INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

ACCURACY ENHANCEMENT OF CLASSIFICATION ON PREDICTIVE DATA MINING MODEL

PRADIP S. INGLE[1], PROF. SHRIKANT P. AKARTE[2]

1. Department Of Computer Science & Engg., Prof. Ram Meghe Institute of Technology & Research , Badnera Amravati, Sant Gadge Baba Amravati University Amravati (maharashtra) India

2. Department Of Computer Science & Engg., Prof. Ram Meghe Institute of Technology & Research , Badnera Amravati, Sant Gadge Baba Amravati University Amravati (maharashtra) India

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: Data mining (DM) often referred as knowledge discovery in database a process of nontrivial extraction of implicit, previously unknown and potentiality useful information from a large volume of data. Data mining is a multi-disciplinary approach comprising of database technology, high performance computing, machine learning, numerical mathematics, statistics and visualization. The primary goal of the classification frameworks is to provide a better result in terms of accuracy. A classification paradigm is a data mining framework containing all the concepts extracted from the training dataset to differentiate one class from other classes existed in data. This research has developed a prototype Intelligent Heart Disease Prediction System using data mining techniques, 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.

Keywords: Data Mining, Prediction System, Neural Network, Heart Disease, Classification.

Corresponding Author: MR. PRADIP S. INGLE

Access Online On:

www.ijpret.com

How to Cite This Article:

Pradip S. Ingle, IJPRET, 2015; Volume 3 (9): 1453-1461

PAPER-QR CODE

INTRODUCTION

A classification paradigm is a data mining framework containing all the concepts extracted from the training dataset. In most of the cases we cannot get better accuracy particularly for huge dataset and dataset with several groups of data[1]. When a classification framework considers whole dataset for training then the algorithm may become unusable because dataset consists of several groups of data. For unknown data, we classify with the best match group/model and attain higher accuracy rate than the conventional Naive Bayes classifier[6].

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.

Objectives:-

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. The main objective of this research is to develop a Decision Support in Heart Disease Prediction System (HDPS) using data mining modeling techniques, namely, Decision Trees, Naïve Bayes and Neural Network. HDPS 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 [5, 6]. A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs.

Literature Review

In today's life the major health problem is heart disease. The main objective of this paper is to predicate about heart diseases by using various data mining techniques.

- An Intelligent Heart Disease Prediction System (IHDPS) is developed by using data mining techniques Naive Bayes, Neural Network, and Decision Trees was proposed by Sellappan Palaniappan et al [13]. 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.

- Everss, Franck Le Duff et al [9] builds a decision tree with database of patient for a medical problem.
- J.Warren et al [10] projected an approach on basis of coactive neuro-fuzzy inference system (CANFIS) for prediction of heart disease. The CANFIS model uses neural network capabilities with the fuzzy logic and genetic algorithm.
- C. S. Pattichis et al [8] uses a classification method for the extraction of multiparametric features by assessing HRV (Heart Rate Variability) from ECG, data pre-processing and heart disease pattern.
- Proposed System Analysis

In many hospitals manage healthcare data using healthcare information system; as the system contains huge amount of data, used to extract hidden information for making intelligent medical diagnosis. The main objective of this research is to build Heart Disease Prediction System that gives diagnosis of heart disease using historical heart database. To develop this system, medical terms such as sex, blood pressure, and cholesterol like 15 input attributes are used. The data mining classification techniques viz. Neural Networks, Decision Trees, and Naive Bayes are used [7].

A. *Existing Systems:-*

Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. There are many ways that a medical misdiagnosis can present itself. Whether a doctor is at fault, or hospital staff, a misdiagnosis of a serious illness can have very extreme and harmful effects. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The National Patient Safety Foundation cites that 42% of medical patients feel they have had experienced a medical error or missed diagnosis. Patient safety is sometimes negligently given the back seat for other concerns, such as the cost of medical tests, drugs, and operations. Medical Misdiagnoses are a serious risk to our healthcare profession. If they continue, then people will fear going to the hospital for treatment [9]. We can put an end to medical misdiagnosis by informing the public and filing claims and suits against the medical practitioners at fault.

B. *Proposed Systems:-*

Thus we proposed that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The main objective of this research is to develop a prototype Heart Disease Prediction System (HDPS) using three data mining modeling techniques, namely, Decision Trees, Naïve Bayes and Neural Network [11]. So its providing effective treatments, it also helps to reduce treatment costs, to enhance visualization of interpretation.

C. *Analyzing the Data set:-*

A data set (or dataset) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the data set in question. It lists values for each of the variables, such as height and weight of an object or values of random numbers. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows.

A total of 500 records with 15 medical attributes (factors) were obtained from the Heart Disease database lists the attributes. The records were split equally into two datasets: training dataset (455 records) and testing dataset (454 records). To avoid bias, the records for each set were selected randomly.

The attribute “Diagnosis” was identified as the predictable attribute with value “1” for patients with heart disease and value “0” for patients with no heart disease [13]. The attribute “Patient ID” was used as the key; the rest are input attributes. It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved. Here in our project we get a data set from .dat file as our file reader program will get the data from them for the input of Naïve Bayes based mining process [14].

D. *Input attributes*

1. Sex (value 1: Male; value 0 : Female)

2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)

3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0:< 120 mg/dl)

4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)

5. Exang – exercise induced angina (value 1: yes; value 0: no)

6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)

7. CA – number of major vessels colored by floursopy (value 0 – 3)

8. Thal (value 3: normal; value 6: fixed defect; value 7:reversible defect)

9. Test Blood Pressure (mm Hg on admission to the hospital)

10. Serum Cholesterol (mg/dl)

11. Thalach – maximum heart rate achieved

12. Oldpeak – ST depression induced by exercise relative to rest

13. Age in Year

14. Height in cms

15. Weight in Kgs.

Naives Baye's in Mining:

I recommend using Probability for Data Mining for a more in-depth introduction to Density estimation and general use of Bayes Classifiers, with Naive Bayes Classifiers as a special case. But if you just want the executive summary bottom line on learning and using Naive Bayes classifiers on categorical attributes then these are slides for you.

Bayes' 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.

Bayes' Theorem:

Prob (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 [7].

Decision Trees in Mining:

The decision tree approach is more powerful for classification problems and prediction. There are two steps in this techniques building a tree & applying the tree to the dataset. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy. Decision trees represent rules, which can be understood by humans and used in knowledge system such as database. It requires no domain knowledge or parameter setting and can handle high dimensional data. Hence it is more appropriate for exploratory knowledge discovery. It still suffers from repetition and replication. Therefore necessary steps need to be taken to handle repetition and replication. The performance of decision trees can be enhanced with suitable attribute selection. Correct selection of attributes partition the data set into distinct classes. Our work uses J48 decision tree for classification. Observations show that Decision trees outperform the other two classifiers but take more time to build the model. This technique gives maximum accuracy on training data [8]. The overall concept is to build a tree that provides balance of flexibility & accuracy.

A Neural Network:

A neural network (NN) is a parallel, distributed information processing structure consisting of multiple numbers of processing elements called node, they are interconnected via unidirectional signal channels called connections.

System Design

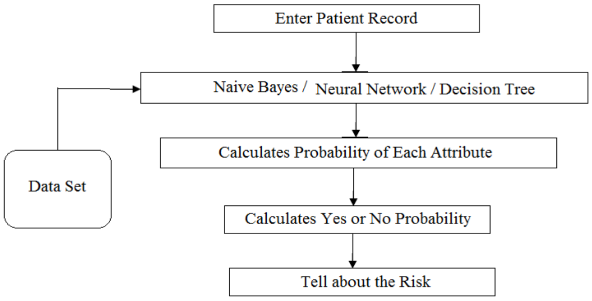
Enter Patient Record
Naive Bayes / Neural Network / Decision Tree
Calculates Probability of Each Attribute
Calculates Yes or No Probability
Tell about the Risk
Data Set

Fig. System Design

This architecture shows all the processes from input implementation to output Predict Heart Disease Using Data Mining Algorithms. Hence, firstly we collect the symptoms from the user and input in it. Then analyse the forwarded data for implementing the algorithm then designing the questionnaires and heart diseases in WEB. After that analysed data is saved in database, in this way by using all architectural processes we get the results.

Benefits and limitations

Heart Disease Prediction System can serve a training tool to train nurses and medical students to diagnose patients with heart disease. It can also provide decision support to assist doctors to make better clinical decisions or at least provide a "second opinion."

The current version of Heart Disease Prediction System is based on the 15 attributes .This attributes may need to be expanded to provide a more comprehensive diagnosis system. Another limitation is that it only uses categorical data. For some diagnosis, the use of continuous data may be necessary. Another limitation is that it only uses three data mining techniques. Additional data mining techniques can be incorporated to provide better diagnosis. The size of the dataset used in this research is still quite small. A large dataset would definitely give better results.

CONCLUSION:-

The objective of our work is to provide a study of different data mining techniques that can be employed in automated heart disease prediction systems. Various techniques and data mining classifiers are defined in this work. Applying data mining techniques to help health care professionals in the diagnosis of heart disease is having some success; the use of data mining techniques to identify a suitable treatment for heart disease patients has received less attention. Three data mining classification techniques were applied namely Decision trees, Naive Bayes & Neural Networks. From results it has been seen that Neural Networks, Decision trees & Naive Bayes provides accurate results.

REFERENCES

1. Shadab Adam Pattekari and Asma Parveen, International Journal of Advanced Computer and Mathematical SciencesISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294.

2. Chaitrali S. Dangare Student, M.E. (CSE) Dept. CSE Walchand Institute of Technology Solapur, Maharashtra, India International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

3. Mrs.G.Subbalakshmi (M.Tech),Kakinada Institute of Engineering & Technology(Affiliated to JNTU-Kakinada),Yanam Road, Korangi-533461,E.G.Dist., A.P., India. G.Subbalakshmi et al. / Indian Journal of Computer Science and Engineering (IJCSE)ISSN : 0976-5166 Vol. 2 No. 2 Apr-May 2011.

4. Jyoti Soni Student, M.Tech (CSE). Raipur Institute of Technology Raipur, Chhattisgarh, India International Journal of Computer Applications (0975 – 8887)Volume 17– No.8, March 2011.

5. S. Koch, —Home telehealth—Current state and future trends,‖ Int. J. Med. Inform., vol. 75, no. 8, pp. 565–576, 2006.

6. C. S. Pattichis, C. N. Schizas, M. S. Pattichis, E. Micheli-Tzanakou, E. C. Kyriakou, and D. I. Fotiadis, —Introduction to the special section on computational intelligence in medical systems,‖ IEEE Trans. Inform. Technol. Biomed., vol. 13, no. 5, pp. 667–672, Sep. 2009.

7. S. G. Mougiakakou, I. K. Valavanis, N. A. Mouravliansky, A. Nikita, and K. S. Nikita, —DIAGNOSIS: A telematics-enabled system for medical image archiving, management, and diagnosis assistance,‖ IEEE Trans. Instrum. Meas., vol. 58, no. 7, pp. 2113–2120, Jul. 2009.

8. P. A. Bath, —Data mining in health and medical information,‖ Annu. Rev. Inform. Sci. Technol., vol. 38, pp. 331–369, 2004.

9. R. Gaikwad and J.Warren, —The role of home-based information and communications technology interventions in chronic disease management: A systematic literature review,‖ Health Inform. J., vol. 15, no. 2, pp. 122–146, 2009.

10. A. Martinez, E. Everss, J. L. Rojo-Alvarez, D. P. Figal, and A. Garcia- Alberola, —A systematic review of the literature on home monitoring for patients with heart failure,‖ J. Telemed. Telecare, vol. 12, no. 5, pp. 234–241, 2006.

11. J. Gonseth, P. Guallar-Castillon, J. R. Banegas, and F. Rodriguez-Artalejo, —The effectiveness of disease management programmes in reducing hospital re-admission in older patients with heart failure: A systematic review and meta-analysis of published reports,‖ Eur. Heart J., vol. 25, no. 18, pp. 1570–1595, 2004.

12. M. Jessup, W. T. Abraham, D. E. Casey, A. M. Feldman, G. S. Francis, T. G. Ganiats, M. A. Konstam, D. M. Mancini, P. S. Rahko, M. A. Silver, L. W. Stevenson, C. W. Yancy, S. A. Hunt, M. H. Chin, H. F. W. Comm, and W. C. Members, —2009 focused update.

13. Sellappan, P., Chua, S.L.: “Model-based Healthcare Decision Support System”, Proc. Of Int. Conf. on Information Technology in Asia CITA’05, 45-50, Kuching, Sarawak, Malaysia, 2005

14. Blake, C.L., Mertz, C.J.: “UCI Machine Learning Databases”, http://mlearn.ics.uci.edu/databases/heart-disease/, 2004.