SJIF Impact Factor: 6.260| ISI I.F.Value:1.241 Volume: 4 | Issue: 10 | October | 2019 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Peer Reviewed Journal

A STUDY ON DATA MINING CLASSIFICATION ALGORITHMS IN VARIOUS DISEASE PREDICTIONS

Mrs.K.Sindhya

Assistant Professor, Department of Information Technology, Nirmala College for Women, Coimbatore, T.N, India

ABSTRACT

Data mining is the process of extracting hidden information from the huge database. Medical domain contains different forms of data like text, numbers and images that can be handled properly to provide the useful information. The patterns obtained from the medical data can be useful for the physicians to detect diseases, predict the accuracy for the treatment. This paper provides the performance of various classification algorithms involved in disease prediction.

KEYWORDS: *Data mining, Classification, Decision Tree*

1. INTRODUCTION

Data mining is most popularly used for getting useful information from high volume raw database. It is used to discovering knowledge out of data and presenting it in a form that is easily understandable by humans. It is process to examine large amounts of data routinely collected. Data mining is more efficient in an exploratory analysis because of nontrivial information in huge amount of data. There are two primary objective of using data mining is: i) Prediction and ii) Description. Prediction involved in various fields to predict the unknown and future values of other variable of interest. But Description mainly focusing of finding patterns describing the data that can be interpreted by humans.

Data mining plays an vital role in medical domain. Data mining techniques are efficient in identifying and predicting various diseases. There are several disease can be predicted by using data mining namely breast cancer, Cardiovascular Disease, Kidney disorders, Heart disease, thyroid diseases etc.. This paper analyzed the performance of the classification algorithms in various disease predictions.

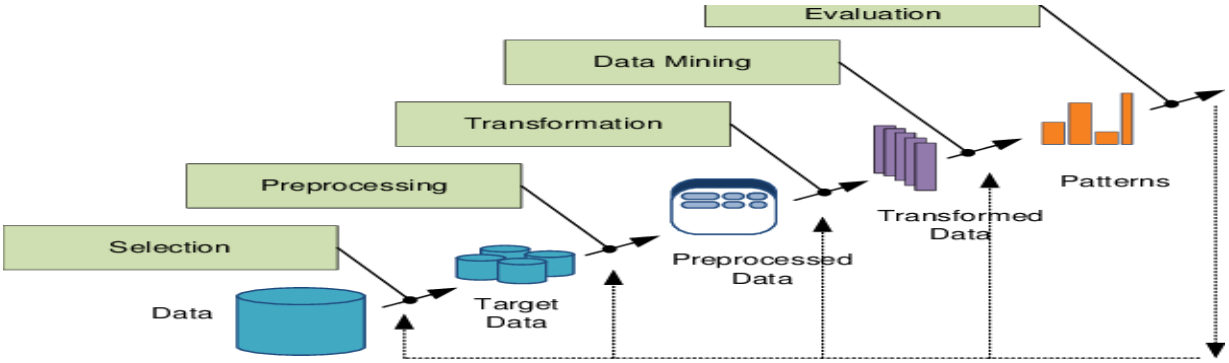

Fig.1: The Steps of KDD process

2. CLASSIFICATION IN DATA MINING

Classification is the process that is used to predict a model that describes and differentiate data classes or concepts, for the purpose of using the model to predict the class of objects whose class label is unknown. The main aim of classification techniques is to analyze the input data. The objective of classification algorithms is to place the data in the appropriate class.

3. CATEGORIES OF CLASSIFICATION ALGORITHMS

In this paper, five major types of classification algorithms are focused. This picture shows the general categories of classification algorithms.

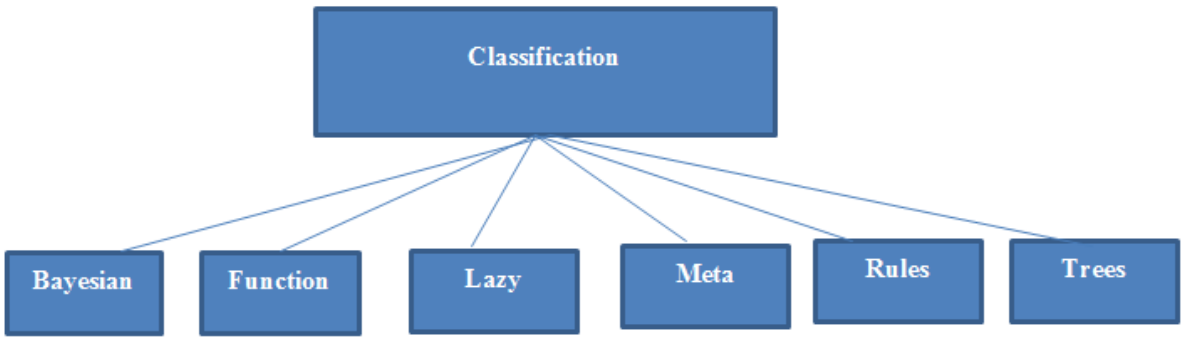

Fig: 2 Categories of the Classification Algorithms

3.1 Naïve Bayes Classification Algorithm

One of the efficient classification algorithm is Naïve Bayes classification algorithm which is based on Bayes Theorem. This Naïve Bayes is not a single algorithm but a family of algorithms where all of them share common principle. In simple, naïve bayes classifier assumes that the presence of a particular feature in a class unrelated to the presence of any feature.

3.2 Decision Tree Classification Algorithm

The decision tree is a widely used classification algorithm which looks like a flow-chart like tree structure. In this tree structure, an internal nodes specifies the feature or attribute, the branch represents a decision rule, and all the leaf node shows the outcome. The top node in the decision tree is called a root node.

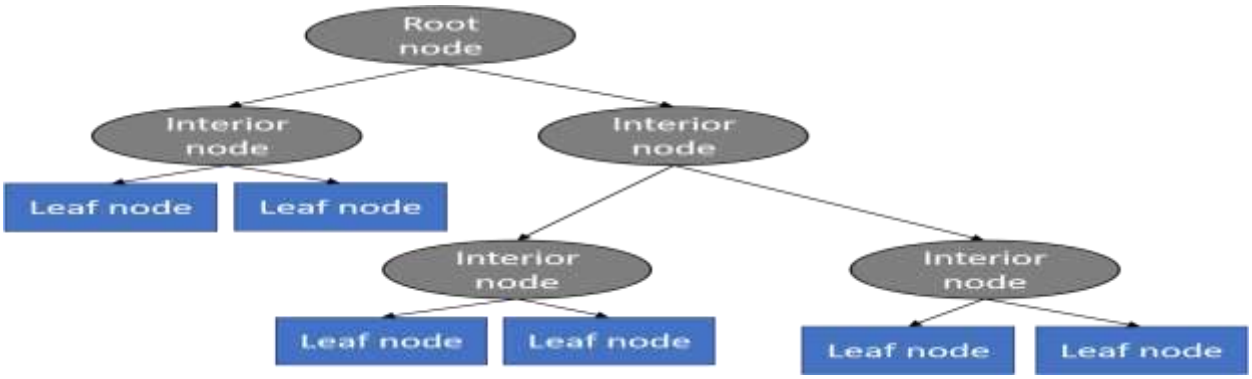

Fig.3: Decision tree Model

3.3 K-Nearest Neighbor Classification Algorithm

K-Nearest Neighbor is the popular classification algorithm because it stores all the available cases and classifies new one based on the similarity measure. It works based on the minimum distance from the query instance to the training samples to determine the k-nearest neighbors.

3.4 Support Vector Machine Algorithm

Support Vector Machine is one of the powerful classification algorithmwhich follows a supervised learning method. In this algorithm, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular co-ordinate. Then we perform classification by finding hyper-plane that differentiate two classes very well.

3.5 Random Forest Classification Algorithm

Random Forest Algorithm consists many decision trees. It uses bagging and feature randomness when building each individual tree to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

4. LITERATURE SURVEY

HuseyinPolat et al[1] used feature selection methods to predict Chronic Kidney Disorder based on Support Vector Machine. They used Wrapper method and Filter method to produce more accurate results in the prediction. They showed the results that the support vector machine classifier by using filter subset evaluator with Best First Search Engine method has high accuracy rate (98.5%) in the diagnosis compared with other selection methods.

Uma N Dulhare et al[2] proposed a prediction system for Heart Disease using Naïve Bayes and Particle Swarm optimization. Their proposed model is efficient to improve the accuracy of Naïve Bayes classifier using Particle Swarm Optimization for feature subset selection which achieves similar or even better classification performance. From the simulation results, it analyzed that this algorithm produced a better results in predicting heart disease. The comparison result showed that Naïve Bayes with PSO produce 87.91% of accuracy.

PratibhaDevishri S et al[3] used Feature Selection method for Chronic Kidney Disease Prediction. Principal Component Analysis is one of the feature selection technique that filters less important attributes; it pick only the important attributes. They examined Various classification algorithms like Decision Stump, Rep tree, IBK, K-star, SGD and SMO classifiers using performance measures. Accuracy measures used to compare classifiers are recall, F-measure and precision by implementing WEKA. They showed the results that the accuracy measure for Decision Stump and Rep tree where mean the absolute error was less with error rate 0.010 and 0.012 respectively.

VincyCherian et al[4] used Naïve Bayes algorithm and a smoothing technique Laplace Smoothing. The proposed decision support system was believed to avoid unnecessary diagnosis test conducted in a patient and delay in starting appropriate treatment by quickly diagnosing heart disease in a patient. This system predicts whether the patient is having heart disease or not. Their proposed work produced 86% of prediction accuracy by using smoothing technique with Naïve Bayes algorithm.

Mustafa S. Kadhm et al [5] proposed a system that used K-Nearest Neighbor algorithm for eliminating the undesired data, thus reducing the processing time. The proposed classification approach based on Decision Tree to assign each data sample to its appropriate class. The proposed system used 768 instances within 8 attributes for each one of PID(Pima Indians Diabetes). The used data is preprocessed in order to reduce the unwanted data, and lead to fast processing time. The proposed system achieved high classification result which is 98.7% comparing to the existing system using Pima Indians Diabetes dataset.

AiswaryaIyer et al [6] proposed a system for predicting diabetes using Decision tree and Naïve Bayes algorithm. Naïve Bayes and J48 decision tree

model were used for prediction system and the results were compared. According to their experimental results, both methods had a comparatively small difference in error rate, though the percentage split of 70:30 for Naïve Bayes Technique gives less error rate as compared to J48 model.

5. CONCLUSION

In this paper a study among classification algorithms have been carried out(Naïve Bayes, Decision Tree, Support Vector Machine, K-Nearest Neighbor etc). Various disease are focused in this study which used classification algorithms and produced good results in predicting the disease. The results of those research showed that, classification algorithms are more effective in healthcare industry. By using classification algorithms, the accuracy of identification and prediction of disease is high and the error rate is low. The time taken for this disease prediction is also very less when comparing with other tests conducted by hospitals. In future, I focus the classification algorithms with some particular disease prediction and compare the results with other existing results.

REFERENCES

1. *HuseyinPolat, HomayDanaeiMehr, Aydin Cetin, "Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods" Journal of Medical System, Spriger, April 2017.*
2. *Uma N Dulhare, "Prediction system for heart disease using Naive Bayes and particle swarm optimization" Biomedical Research,2018; 29(12): 2646-2649.*
3. *PratibhaDevishri S, Ragin O R, Anisha G S, "Comparative Study of Classification Algorithms in Chronic Kidney Disease" International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-1, May 2019.*
4. *VincyCherian, Bindu M.S, "Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique" International Journal of Computer Science Trends and Technology (IJCST) – Volume 5 Issue 2, Mar – Apr 2017.*
5. *Mustafa S. Kadhm, IkhlasWatanGhindawi, DuaaEnteeshaMhawi, "An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach" International Journal of Applied Engineering Research, Volume 13, Number 6 (2018).*
6. *Aiswarya I., S. Jeyalatha and Ronak S., "Diagnosis Of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), vol.5, ,No. 1, pp. 1-14, 2015.*
7. *V. Anuja and R.Chitra., "Classification Of Diabetes Disease Using Support Vector Machine", International Journal of Engineering Research and Applications (IJERA), vol.3,Issue 2, pp. 1797-1801, 2013.*
8. *K.Thenmozhi, P. Deepika, M.Meiyappasamy ," Different Data mining Techniques involved in Heart Disease Prediction: A Survey" International Journal of Scientific Research, Volume-3, Issue-9, September 2014.*
9. *Bramer, M., Principles of data mining. 2007: Springer.*
10. *Arun K Pujari , 'Data Mining Techniques', University press , Edition 2001.*
11. *G. Krishnaveni*, T. Sudha," A Novel Technique To Predict Diabetic Disease Using Data Mining Classification Techniques" in International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT2017), vol. 3, Issue 1, pp. 5-11, 2017.*
12. *A.Kiruthika, P.Deepika, S.Sasikala, S.Saranya," Predicting Ailment of Thyroid Using Classification and Recital Indicators "International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume-3, Issue-3, 2018.*