ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

A Comparative Study of Classification Algorithms for Disease Prediction in Health Care

Isha Vashi[1], Prof. Shailendra Mishra[2]

M.Tech Student, Dept. of CSE, Parul University, Vadodara, Gujarat, India[1]

Assistant Professor, Dept. of CSE, Parul University, Vadodara, Gujarat, India[2]

ABSTRACT: The HealthCare Organization collects large amounts of healthcare information which can not be mined to find unknown information for efficient result. Now a days, Health Services has been converted from an offline paper to online system. This online system consists patients' personal and medical information. Data Mining methods can help to find successful analysis methods and connections and hidden patterns from that patients' information and large volume of data. Decision tree classification algorithms are suitable and popular methods for the medical diagnoses problems. This paper presents a survey of various decision tree classification algorithm for disease prediction in E-Health environment and introduces the reader to the most well known classification algorithms that can be used to predict disease.

KEYWORDS: Health Care; Diseases Prediction ; Classification algorithm; Data Mining ;Decision tree representation

I. INTRODUCTION

Now a days, Health organization has a large volume of data that needs to be collected and stored , such as patients' information , Laboratory results , treatment results and much more. Healthcare Organizations has a large volume of data that requires automatic way for these data to be extracted when needed.

Integration of Data Mining techniques with health systems would help to improve the efficiency and effectiveness of healthcare organizations. Healthcare industries can reduce cost by using computer based data and decision support system.

The main aim of this paper is to find out best classifier from different classification algorithm that can be used to predict disease on applying patients' data.

Data is the most important asset in the latest information economy. Mining useful information from large collection of data is critical task for all organizations. Information technology is being increasingly implemented in medical industry to responds to the needs of doctors in their daily decision making process. Data mining tools are very useful to control limitations of people such as subjectivity or error due to fatigue, and to provide indications for the decision making process. Data mining consists number of techniques that can be used to understand and analyze data in order to find hidden patterns and connections that would make it easy for healthcare industry to make decision based on the knowledge.

Data mining preprocessing and transformation process is required before one can apply their data mining technique to clinical data. Without Data mining, it is difficult to understand the full potential of data collected through various resources.

Data mining is an important step of KDD i.e. knowledge discovery from database. KDD consists of an iterative sequence of Data cleaning, Data integration , Data Mining pattern recognition and knowledge presentation. Data mining can be accomplished by using Classification , Clustering , Prediction , Association and Time series analysis.

This paper is organized as follows : The proposed model of adopting different data mining classification algorithms on a patients' data set is explained in section 2. Section 3 consists of different classification algorithms and its evaluation . Conclusion of this study is presented in section 4.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

II. RELATED WORK

In [1] ,Lakshmi B.N. propose a novel health monitoring approach to predict risk and accuracy for prediction on applying pregnant women's dataset by C4.5 algorithm. Applying C4.5 algorithm on pregnant woman dataset, accuracy for risk prediction can be improved. As a result , C4.5 algorithm gives accuracy percentage is 71.9043 and error percentage is 28.6957 respectively and correctly classified datasets are 164 and incorrectly classified instance are 66 out of total 230 test data. It shows C4.5 classifier can be used to predict disease on patients' data set.

In [2] ,Mennat Allah Hassan compares different classification algorithm to obtain best classifier for prediction of disease in E-Health environment. In this case study , ten classification algorithms like, Bayes Net, Logistic, K star, stacking , JRIP , One R, PART , J48, LMT and Random forest are considered to predict disease and based on performance metrics such as training time, confusion matrix , Precision etc. After classification results, conclusion is bayes net is the best classifier among all classifiers. Therefore this can aid in the decision making process and prediction making it more efficient.

In [3] ,Sankaranarayanan.S , Dr Premananda Perumal.T suggest an approach to predict disease through ID3 and C4.5 algorithm. In this approach, using classification algorithm, diabetes can be predicted on applying patients' data. In this approach, rule classification method and decision tree representation is used to obtain prediction and result shows that decision tree method gives efficient evaluation to predict disease.

In [4] ,Lakshmi B.N. compares C4.5 classification algorithm and Naive bayes classifier on applying pregnant women dataset to obtain risk prediction with accuracy. Comparison of algorithm shows that C4.5 classification algorithm gives best result in compare to naive bayes classifier. As C4.5 algorithm gives 152 correctly classified instances and naive bayes classifier gives 137 correctly classified instances. C4.5 algorithm gives 74.2838% relative absolute error and naive bayes classifier gives 99.4254% relative absolute error. Result shows that in all parameters , C4.5 gives efficient result compare to naive bayes classifier.

In [5] ,Monika Gandhi proposed approach to predict heart disease using different data mining techniques. Author suggests classification methods for prediction of heart diseases. Decision tree representation , Neural network and Naive Bayes Classifier are suggested to predict diseases from patients' dataset. According to this case study , Decision tree representation can be helpful to implement an algorithm in healthcare organizations.

In [6] ,M.A. Nishara Banu proposed diseases forecasting system using data mining methods. In disease forecasting system , k-means algorithm , ID3 algorithm and C4.5 algorithm is implemented to achieve highest accuracy. The result of experimental analysis shows that with integration of ID3 and C4.5 algorithm with K-means algorithm gives greater accuracy in compare to K-means algorithm.

III. PROPOSED MODEL

Mining patients' data to predict disease or to make decision by using patients' personal and medical information , it requires steps to be followed. Steps for proposed model are as shown in figure 1.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

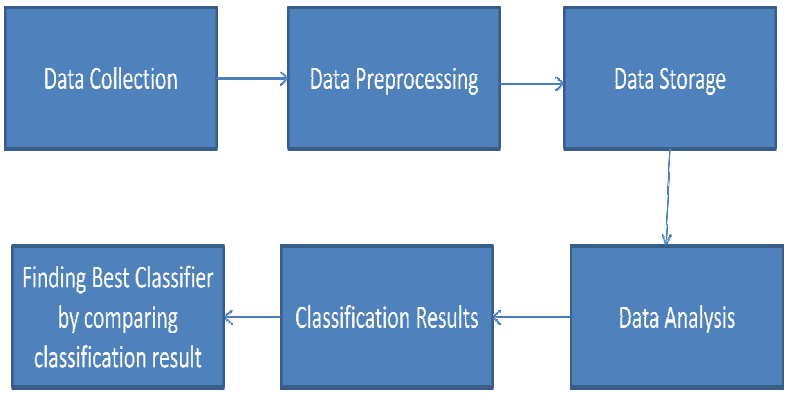

Figure 1 : Proposed Model

The steps undertaken to apply data mining algorithms shown in figure 1 are presented as follows-

a) Data Collection:
The First step consists data collection from healthcare organizations such as hospitals, laboratories and medical centers. This data consists personal data about the patients such as patient name , age , address , height , weight and medical information such as blood group , blood pressure level , sugar level , and symptoms that they had such as high fever , headache , etc. and their laboratory results.

b) Data Preprocessing:
The Second step encloses transforming and preprocessing of data. Data collected from healthcare organizations obtained into one single form understandable by data mining tool. Data comes from different resources each with its own different form. This different form of data needs transformation and preprocessing. This transformation and preprocessing consists following steps –
- Data Cleaning,
- Matching,
- Combining,
- Removing noisy and relevant Data
- Standardization.

c) Data Storage:
The third step consists data storage process. In data storage, transformed data is stored into a single database with the same format. So, this data can be used to apply data mining techniques on.

d) Data Analysis:
After data storage, next step is data analysis. Data analysis is most important phase in this proposed model. It encloses following procedure: First, It includes applying data mining techniques or classification algorithms on patients' data being loaded dataset into data mining tool. Second, it computes classification results of algorithm.

e) Classification Results:
Next step of proposed model consists classification result of algorithm and it includes computing the best classifier for the dataset obtained according to the study. After classifying result, comparison of this result will be occurred and according to different parameters of algorithm like accuracy, efficiency, best algorithm will be chosen which will be helpful to doctors, physicians to predict diseases and help them to make decision for patients' data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

IV. DIFFERENT DATA MINING TECHNIQUES FOR DISEASE PREDICTION

After the data set has been prepared it has been uploaded to the data mining tool for classification. The data mining tool used in this research is WEKA tool and version 3.2.13 which was written in java, developed at the University of Waikato, New Zealand in 1999 for knowledge analysis. WEKA stands for Waikato Environment for Knowledge Analysis.

A. *USED INTERFACE IN WEKA TOOL*

In this research, WEKA's main interface, explorer is used. Explorer interface has several panels like pre-process, classify, associate, cluster, select attributes and visualize. In this study, classification panel is the one focused on and applied on patients' dataset. In classification, patients' data are divided into two set, training data and test data. The accuracy of the classification of each algorithm is tested through the ten-fold cross validation which divides dataset into ten folds with equal distribution. In each test nine of these folds are tested as a training set and the remaining fold is used for testing. The test is repeated ten times and the average results are computed.

B. *CLASSIFICATION ALGORITHMS USED*

In the prediction of diseases, we will use following classification methods:

a) Decision Trees
b) Naive Bayes Classifier

a) Decision Tress :

Decision tree learning uses a decision tree as a predictive model used in data mining. In this research, decision tree is used to predict disease from patients' data through classification algorithm. Decision tree consists various algorithm like ID3 algorithm, C4.5 algorithm etc.

Iterative Dichotomized Algorithm: ID3 is an algorithm was developed by Ross Quinlan used to generate decision tree from given dataset. ID3 algorithm produces decision tree using Shannon entropy. This algorithm consists information gain and entropy to generate Decision tress. ID3 algorithm consists for short decision tree out of set of learning data and shortest is not best Classification. Due to this limitation, it is succeeded by Quinlan's C4.5 and C5.0 algorithm.

C4.5 algorithm: c4.5 is an algorithm is used to generate decision tree using information gain in the same way as ID3 algorithm developed by Ross Quinlan as a successor of ID3 algorithm. It is used for great volume of data so that it will be helpful to generate best classification and consists for large decision trees. C4.5 algorithm can handle missing attribute value and handles attributes with different cost. This algorithm is easy to understand as compare to ID3 algorithm.

b) Naive Bayes Classifier :

Naive Bayes Classifier is a simple technique for classifier that depends on Bayes' theorem with independent assumptions. It provides data structure and facilities common to Bayes network learning algorithms. An advantage of Naive Bayes Classifier is that it only requires a small number of training data to estimate the attributes for classification.

V. CONCLUSION AND FUTURE WORK

This paper presents various classification algorithms to predict diseases in Healthcare. Different classification algorithms are introduced to predict diseases on applying patients' dataset. Different classification algorithm gives different result on base of accuracy, training time, precision etc. To get efficient result, comparison is needed

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

between these algorithms and we can decide best algorithm among all of them. As paper presents , two classification methods , decision tree representation and naïve bayes classifier , conclusion of paper after referring both methods and existing research paper is that Decision tree representation gives efficient result , greater accuracy and less training time compare to naïve bayes classifier. According to this survey, we can conclude that the most efficient method for disease prediction is Decision Tree Representation method. In future work, different data mining technique can be used with decision tree classification algorithm to improve performance and high accuracy.

REFERENCES

1. Lakshmi.B.N,Dr.Indumathi.T.S.,Dr.Nandini Ravi,'A Novel Health Monitoring approach for pregnant women', International Conference on emerging Reseach in Electronics , Computer Science and Technology-2015,978-1-4673-9563-2/15,pp.324-328,2015.
2. Mennat Allah Hassan, M.Elemam. Shehaband Essam, M.Ramzy Hamed,' A Comparative Study of Classification Algorithms in E-Health Environment', ISBN: 978-1-4673-7504-7 ©2016 IEEE, pp.42-47, 2016.
3. Sankaranarayanan.S, Dr Pramananda Perumal.T,' A Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies', 2014 World Congress on Computing and Communication Technologies,pp.231-233,2014.
4. Lakshmi.B.N,Dr.Indumathi.T.S,Dr.Nandini Ravi,' A Comparative Study of Classification Algorithms for Risk Prediction in Pregnancy', 978-1-4799-8641-5/15/$31.00 ©2015 IEEE,pp.251-256,2015.
5. Monika Gandhi, Dr. Shailendra Narayan Singh,' Predictions in Heart Disease Using Techniques of Data Mining', 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015),pp.520-525,2015.
6. M.A.Nishara Banuand, B.Gomathy,'Disease Forecasting System Using Data Mining Method', 2014 International Conference on Intelligent Computing Applications,pp.130-133,2014.
7. Dr.B.L.Shivakumarand, S. Alby,' A Survey on Data-Mining Technologies for Prediction and Diagnosis of Diabetes', 2014 International Conference on Intelligent Computing Applications,pp.167-173,2014.
8. Sankaranarayanan.S, Dr Pramananda Perumal.T,' Diabetic prognosis through Data Mining Methods and Techniques', 2014 International Conference on Intelligent Computing Applications,pp.162-166,2014.
9. Ramana Nagavelli, Dr.C.V.Guru Rao,' Degree of Disease Possibility (DDP): A mining based statistical measuring approach for disease prediction in health care data mining', IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), May 09-11, 2014, Jaipur, India.
10. Boris Milovic, Milan Milovic,' Prediction and decision making in health care using data mining', Kuwait Chapter of Arabian Journal of Business and Management Review Vol. 1, No.12,pp.126-136; Aug 2012.
11. George-Nektarios, "Weka Classifiers Summary," www.academia.coml5167325/Weka_Classifier_Summary. 2013.
12. Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.
13. Ho, T. J.: "Data Mining and Data Warehousing", Prentice Hall, 2005.

BIOGRAPHY

Isha R. Vashi is an M.Tech. Student in the Computer Science and Engineering Department, Parul University, Vadodara,Gujarat, India and Pursuing Master of Technology (M.Tech) degree from Parul University, Vadodara, Gujarat, India. She received Bachelor of Engineering (B.E.) degree in Computer Science and Engineering in year 2015 from SVMIT, Bharuch, Gujarat, India. Research interests are Data Mining, Web mining, Disease prediction, etc.