A REVIEW OF THE MODERN DATA MINING TECHNIQUES FOR THE MULTIPLE DISEASE PREDICTION

Mahendra Sharma[1],Student, Yashashvi Hirwale[2], Student, Neeraj Mehta[3] Assistant Professor
Student, Institute Of Engineering And Science, IPS Academy, Indore
Email Addresses: [1]mahendra.ms05@gmail.com, [2]yashashvi.hirwale1@gmail.com,
[3]neerajmehta@ipsacademy.org

***Abstract*— Data Mining is an attractive and impressive tool for obtaining valuable knowledge from the vast amount of available data that can be used further for taking right decisions. A number of means and approaches are available for deriving profitable outcomes from the possible data. Mining of data by applying if-else rule has conventionally used for the objective of reveling rules in medical applications. The detection of distinct disease such as diabetes, heart attack etc. from large number of guess and evidences is a subject of great interest for the researchers which is not free from false assumption and unpredictable outcomes. Hence there was terrific requirement to utilize the valuable conclusions resulting from the information of the patients gathered in our data storehouse. Many predictors had worked for improving the result of the disease prediction systems using the approach of data mining. The systems are designed for predicting single as well as different diseases still there is probability for improving the result of the currently used disease prediction system in terms of accuracy and efficiency. Here, we have presented an analysis of current data classification algorithms and will encounter the chances of occurrence of distinct diseases on the basis of data stored in our database.**

***Index Terms*— Data Mining, Disease Prediction, Decision Tree, KDD**

I. INTRODUCTION

The efforts made by the analyst to establish advance and influential tool for converting accessible data into valuable and beneficial knowledge have lead to the evolution of a unique exploration area known as data mining or knowledge discovery from data.

Data mining is a multifaceted field inferencing its perception from diverse fields. Hospitals usually aggregate a large quantity of data over some years. These data is used to contribute as a support for the review of the threat aspects for different diseases. For example we can find out the level of causing cancer to find common arrangements linked with the disease. One of the important aspect of Data mining is to provide the venture with sense. One of the prominent function of data mining is the inspection of effective and fascinating arrangements in the data. A majority of areas related to medical field such as the Family and General Practice, public healthcare, surgery, research department and the discovery of relations between the medical and examination data also make the use of the data mining techniques Data mining is capable of delivering the analysis of the disease which disease can be caused by evaluating the symptoms and assumptions and can contribute better results in future for predicting the risk analysis of the diseases. In this paper, we presented prediction of the risk levels of multiple diseases from the Disease database. The database consists of mixed attributes containing both the numerical and categorical data. The records are first cleaned and filtered with the intention that the irrelevant data can be removed from the database before the mining process starts and then the classification of data takes place. A number of classification algorithms can be used to classify the data. Classification is a form of data analysis that extracts models describing important data classes. These models are known as classifiers that are used to predict categorical labels. For performing the task of classification we are going to make the use of an updated ID3 algorithm to show the effective risk factor with decision tree based classification algorithm. The remaining section of this paper is organized as follows section 2 deals with a review of the previous work that has been done on the disease prediction system. Section 3 deals with how the fragmentation of risk level and the extraction of the significant patterns from the database. Section 4 deals with the proposed model and experimental outcomes. The conclusion obtained is depicted in section 5.

I. SYSTEM MODEL

In this study the concept of knowledge is used to develop a concise disease prediction system.. The knowledge discovery process basically consist of nine steps. The steps are repetitious and interactive, with many decisions made by the user. The process is iterative means we are provided with a way to switch to the previous state if a mistake is committed.

The procedure starts initially with determination of the goals that are to be achieved, and lastly ends with the worthy usage of the discovered knowledge. Based on the KDD methodology, the nine steps were undertaken as:

A. Knowing the application area

In order to specify the problem faced and to describe the medical goals, the researcher has worked closely with the hospital's departmental head. The discussion with the specialists and experts had helped out the researchers to better understand the problem and to know about ways that are used as the solution for the problems. An important objective in this process is determination of data mining goals and their success implementation of these goals

B. Electing and Establishing the Target Dataset

Hospitals usually accumulate a huge amount of data and maintains a separate file for the records of each patient. As these records are stored in a separate word file . so we need to integrate the important records at same place as they are used for effective for prediction of diseases.

C. Preprocessing and Clarifying data

The target data is checked for distorted, incompatible and misplaced values using the concept of outlier analysis. Misplaced values and incompatibility determined in the dataset are corrected, and misplaced values were exchanged with the most probabilistic value

D. Data conversion process

In data consolation, a few conversions in the data are made so to make it more convenient for the mining process. The conversion is carried out to make the process fast and reliable and the data is converted into form that is appropriate for the mining process.

.

E. Selecting the appropriate Data Mining Technique

Here it is decided that which data mining technique is suitable for achieving the defined goals which is the result of our study. As it is one of the essential step so different available techniques are evaluated to obtain the desired results and accurate prediction of diseases.

F. Selection of the Data Mining Algorithm

The data classification algorithms are compared and evaluated based on their performance on the dataset. The assessment procedure is considered to find the best algorithm for achieving the defined aim. The applicability of the algorithm in its application area and the structure of data also play an important role in the assessment of the algorithm.

G. Implementing the Data Mining Algorithm

For implementing the best classification algorithms that had been decided on the basis of their performance experiments are designed and these experiments are performed on the data set

H. Assessment of the system

For assessment purpose we evaluate the proposed model with different models so to find out that whether it met our proposed data mining goals which were identified at the first step of the complete process. . The steps that are involved in this process includes understanding the results, identifying if any new information comes then it is novel and useful for us, analysis of the results, checking their effects and the fulfillment of the defined objectives.

II. PREVIOUS WORK

A lot of work has been done by the researchers for the prediction of both single and multiple diseases. The methods that were used previously for determining similar and beneficial patterns in data include Navie Baye's theorem and regression analysis . This method has laid to the discovery of other unique methods in data mining such as neural networks, clustering genetic algorithms , decision trees and support vector machines.

Hnin Wint Khaing had represent an efficient approach to predict the risk factor of causing the heart attack from the heart disease database by making the use of data mining techniques [1].

S.Vijendra had discussed about the shortcomings of the ID3 algorithm in choosing the attributes and had proposed the improved version of the ID3 algorithm. In his proposed algorithm attributes are divided into groups and then he had applied the selection measure 5 on these groups to improve information gain and these process is repeated until good classification ratio is achieved

The various data mining classification techniques and about the regression tree had been discussed by Breiman et al. [3] in his paper. CART is a 3-2 nonparametric technique that can select from among a large number of variables those and their interactions that are most important in determining the outcome variable to be explained

C. Ruimin and W. Maio had proposed a more accurate nad efficient algorithm for ID3[4] . In these approach they made the use of a greedy technique to select the best attribute which has maximum information gain.

In [5] X.nuinui et al. had proposed an improved ID3 algorithm that uses the concept of information gain to effectively predict the risk factors.

A detailed review on data mining concepts and the techniques of data mining had been discussed by kamber in his paper [6].

II. PROPOSED METHODOLOGY

The wide acceptance of using the relational based approach for handling data in various commercial and data processing applications and their continuous improvement and advancement in the relational databases had made people to use the relational databases in non-commercial applications too. The database is used in our approach to store the collected data. The result of our analysis is discussed in this section.

With the help of the database the patter significant for the prediction of various disease are mined using the defined approach. The database is first preprocessed which result in the removal of inconsistent and incompatible values and the misplaced values are replaced with the most frequently occurring ones. The classified dataset is then clustered and the unique patterns are mined and the updated ID3 algorithm is applied on it

In our proposed technique we had made the use if a greedy method to select the fittest attribute from all the attribute using which we can divide the dataset into smaller partitions. The proposed greedy method also uses the concept of information gain. our proposed algorithm also chooses the attribute with highest information gain that is similar to the concept of ID3 but we have modified the formulae of information gain. The modified formulae contain utility value of each attribute. In this the selection criteria has improved, which ultimately will result is more classification and prediction. The Expected outcomes are as follows:

- Classification accuracy will be improved.
- Time and space consumption will be reduced.
- Error rate will be reduced.

III. EXPERIMENTAL RESULT

The experiment is carried out on a publicly available database for heart disease. The dataset contains total 573 records. The dataset is divided into 2 sets training (303 records) and testing set (270 records). A data mining tool Weka 3.6.6 is used for experiment. Parameters used for experiment are listed below.

Patient ID: Patient Identification number.
Diagnosis: Value 1: =<50% (no heart disease)

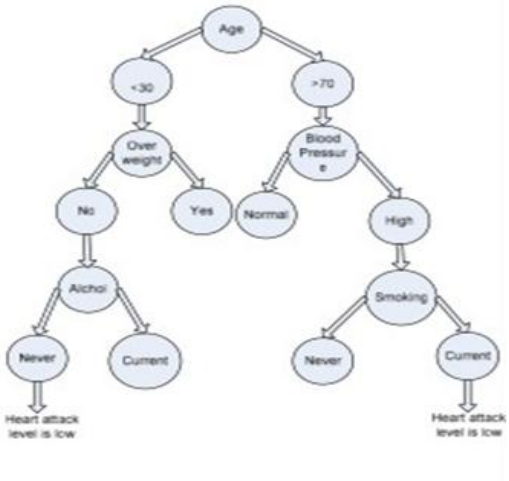

Figure 1: A decision tree for the concept heart attack level by information gain (updated ID3)

The experimental result of our approach is presented. The goal is to have high accuracy, besides high precision and recall metrics. These metrics can be derived from the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics.

TP (True Positive): It denotes the number of records classified as true while they were actually true.

FN (False Negative): It denotes the number of records classified as false while they were actually true.

FP (False Positive): It denotes the number of records classified as true while they were actually false.

TN (True Negative): It denotes the number of records classified as false while they were actually false.

IV. CONCLUSION

The decision tree based approach serve as a more suitable way to make right conclusions about the diseases in a precise and adequate manner as they are capable of representing numerical and categorical dataset. The data mining approaches are commonly used for the prediction of multiple diseases The outcome of this work is very important for the medical field as it may help the patients and doctors up to a great extent. To improve the accuracy of the proposed system further research should be conducted using different data mining classification algorithms like- rule based induction, expectation maximization, etc. In this paper, a critical review of the modern decision tree based technique has been performed. Their working is discussed with advantages and disadvantages in brief.

V. FUTURE SCOPE

In our future work, we have had planned to design and develop an efficient heart attack prediction system with the aid of xml data using X-Parser and X-Query language.

REFERENCE

[1] HW Khaing –" Data mining based fragmentation and prediction of medical data" Research and Development (ICCRD), 2011 3rd …, 2011 - ieeexplore.ieee.org.

[2] Singh Vijendra. Efficient Clustering For High Dimensional Data: Subspace Based Clustering and Density Based Clustering, *Information Technology ""Journal; 2011*, 10(6), pp. 1092-1105.

[3] D Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J."Classification and Regression Trees". Wadsworth International Group. Belmont, CA: The Wadsworth Statistics/Probability Series1984.

[4] Chai Rui-min and Wang Miao, "A more efficient classification scheme for ID3",Sch. of Electron. & Inf.

Eng., Liaoning Tech. Univ., Huludao, China; 2010,Version1, pp. 329-345.

[5] Huang Ming, Niu Wenying and Liang Xu , “An improved Decision Tree classification algorithm based on ID3 and the application in score analysis”, *Software Technol. Inst*., Dalian Jiao Tong Univ., Dalian, China, June 2009.

[6] Iu Yuxun and Xie Niuniu “Improved ID3 algorithm”,*Coll. of Inf. Sci. & Eng*., Henan Univ. of Technol., Zhengzhou, China;2010,pp. ;465-573.

[7] Jiawei Han and Micheline Kamber**, “***Data Mining: Concepts and Techniques*”, 2nd edition, Morgan Kaufmann, 2006, ch-3, pp. 102-130.

[8] Chen Jin, Luo De-lin and Mu Fen-xiang,” An im pr oved ID3 decision tree algorithm”,Sch. of Inf. Sci. & Technol., Xiamen Univ., Xiamen, China, page; 2009, pp. 127-134.

[9] Quinlan, J. R. “Induction of Decision Trees”. *Machine Learning;* 1986,pp. 81-106.

[10] Quinlan, J. R. Simplifying “Decision Trees. International Journal of Man-Machine Studies" ;1987, 27:pp. 221-234.

[11] Gama, J. and Brazdil, P. “*Linear Tree. Intelligent Data*Analysis”,1999,.3(1): pp. 1-22.

[12] Langley, P. “*Induction of Recursive Bayesian Classifiers*”. In BrazdilP.B. (ed.), Machine Learning: ECML-93;1993,pp.153-164.
Springer,Berlin/Heidelberg~lew York/Tokyo.

[13] Witten, I. & Frank, E,"*Data Mining: Practical machine learning toolsand* techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.ch. 3,4, pp 45-100.