ISSN:0975-9646

Diagnosis of Heart Disease Using Data Mining Algorithm

Deepali Chandna

Department of Computer Science & Engineering
IFTM University, Moradabad(UP)

Abstract -In health concern business, data mining plays a significant task for predicting diseases. Numeral number of tests must be requisite from the patient for detecting a disease . However using data mining technique can reduce the number of test that are required. Cardiovascular disease is the principal source of deaths widespread and the prediction of Heart Disease is significant at an untimely phase. In order to reduce number of deaths from heart diseases there have to be a quick and efficient detection technique. The principle of this study is, hence to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases, from a data collected together by an International Cardiovascular Hospital.

Keywords
Data Mining, Heart Disease, k-nearest neighbour, ANFIS, information gain.

1.0 INTRODUCTION

The advancement of information technology, system integration as well as software development, techniques have shaped a innovative generation of multifaceted computer systems. Information technology researchers have been offered several challenges by these systems. An instance of such system is the healthcare system. Newly, there has been an enlarged awareness to make use of the advancement of data mining technologies in healthcare systems. Consequently, the objective of the present effort is to explore the aspects of making use of health data for the assistance of humans by means of new machine learning and data mining techniques. The thought is to recommend an computerized method for diagnosing heart diseases based on prior data and information.

Data mining is a discipline to realize knowledge from databases. The database contains a set of instances (records or case). Machine learning can be defined as a scientific field so as to plan and develop algorithms that let computers to enhance acquaintance of real time problem based on earlier statistics, and perform to resolve a real time problem beneath definite instructions and rules. At hand there are numerous presentations of machine learning; data mining is the largely used application of machine learning. Every illustration used by machine learning and data mining algorithms is formatted by means of same set of fields (features, attributes, inputs, or variables). When the instances contain the correct output (class label) then the learning process is called the supervised learning. On the other hand, the process of machine learning without knowing the class label of instances is called unsupervised learning. Clustering is a common unsupervised learning method (some clustering models are for both). The objective of clustering is to describe data. On the other hand, classification and regression are predictive methods. In the present research, my focus is on supervised machine learning.

This thesis proposes new methods intended for investigating feature selection techniques as well as develop new machine learning algorithms designed for providing automatic computer aided analysis and decision support system for heart disease diagnosis. The aim is to build up an integrated structure with a righteous workflow (constructing missing features values, feature selections, and classification algorithms).

In requisites of features selection techniques, the research decided on features selection technique as a process to increase high superiority attributes to improve the mining process.

In regards to analysis approach the present work projected a new means for diagnosis based on a combination of learning algorithm and feature selection technique. The thought is to get hold of a hybrid incorporated approach so as to merge the most excellent performing learning algorithms and the finest performing feature selection technique by means of an experimental estimate on the dataset obtained from UCI (University of California, Irvine C.A) Centre for machine learning and intelligent systems.

2.0 LITERATURE SURVEY

In this segment, we reassess the existing literature and confer about different aspects of data mining applications in prediction of heart diseases.

In Year 2011, A.Q. Ansari et. al. [02] performed a work**, "Automated Diagnosis of Coronary Heart Disease Using Neuro-Fuzzy Integrated System"**. In this paper, the author offered a Neurofuzzy integrated system for the analysis of heart diseases. To show the effectiveness of the projected system, Simulation for computerized diagnosis is performed by means of the realistic causes of coronary heart disease. The author concluded that this kind of system is suitable for the identification of patients with high/low cardiac risk.

In year 2011 Mrs.G.Subbalakshmi et. al. [02] performed a work "**Decision Support in Heart Disease Prediction System using Naive Bayes**" published in 2012**.** The main objective of this research is to develop a Decision Support in Heart Disease Prediction System using Naïve Bayes algorithm. The system extracts hidden useful information from the heart disease database. This model may possibly answer difficult queries, each one with its own potency with respect to ease of model analysis, access to complete information and accurateness. This model can be further enhanced and expanded by incorporating other data mining techniques.

In year 2012 Mai Shouman, Tim Turner, and Rob Stocker et. al. [03] performed a work "**Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients**". In this paper the author details work that applied KNN on a Cleveland Heart Disease dataset to investigate its efficiency in the prediction of heart disease. The author also investigated if the accuracy could be enhanced by integrating voting with KNN. The results show that applying KNN achieved an accuracy of 97.4% . The results also show that applying voting could not enhance the KNN accuracy in the diagnosis of heart disease.

In year 2013, S. Vijiyarani et. al. [04] performed a work, **"An Efficient Classification Tree Technique for Heart Disease Prediction"**. This paper analyzes the classification tree techniques in data mining. The classification tree algorithms used and tested in this work are Decision Stump, Random Forest, and LMT Tree algorithm. The objective of this research was to compare the outcomes of the performance of different classification techniques for a heart disease dataset. This work is done by using Waikato Environment for Knowledge Analysis(WEKA). It is open source software which consists of a collection of machine learning algorithms for data mining tasks.

In year 2013 Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg et. al. [05] performed a work **"Data Mining in Clinical Decision Support Systems for Diagnosis, Prediction and Treatment of Heart Disease"**. With the help of this study the author concluded that there is large amount of data available in medical institutions, but this data is not properly used. This medical data lacks in the quality and completeness because of which highly sophisticated data mining techniques are required to build up an efficient decision support system. The studies reveal the fact that the systems should be built which not only are accurate and reliable but also reduce cost of treatment and increase patient care. Also the build systems should be easy to understand to enhance human decisions. The author also suggested that work should be done for proposing treatment plans for patients because data mining techniques have shown significant success in prediction and diagnosis of diseases and especially heart diseases, hence these techniques could be applied for treatment purposes also.

In year 2013 Ashish Kumar Sen, Shamsher Bahadur Patel, Dr. D. P. Shukla et. al. [06] performed a work "**A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level**" .In this work, the author has designed a system which could identify the chances of a coronary heart disease. He has divided all parameters into two levels according to criticality of the parameter and assigned each level a separate weightage. Finally both the levels are taken into consideration to arrive a final decision. The author has implemented neuro-fuzzy integrated approach at two levels. So, error rate is very low and work efficiency is high. The author concluded that this same approach could be used to perform the analysis on some other diseases also.

3.0 PRESENT WORK

In this research, our method involves different data mining processes as shown in figure 1.

Data used for current effort is obtained from UCI (University of California, Irvine C.A) Centre for machine learning and intelligent systems. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them.

The aim after feature selection is to decide on a subset of attributes by ignoring features with less significant information. In the present research, feature selection methods are used to lessen the amount of features in the dataset prior to initiation of the mining method.

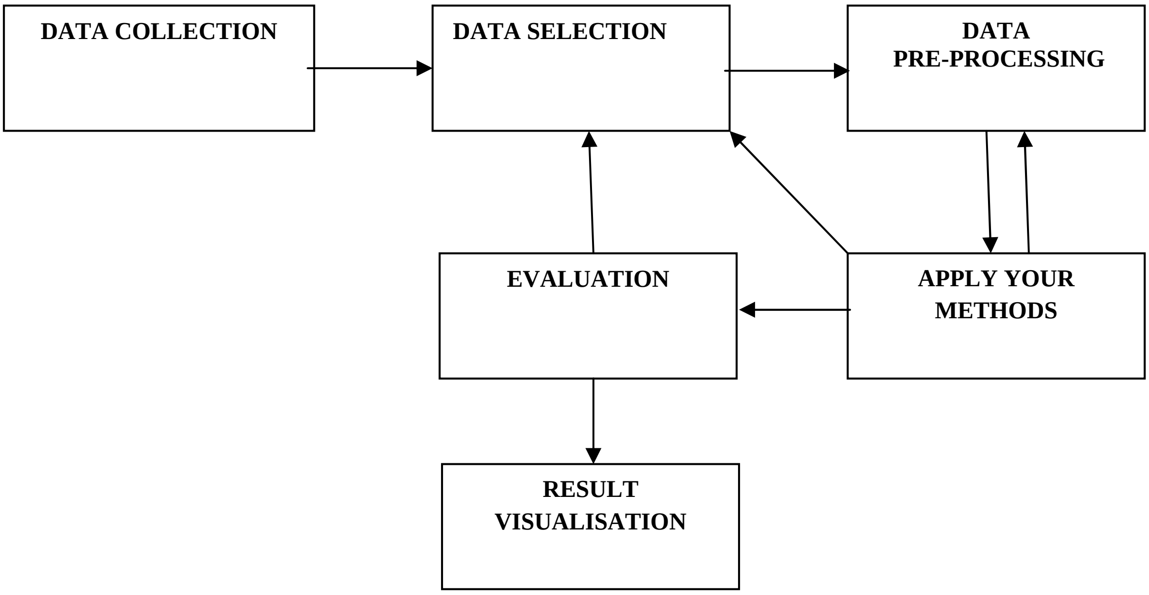

Figure 1

The data set obtained by the data selection phase may contain incomplete, inaccurate, and inconsistence data. Data pre-processing is an essential step in data mining process to assure superiority data elements. The planned approach uses the weighted **k- nearest neighbour's algorithm**. The most important thought is to spread the classification accurateness to a certain threshold set by the researchers and users. The planned scheme showed minor enhancement of 0.005 classification accuracy on the new dataset with no missing values, than the original dataset which contain some missing features values. The data must be prepared for the mining process , at the end of the current phase.

This work projected a system that uses method called **Information Gain** and **Adaptive Neuro-Fuzzy Inference System** for heart disease diagnosis.

The information gain method was proposed to calculate approximately superiority of each one attribute by means of the entropy by estimating the differentiation among the prior entropy and the post entropy. The information gain method is one of the simplest attribute ranking methods and is frequently used in text categorization.

Adaptive Neural Fuzzy Inference System (ANFIS), projected by Jang in 1993, is a grouping of two machine learning approaches: Neural Network (NN) and Fuzzy Inference System (FIS).

The proposed approach is to join the information gain method and ANFIS method for the analysis of diseases (in this case; heart diseases). The information gain will be used for selection of the quality of attributes. The production of applying the information gain method is a set of characteristics with high ranking values and this set of high ranked characteristics will be the input for ANFIS. The selected characteristics will be applied to ANFIS to train and test the planned approach. The arrangement of the proposed approach is shown in Figure 2, where X = {x1,x2,..,xn} are the new features in dataset, Y = {y1,y2,..,yk} are the features subsequent to applying the information gain (features selections), and Z denotes to the concluding output subsequent to applying Y on ANFIS (the diagnose).

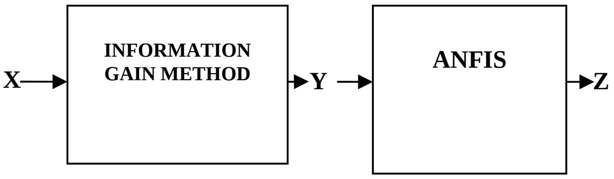

Figure 2

This is an approach for heart disease diagnosis that uses the advantages of Adaptive Network based Fuzzy Inference System (ANFIS) and the Information Gain method. In this approach, the ANFIS is used to construct an input-output mapping using together the human knowledge and machine learning ability. The information gain method is used to decrease the number of input features to ANFIS.

In this study, the evaluation of projected methods is performed by comparing the results with the real data values. According to that, the classification accuracy and error rate are calculated. The error rate (Err) of the classifier is defined as the average number of misclassified samples divided by the total number of records in the dataset. On the other hand, the classification accuracy of the model can be calculated as one minus the error rate. If the classification accuracy is less than a certain threshold, then some changes has to be perform to the method, the feature selection, or the pre-processing phase until obtaining satisfying outcomes.

The current work has used two well-known machine learning tools; WEKA and MATLAB.

4.0 CONCLUSIONS

The present work showed how information gain method, feature selection technique, can be used in collaboration with adaptive neuro fuzzy inference systems in diagnosing new patient cases. The combination created a new approach for diagnosing the breast cancer by reducing the number of features to the optimal number using the information gain and then applied the new dataset to the adaptive neuro fuzzy inference system (ANFIS). The study found that the accuracy for the proposed approach is 98.24% compared with other methods. The proposed approach showed a very promising results which may lead to further attempts to utilise information technology for diagnosing patients for heart diseases.

The current research resided mainly on classification accuracy as the main criteria for measuring the performance of proposed approaches. However, future work will focus in other criteria such as classification speed and computational cost. Future work can also broaden disease options.

REFERENCES

[1] Mrs. G.Subbalakshmi, Mr.M.Chinna Rao "Decision Support in heart disease prediction system using naïve bays",IJCSE Indian journal of computer sciense and engineering, ISSN : 0976-5166 Vol. 2 No. 2 Apr-May 2011

[2] A.Q. Ansari et. al.," Automated Diagnosis of Coronary Heart Disease Using Neuro-Fuzzy Integrated System", 2011 World Congress on Information and Communication Technologies 978-1-4673-0125-1@ 2011 IEEE (pp 1383-1388)

[3] Mai Shouman, Tim Turner, and Rob Stocker, "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients", International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012

[4] S. Vijiyarani et. al., "An Efficient Classification Tree Technique for Heart Disease Prediction" ,International Conference on Research Trends in Computer Technologies (ICRTCT - 2013) Proceedings published in International Journal of Computer Applications (IJCA) (0975 – 8887), 2013

[5] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg," Data Mining in Clinical Decision Support Systems for Diagnosis, Prediction and Treatment of Heart Disease", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)Volume 2, Issue 1, January 2013

[6] Ashish Kumar Sen, Shamsher Bahadur Patel, Dr. D. P. Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level", International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 9 Sept., 2013 Page No. 2663-2671