event. The algorithm is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. For example, a patient may be observed to have certain symptoms. Based on the observation, Bayes' theorem can be used to compute the probability that a proposed diagnosis is correct.

Baye's Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes theorem can be stated as follows,

P (B given A) = Prob(A and B) / Prob (A)

To calculate the probability of B given A, the algorithm counts the number of cases where A and B occur together and divides it by the number of cases where A occurs alone. The algorithm works as follows:

- Each data sample is represented by an n dimensional feature vector, X = (x1, x2….. xn), depicting n measurements made on the sample from n attributes, respectively A1, A2… An.
- Suppose that there are m classes, C1, C2……Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive probability assigns an unknown sample X to the class Ci if and only if:

P (Ci/X)>P (Cj/X) for all 1< = j< = m and j!= i

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem,
P (Ci/X) = (P(X/Ci)P(Ci))/P(X)

As P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1) = P(C2) = …..= P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci) = si/s , where Si is the number of training samples of class Ci, and s is the total number of training samples.

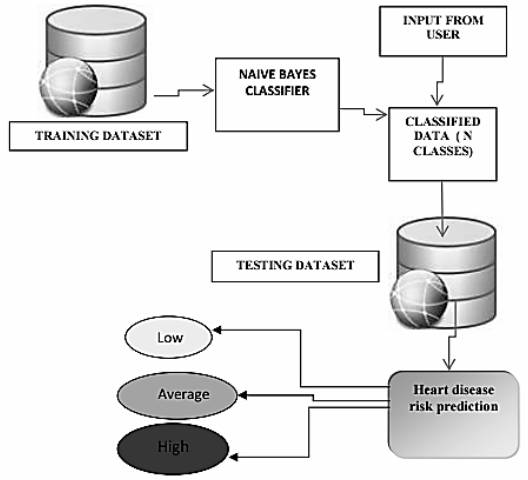

Fig.1 System Architecture

7.1 Implementation of bayesian classification

The Naïve Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state.

Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. We prefer naive bayes implementation for the following reasons

- When the data is high.
- When the attributes are independent of each other.
- When we want more efficient output, as compared to other methods output.

Issues and Challenges

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor's intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

Conclusion

Decision Support in Heart Disease Prediction System is developed using Naive Bayesian Classification technique. The system extracts hidden knowledge from a historical heart disease database. This is the most effective model to predict patients with heart disease. For, example it can incorporate other medical attributes besides the above list. It can also incorporate other data mining techniques. Continuous data can be used instead of just categorical data. The overall objective is to study the various data mining techniques available to predict the heart disease and to compare them to find the best method of prediction.

References

Frawley and G. Piatetsky –Shapiro, (1996), Knowledge Discovery in Databases: An Overview. Published by the AAAI Press/ The MIT Press, Menlo Park, C.A.

Yanwei, X.; Wang, J.; Zhao, Z.; Gao, Y.,(2007), Combination data mining models with new medical data to predict outcome of coronary heart disease, Proceedings International Conference on Convergence Information Technology, pp. 868 – 872.

Sellappan Palaniappan, RafiahAwang, (2008), Intelligent Heart Disease Prediction System Using Data Mining Techniques, IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8.

Ho, T. J.,(2005), Mining and Data Warehousing, Prentice Hall.

Sellappan, P., Chua, S.L., (2005), Model-based Healthcare Decision Support System, Proc. Of Int. Conf. on Information Technology in Asia CITA'05, 45-50, Kuching, Sarawak, Malaysia.

Tang, Z. H., MacLennan, J., (2005),Data Mining with SQL Server 2005", Indianapolis: Wiley.

Han, J., Kamber, M.,(2006), "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers..

Boser, B. E., I. Guyon, and V. Vapnik, (1992), A training algorithm for optimal margin Classifiers, In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pp.144 -152., ACM Press.

V. Vapnik, (1995). The Nature of Statistical Learning Theory, NY, Springer Verlag.

Christopher J.C. Burges, (1998), A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, Springer, 2(2), pp.121-167.