APPLICATION OF DATA MINING CLASSIFICATION TECHNIQUES IN PREDICTION OF HEART DISEASE: A REVIEW

[1]Bhabesh Deka, [2]Sangeeta Kakoty
[1]Research Scholar, [2]Dy. Director (Multimedia, K.K.H. State Open University)
[1]Dept. of Computer Science & Engineering
[1]Assam down town University, Guwahati, India

Abstract: The invention of an efficient and accurate diagnosis technique for treatment of any disease is always become a blessing for the society. One of the major goal of research and scientific activity of today's world is to improve the health sector. It has seen that modern Bio-medical science today uses data mining algorithms to study the history of major diseases, which is obviously a critical job. This paper primarily focuses on the research related to diagnosis of heart diseases that uses data mining classification techniques. Different important data mining classification algorithms specially Decision tree, SVM, k-NN are discussed through this study.

Keywords: *Data mining, Heart disease, Decision tree, SVM, k-NN.*

I. INTRODUCTION

The development of modern techniques in the field of Computer Science and Information Technology leads to new generation where we can use the data mining techniques to predict different diseases in the health care sector. In the field of computer science and engineering, Data mining is also called knowledge discovery in databases (KDD). It is the process of discovering interesting and useful patterns and hidden relationships in data warehouses. It includes various tools from Statistics as well as Artificial Intelligence (such as neural networks and machine learning) with database management for analysis of huge digital data sets. Due to increase of population as well as pollution in every direction, it primarily affects the human health, which leads to different diseases in our body at different age levels. Today, most of us have become too much busy in our work that results lack of physical exercise and also consumption of unhygienic foods makes us lazy and abnormal weight gain. This is why a common and popular disease that affects most of us is the Heart disease, which is also known as Cardiovascular Disease (CVD). According to World Health Organization (WHO), around 17.9 million people die every year due to CVD, which is approximately 31% of all global deaths [12]. This makes it number one cause of death globally. In 2015, a study shows that in India, people of ages between 30 to 69 years, out of 1.3 million cardiovascular deaths, 0.9 million (68.4%) were caused by coronary heart disease and 0.4 million (28.0%) by stroke. It also shows that Adults born after the 1970s are much more vulnerable to such deaths than those born earlier [13].

Different survey shows that major risk factor related to heart disease includes: Previous heart disease history in family, Excessive smoking habit, Increase level of cholesterol, High blood pressure, abnormal weight gain, physical inactivity etc.

II. RELATED WORK

In recent years due to development of technology both in the field of medical science and computer application, it is possible to analyse and retrieve useful facts from the medical databases. A review of different data mining techniques especially in heart disease prediction is reflected in this study. Here, we have also mentioned some other papers related to prediction of different other diseases that have used the similar tools and techniques. Most of the papers have used the WEKA tool in their study and compared the dataset with more than one data mining techniques.

Shilpa M. Satreet. et. al [1],have presented a medical diagnosis system for predicting the risk of CVD. They tried to build the system using the Back propagation algorithm of Artificial Neural Network (ANN) and they found that the classification accuracy is around 90.17% by that approach.

Ajad Patel et. al [2],described a prototype with the help of two data mining techniques Naïve Bayes and Weighted Associated Classifier (WAC). They used a data set which consist different attributes like Age, Sex, Height, Weight, Blood pressure, Sugar, Cholesterol etc. The prime objective of their proposed system was to find out the risk of heart disease in a patient.

Prof. Mamta Sharma et. al [4], in their paper attempts to compares different data mining techniques specially Decision tree and Neural network algorithms to predict heart disease. They proposed a feed-forward back propagation neural network model that consists of three layers. The outcome of their study shows that Neural Network algorithm which consists of 15 attributes gives the highest accuracy over other algorithms like Decision Tree. Also, they found that Naïve Bayes algorithm provides accuracy around 90% which was at average level.

Prof. Priya R. Patil and Prof. S.A. Kinariwala [5], presented a research work where they have proposed a system for automatic diagnosis of heart disease by using the improved random forests classification algorithm. Their primary aim of the proposed system was to construct an ensemble that provides optimal accuracy and correlation. They also mentioned that their main objective of building an ensemble classifier was to reduce variance and reduce bias.

Megha Shahi and Prof. R.K Gurm [6], in their research explore different data mining algorithms that have been used in health care to predict heart disease. They found that SVM algorithm provides a good level of accuracy which is around 85% in compared to other techniques.

N. Deepika et. al [7], through this paper the authors presented the way of extraction of hidden patterns related to heart disease from data warehouse, which contains various important data sets of heart patients. In the first stage, they have prepared the training data set by removing unwanted data values, so that it gives more accuracy during mining process.After this process they applied association rule on that data.

B. Venkatalakshmi and M.V Shivsankar [8], in there research design and develop a predictive diagnosis system for heart disease base on predictive data mining. They have conducted different experiments to compare the performance of various predictive data mining techniques like Decision tree and Naïve Bayes algorithms. They used a clinical data set with 13 attributes and applied WEKA data mining tool.

In most of the research it is observed that they have taken different attributes in their data set and tested the accuracy level with more than one data mining algorithms. Their performance varies according to the test dataset that they have applied. It is difficult to predict the efficiency level of a particular algorithm. It means that the performance of an algorithm directly depends on the selection and filtration of the collected dataset with appropriate number primary attributes.

III. ALGORITHMS AND TOOLS

3.1 ALGORITHMS

The main goals of classification algorithms are to maximize the predictive accuracy obtained by the classification model [3]. There are different types of classification algorithms in data mining but this paper elaborates three of them, which are mentioned below:

- Decision Tree: It provides several distinct advantages in many classification and prediction applications. Decision Tree Classifier, repetitively divides the working area (plot) into sub part by identifying lines. It is a predictive machine-learning model that selects some target values which is based on different attribute values of the training dataset. Here, the internal nodes of the tree denote the different attributes and the edges between the nodes indicate the possible values. The dependent variable is an attribute which is to be predicted and its value is calculated based on the other attribute values. On the other hand the independent variables are those attributes which helps in predicting the value of the dependent variable. Decision tree is similar to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label [4]. B.Venkatalakshmi, M.V Shivsankar [8], in their experiment used WEKA tool and their data set consists of 294 records with 13 attributes and they found that Decision Tree algorithm provides an accuracy with 84.1% where as Naïve Bayes provides 85.03% accuracy.

- SVM: Support Vector Machine (SVM) is developed during 1990 and it is the one of the most popular machine learning algorithms in data mining. It can perform tasks like binary classification as well as estimation of regression. The prime reasons behind its popularity are that it is more efficient to minimise the expected error then minimizing the classification error like other techniques do. Also, it is more efficient in computational methods as because it supports duality theory of Mathematical programming. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge [3]. SVM is also called as a hyper-plane and it can separate the classes with more optimality and creates the largest possible distance between the separating hyper-plane.

- k-Nearest Neighbour: k-NN is one of the simple algorithms which accepts all available datasets and classifies new datasets depending on some similarity measure. The K-nearest-neighbour classification was developed from the need to perform different analysis when parametric estimates of the probability densities are unknown or difficult to find [6]. It has been used in many important activities like pattern recognition, statistical analysis etc. from a long time ago. It is the non-parametric Machine learning algorithm that works on the basis of its closest neighbour whose class is already known. In this algorithm the nearest neighbour is calculated by estimating the value of k, which is the number of nearest neighbours that are to be considered to characterise. To improve the accuracy and performance the k-NN algorithm has been modified several ways by different authors.

These models are applied using different data mining tools like WEKA, Orange, Rapid Miner etc. Generally, after preparation of the test dataset, the testing phase starts where these models are applied. And in the final stage, the result can be analysed to get the prediction accuracy and efficiency of each algorithms separately.

3.2 WEKA TOOL

Waikato Environment for Knowledge Analysis (WEKA) is a collection of machine learning algorithms which are used for data mining purpose. WEKA is a state of the art facility for developing machine learning techniques for real world data mining problems [6]. It is developed at the University of Waikato, New Zealand and is free software licensed under the GNU General Public License. WEKA includes different visualization tools and algorithms with well-designed graphical user interface (GUI)

which can be used in data analysis and predictive modelling. The algorithms can either be applied directly to a dataset or we can call it in our own Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

In this study, we have not only considered the prediction accuracy level of heart disease diagnosis but also some other cases like Diabetic diagnosis prediction, Students' performances etc. are considered. Since our goal is to find out the most efficient technique that will show result with higher accuracy, we have studied the results of other research activities that have used data mining techniques.

IV. CONCLUSION

There are various tools and techniques available which are used in early prediction of Heart disease and also different research activities have already been done in this sector till now. It is observed that if one algorithm performs well in one situation, it may not perform well in some other situation. That is why selection of a particular algorithm is not so easy. But in most of the cases SVM, k-NN and Decision tree has performed well in comparison with other algorithms.

V. ACKNOWLEDGEMENT

Authors acknowledge the immense help received from the scholars whose articles are cited and included in references of this manuscript. The authors are also grateful to authors/ editors / publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed. The first author also likes to thank Assam down town University for providing us an better opportunity in this research.

VI. REFERENCES

[1] Shilpa M. Satre, Sunaina Manohar Bhagat, Shalini Manoj Thakur. February 2018. Heart Disease Prediction System Using Data Mining. International Journal of Engineering Science and Computing. Volume 8 Issue No.2

[2] Ajad Patel, Sonali Gandhi, Swetha Shetty, Prof. Bhanu Tekwani. Heart Disease Prediction Using Data Mining. Jan -2017. International Research Journal of Engineering and Technology (IRJET). Volume: 04 Issue: 01, e-ISSN: 2395 -0056

[3] Mr.Sudhir M. Gorade, Prof. Ankit Deo, Prof. Preetesh Purohit. April -2017. A Study of Some Data Mining Classification Techniques. International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 04, e-ISSN: 2395 -0056

[4] Prof. Mamta Sharma, Farheen Khan, Vishnupriya Ravichandran. June -2017. Comparing Data Mining Techniques Used For Heart Disease Prediction. International Research Journal of Engineering and Technology (IRJET). Volume: 04 Issue: 06, e-ISSN: 2395-0056,

[5] Patil R Priya, Kinariwala A S. Automated Diagnosis of Heart Disease using Data Mining Techniques. International Journal of Advance Research, Ideas and Innovations in Technology. Volume-3, Issue-2, ISSN: 2454-132X

[6] Megha Shahi, Er. Rupinder Kaur Gurm. April-2017 . Heart Disease Prediction System Using Data Mining Techniques - A Review. International Journal of Technology and Computing (IJTC), ISSN-2455-099X, Volume 3, Issue 4.

[7] N. Deepika and K. Chandrashekar. 2011. Association rule for classification of Heart Attack Patients. International Journal of Advanced Engineering Science and Technologies. Vol.11, No.2, p253-257.

[8] B.Venkatalakshmi, M.V Shivsankar. March 2014. Heart Disease Diagnosis Using Predictive Data mining. International Journal of Innovative Research in Science, Engineering and Technology. Volume 3.

[9] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra. 2012. Performance Analysis of Classification Data Mining Techniques over heart Diseases Data base. International Journal of Engineering Science and Advanced Technology.

[10] Deepali Chandna. 2015. Diagnosis of Heart Disease Using Data Mining Algorithm. International Journal of Computer Science and Information Technologies (IEEE Conference). p1678-1680.

[11] V. Krishnaiah, Dr. G. Narsimha, Dr. N. Subhash Chandra. 2013. Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. International Journal of Computer Science and Information Technologies. Vol. 4(1), 39 – 45, ISSN: 0975-9646

[12] 26-10-2018. URL: https://www.who.int/cardiovascular_diseases/world-heart-day/en/

[13] 26-10-2018. URL: https://www.livemint.com/Politics/fKmvnJ320JOkR7hX0lbdKN/Rural-India-surpasses-urban-in-heart-disease related-deaths.html