Asian Journal of Convergence in Technology *Volume III, Issue III*
ISSN No.:2350-1146, I.F-2.71

A Comparative Study of Classification Algorithms for Diseases Prediction in medical domain

Isha R.Vashi[1]
Computer Science & Engineering Department
Parul University
Vadodara, India.
ishavashi05@gmail.com

Shailendra Mishra[2]
Computer Science & Engineering Department
Parul University
Vadodara, India.
shailendra.mishra@paruluniversity.ac.in

Swapnil Andhariya[3]
Computer Science & Engineering Department
Parul University
Vadodara, India.
Swapnil.andhariya@paruluniversity.ac.in

***Abstract*—The healthcare industry collects great volume of information which cannot be mined to find unknown information for sufficient result. Now days, Health Services has been converted from an offline paper to online application. This online application consists patients' personal and medical information. Data mining methods can help to find successful analysis methods and connections and hidden patterns from those 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 algorithms 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*—Classification algorithm; Data Mining; Decision tree representation; Diseases Prediction; Health Care**

I. INTRODUCTION

Now days, Health organization has a large collection of data that needs to be collected and stored, such as patients' information, Laboratory results, treatment results and much more. Healthcare industries can reduce cost by using computer based data and decision support system.

The main objective of this paper is to find out best classifier from different classification algorithm that can be used to predict disease on applying patients' data. Mining useful information from large collection of data is complicated task for all organizations. 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[1]. Data mining consists number of methods that can be used to find hidden patterns that would make it easy for healthcare organization to make decision based on the information.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information [2]. Data mining brings a set of tools and techniques that can be applied to processed medical data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions [3].

Medical data mining is applying data mining techniques and methods in medical data. The challenge faced by healthcare industry with regard to the massive data-rich but information-poor collection is to extract valuable information to be available at a particular time, place in the form needed to support the decision-making process [4].

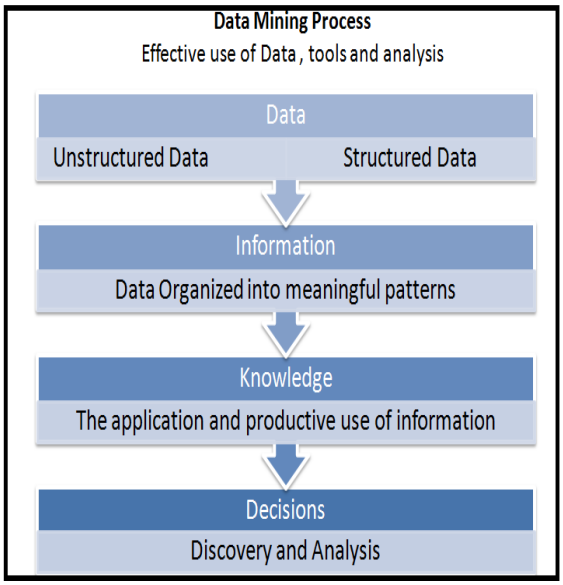

Fig 1: Data mining process

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.

II. LITERATURE SURVEY

Lakshmi.B.N.[5] used C4.5 decision tree classification algorithm to predict risk from pregnant women's dataset. The performance of C4.5 decision tree classification algorithm selected for study to obtained accuracy in risk prediction. C4.5 classification algorithm is implemented on WEKA toolkit. Accuracy and error percentage obtained by algorithm shows the efficiency of proposed method to predict risk.

Mennat Allah Hassan[6] used different ten classification algorithms to predict diseases from patients' dataset. These algorithms are performed on data mining weka tool to optimize best classifier and that classifier can be implemented to predict diseases. True positive rate, false positive rate, Confusion matrix, precision, training time are parameters which are analyzed to obtain best classifier.

Monika Gandhi[7] proposed classification methods like decision tree representation, neural network and naïve bayes classifier to predict heart diseases.

Dr.Indumathi.T.S.[8] proposed an approach to find prediction of disease using C4.5 decision tree classification algorithm. In proposed model, author compares the result of naïve bayes classifier to C4.5 decision tree algorithm. Comparison shows greater efficiency and accuracy can be obtained by C4.5 algorithm.

Sankaranarayanan.S[9] proposed approach to predict diabetes from patients' data through ID3, C4.5 algorithm and rule set classifiers. ID3 and C4.5 algorithm is decision tree algorithm that creates decision tree based on attributes in dataset.

Dr. Nandini Ravi[10] proposed health monitoring approach through C4.5 algorithm to predict gestational diabetes state, premature birth and risk at the stage of pregnancy. Accuracy and training time of dataset are parameters to predict the risk.

M.A.Nishara Banu[11] used K means clustering algorithm with C4.5 classification algorithm to predict heart diseases. In this method, data set is divided into clusters and C4.5 algorithm is applied on each cluster to predict disease on given attributes. Training time, confusion matrix and True positive rate are parameters to find accuracy.

III. PROPOSED WORK

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 work are as shown in fig 2.

The steps shown in fig 2 are described 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 consists transforming and preprocessing of data. Data collected from healthcare organizations obtained into one 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.

c) Data Storage:

The third step consists data storage process. In data storage, transformed data is stored into a one database with the same format.

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 methods or classification algorithms on patients' data being

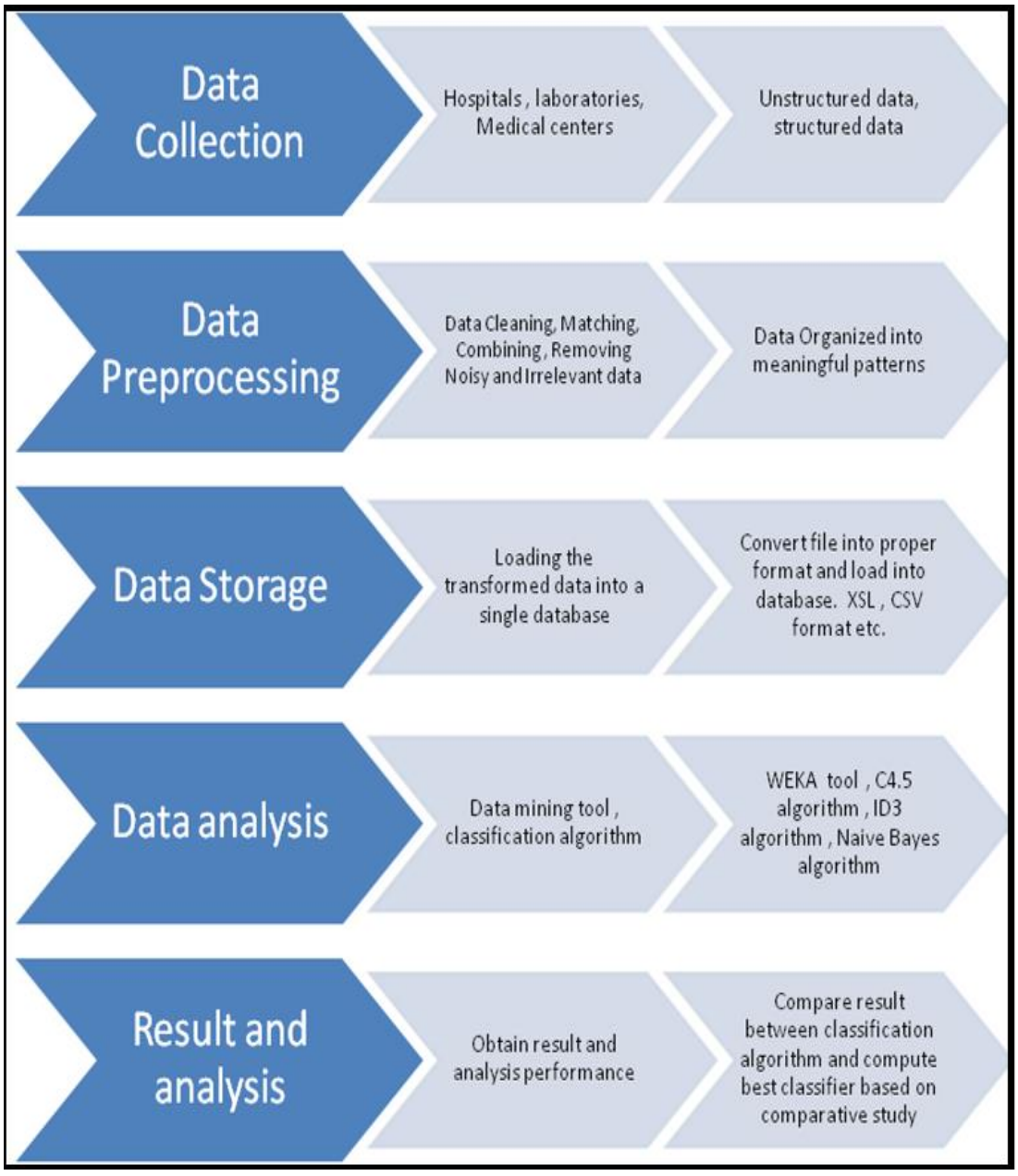

Fig 2: Proposed work

loaded dataset into data mining tool. Second, it computes classification results of algorithm.

e) Obtain Result and Analyse Performance :

Next step consists classification result of algorithm and it includes computing the best classification algorithm for the dataset obtained according to the analysis. 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 to predict diseases and help them to make decision for patients' data.

IV. DIFFERENT DATA MINING TECHNIQUES FOR DISEASE PREDICTION

A. *Use Interface in WEKA tool*

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[13].

In this research, WEKA's main interface, explorer is used. Explorer interface has several panels like preprocess, classify, associate, cluster, select attributes and visualize. In this study, classification panel is the one focused on and applied on patients' dataset[13].

B. *Classification algorithms*

Different classification models which are analyzed for diseases prediction are shown in fig 3.

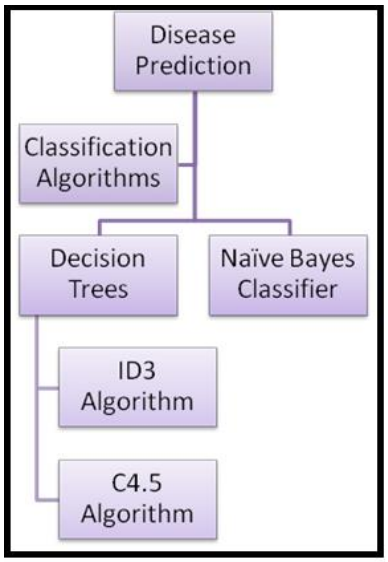

Fig 3: Classification algorithms for disease prediction

In the prediction of diseases, we will use following classification methods of data mining are analysed:

a) Decision Tress :

Decision tree learning uses a decision tree as a predictive model used in data mining. Decision trees are powerful and popular tools for classification and prediction in data mining[1]. Decision tree represent rules which can be understood by humans and used in knowledge system such as database. In this research, decision tree is used to predict disease from patients' data through classification algorithm. There are various classification algorithm are used to represent decision tress. Most preferred algorithms are ID3 and C4.5 algorithm.

1. Iterative Dichotomized 3 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[7]. ID3 algorithm consists for short decision tree out of set of learning data and shortest is not best Classification.

Steps:

a) Build classification attribute from patients' dataset.
b) Compute classification entropy (Shannon entropy).
c) For each one attribute compute information gain.
d) Select attribute with the highest gain in tree. (Beginning from root).
e) Remove node, attribute making decreased table.
f) Repeat steps 3-5 until all attributes have been generated. And at that point, smallest tree is preferred.

Advantages of ID3 algorithm:

- It is easy to understand and interpret.
- Rules are easily generated with ID3 algorithm.
- It consists implicit perform feature selection.
- Allows addition of number of new data.

2. 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[4].

Steps:

a) All the attributes in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.

b) None of the features provide any information gain. C4.5 creates a decision node higher up the tree using the expected value of the class.
c) Instance of previously-unseen class detected. Again, C4.5 creates a decision node higher up the tree using the expected value.
d) For each attribute a, find the normalized information gain ratio from splitting on a.
e) Let a_best be the attribute with the highest normalized information gain.
f) Create a decision node that splits on a_best.
g) Recur on the sub lists obtained by splitting on a_best, and add those nodes as children of node.

Advantages of C4.5 algorithm:

- It handles both continuous and discrete items.
- It handles training data with missing attributes.
- Handling attributes with different costs.
- Pruning trees after creation of decision tree.

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[1].

Bayes' Theorem:

Probability (B given A) = (Probability (A and B) / Probability (A))

Advantages of Naive Bayes Classifier:

- Easy handle of large amount of data.
- Handles real and discrete data.
- Handles streaming data well.

V. CONCLUSION

Data mining gives a lot of techniques to extract hidden pattern from the healthcare industry. This proposed study given an overview of data mining techniques like classification algorithms and tools like WEKA tool. This study analyses advantages of each classification algorithms and provides information about that techniques and also helps to choose suitable classification algorithm for disease prediction. Furthermore, Decision trees or Naïve Bayes classifier can be studied in more detail to implement an algorithm that is helpful in healthcare organizations.

REFERENCES

[1] Ian H. Witten, Eibe Frank and Mark A. Hall, 'Data Mining - Practical Machine Learning Tools and Techniques", Third Edition, Morgan Kaufmann Publishers.

[2] Dr.T.Karthikeyan, Dr.B.Ragavan and V.A.Kanimozhi 'A Study on Data mining Classification Algorithms in Heart Disease Prediction',International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol. 5, Issue 4, April 2016, ISSN: 2278 – 1323.

[3] V.A. Kanimozhi and Dr. T. Karthikeyan, 'A Survey on Machine Learning Algorithms in Data Mining for Prediction of Heart Disease', International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2016, ISSN (Online) 2278-1021 ISSN (Print) 2319 5940.

[4] Arun K Pujari , 'Data Mining Techniques', University press , Edition 2001.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Lakshmi.B.N,Dr.Indumathi.T.S.,Dr.NandiniRavi,'Prediction Based Health Monitoring in Pregnant Women', International Conference on emerging Reseach in Electronics,Computer Science and Technology-2015,978-1-4673-9223-5/15/$31.00_c 2015, IEEE.

[11] M.A.Nishara Banuand, B.Gomathy,'Disease Forecasting System Using Data Mining Method', 2014 International Conference on Intelligent Computing Applications,pp.130-133,2014.

[12] 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.

[13] Svetlana S. Aksenova , 'Machine Learning with WEKA - WEKA Explorer Tutorial for WEKA Version 3.4', 2004.