North Asian International Research Journal of Sciences, Engineering & I.T.

ISSN: 2454-7514 Vol. 7, Issue-3 March -2021

IRJIF. I.F. : 3.821

Index Copernicus Value: 52.88 Citation Index

A Peer Reviewed Refereed Journal

DATA MINING TECHNIQUES FOR PREDICTION OF HEART DISEASE

SHEIKH AMJAD*

****Daffodil International University, Dhaka, Bangladesh***

Abstract: Heart disease is one of the deadliest diseases not only in a particular region but also all over the world. Millions of people die in Europe, North America and Indian subcontinent cause of heart disease. Heart diseases are various conditions which may affect one's heart. If the heart diseases can diagnosed earlier the possibility of death can be reduced. Along with medical researcher many computer sciences researchers shows interest in this field to find better way to diagnosis heart disease. Many researchers proposed different Data Mining technique for this. Data Mining is a great way to discover knowledge form the invisible pattern of huge collection of data. This paper reviewed several data mining technique to detect heart diseases using some machining learning algorithms which used different attribute of patient data. Many attribute like patient age, gender, CP, Fbs, Trestbps, Chol, Smoking history etc. are considered.

Keywords— Data mining, Machine Learning, Heart Disease, Naïve Bayes, Decision tree, Classifier, Prediction.

1 Introduction

The process of finding pattern in massive collection of data is called Data Mining. It is a subset of statics and computer science. It is a very useful technique to fetch knowledge from a big dataset; data mining uses statistical and machine-learning models to discover invisible patterns in a large collection of data. Data Mining gain a vast popularity in field like Education, Business Analysis, Fraud Detection, Manufacturing Engineering, The success of data mining in particular field spread the use of data mining in different sector. In field like Health Care and Bio-informatics the application of data mining is slightly visible. Data mining seems a promising way to improve the future healthcare system. Data mining can play a very crucial role in disease prediction. Many researcher uses data mining technique to predict the heart disease.

Heart disease is a term included any disorder of the heart. The report of Centers for Disease Control (CDC) says, heart disease is the major cause of death in country like United States, Canada, United Kingdom and Australia. One in each four deaths in the U.S. happens as the cause of heart disease [1]. There are many reasons which may lead to heart disease. As example, research's writing in BMJ found stress disorder may cause cardiovascular diseases, according to their research a person with stress has additional 37 percent risk of cardiovascular risk then a person

having without stress disorder [2]. The diagnosis often made on the basis of the doctor's prior experiences and knowledge. Which sometime may not achieve accurate results and this process is very time consuming and costly as well. Data mining's ambition is to detect heart disease using several methods with the help of medical data. Machine Learning algorithms like Decision Tree, Naïve Bays, Support Vector Machine, Random Forest, are repeatedly used by researcher to predict heart disease. Along the classification model, selection of attribute played a very important role, in this paper also discusses how the attribute selection impact on the accuracy of detection.

2 Literature Review

They several papers published on prediction and early diagnosis of heart disease. In this section few previously performed work to predict heart disease using different technique will be discussed. Those papers use different data mining techniques with various machine learning methods.

K. Aravinthan and Dr. M.Vanitha [3] on their paper described how Naïve Bayes can use to predict heart disease. They use Naïve Bayes classifier, this classifier is based on Bayes' theorem, the formula given bellow.

$$P_{A/B} = P_{B/A} \times P_{APB}$$

Where P(A) and P(B) is the probability of A and B's independent observation of each other. Again P(A/B), P(B/A) are both conditional probability, here P(A/B) is the probability of A at the presence of B and P(B/A) is the probability of B at the presence of A [5].

They collected medical data which contained records of both person having heart disease or no heart disease. Different attribute are chosen from the data set for classification. Where attributes like persons gender is Discrete attribute only have two limited values male and female. Also having Continues attributes like Thalach which may have numerous value with respect of each person's maximum heart rate, continues attribute's values are not limited in particular range. After classification the classifier produce value between 0 and 1 for each person. If the value is between 0 – .49 consider having no heart disease on the contrary value .50 – 1.0 is denoted person having heart disease. In this paper they include the Accuracy Rate of the classifier with parameter like Sensitivity, Specificity and Precision. Here sensitivity means rate of True Positive and Specificity means rate of True Negative. Where Precession is rate of True Positive with respect of the summation of both True Positive and False Positive.

Sonam Nikhar1, A.M. Karandikar [4] uses two different techniques, one is Naïve Bayes and the other is Decision Tree. They also go through another vital process, which is called feature selection. Feature selection is the way of selection the best attribute on the basis of the information gain among numbers of attribute. To find the gain they uses those formula noted bellow.

$$IS1,S2,\ldots\ldots..Sm = -\sum_{k=1}^{m} pk \log_2 pk$$

Hence, IS1,S2,……..Sm

is required information for classify a particular sample

EA= ∑j=1vS1j+……..+SmjSIS1j,……..,Smj

Here E (A) is entropy, which measures the purity of attribute from particular instance of dataset.

Information Gain (A)= IS1,S2,……..Sm– E (A)

Thus information gain is calculated. After achieving gain for each and every attribute, some attribute with lower gain are eliminated and remaining were selected for classification.

As previous section discussed about Naïve Bayes, here start from Decision Tree. Decision tree one of popular tool for machine learning, researcher often use it operational research. The basic of this algorithm is greedy approach that makes tree in top down recursive manner. The structure of decision tree seems like flow chart, where each node represents a test of a particular attribute, the branches illustrate outcome of a test. The class labels are denoted by each leaf. The way to the leaf from root illustrate the classification rules [6]. The algorithm start with all instance of the dataset with the selected attribute. After successful implementation the accuracy result of heart disease is obtained. This paper found best performance using Decision Tree among those two algorithms.

Kanika Pahwa, Ravinder Kumar on their paper [7] evaluated two different data mining technique. They applied Random Forest and Naïve Bayes technique to predict heart failure. They initially take 14 feature like patient “Age”, “Sex”, “Thalach”, “Oldpeak”, “Thal” etc. Then with feature selection to reduce the feature for the classifier. Here for feature selection the follow slightly different approach then previous paper [4]. While that one use only total gain of the attribute, this paper calculate weight for each feature. For weight of each attribute Gain Ratio and feature rank is used. For feature rank SVM-RFE is applied. SVM-REF is a recursive feature eliminating algorithms. The equation for determining Weight is bellow,

Weight A= score A rank A+1*1000

Where score A

Is the Gain Ratio and rank A

Is the Feature Rank. After calculation of the feature with bellow threshold are eliminated and Random Forest and Naïve Bayes technique are applied on remaining features.

3 Data Sources

Collection data is a vital part for any research. In [3] and [4] data collected from Cleveland Heart Disease database. In this database several records and medical attributes are acquired. After processing the data, a portion of data used for training and remaining used for test. Description of few attribute of dataset show in the table below.

Table 1: Attributes of Heart Diseases Data Set

Attribute	Type	Description
cp	Discrete	Type of chest pain (4 types).
thal	Discrete	3 for “normal”, 6 for “fixed defect” and 7 for “reversible defect”s
sex	Discrete	Male = 1, Female = 0
exang	Discrete	Exercise induced angina: 1 = True 0 = False
ca	Discrete	“major vessels colored” number by fluoroscopy that range 0-3
age	Continues	Age of individuals.
slope	Discrete	peak exercise segment : 1 means up sloping, 2 means flat and 3 denotes down sloping
oldpeak	Continues	Depression induced by exercise relative to rest
thalach	Continues	Max heart rate
trestbps	Continues	Resting Blood Pressure
chol	Continues	Serum cholesterol (mg/dl)
restecg	Discrete	Resting electrocardiographic results:
fbs	Discrete	Fasting Blood Sugar > 120 mg/dl: 1= true 0 = false

4 Results

In this section of this paper overall result of different methodology will be discussed. Firstly in [3] using Naïve Bayes classifier the average accuracy rate for predict heart disease is 83.6296 and error rate is 16.3703. They uses total of 13 attribute. In [4] found better accuracy rate Decision Tree with comparison of Naïve Bayes. They also found that pre-calculation of Gain Ratio of the attributes of the dataset bring more accuracy then selection of random attribute. In [7] the accuracy rate of Naïve Bayes and Random Forest in random approach is respectively 78.2178 and 76.8977, where is their propose approach called “Hybrid Approach” accuracy rate for those algorithms is 83.8284 and 82.1782. They test with different number of attribute at a time. The attribute also chosen from feature selection process. They shows that number of attribute can give different rate in accuracy. More attribute can give better accuracy then less number of attribute.

5 Future Scopes and Conclusion

Over the years data mining is dealing with forecasting problem and gain valuable information from large collection of data on different field. In medical and healthcare huge data is exists, all we need to utilize them. Data mining can discover pattern from previous data and measure risk of occurring life threatening heart diseases. This paper discussed about three different data mining methods which can be a very useful for predicting heart disease based on different attribute of selected dataset. Different technique works different way and shows variation in produced result. There are still many data mining technique and algorithms, use of them can bring much better result as well as additional attribute can impact the performance significantly. As previously said that data mining have success in

several field, data mining can be a great way for future healthcare system. This may reduce the required time and cost of diagnosis of heart disease.

Acknowledgements

- Here I like to state that i am very thankful to our honorable teacher Mr. Sheikh Abujar, Lecturer, Department of Computer Science and Engineering for his proper guidance's and motivation throughout this work. Author also express gratitude to Daffodil International University for the great facilities and support provided to conduct this work.

References

1. Felman, "Medical News Today," https://www.medicalnewstoday.com, 07 02 2018. [Online]. Available: https://www.medicalnewstoday.com/articles/237191.php. [Accessed 20 08 2019].
2. N. Bakalar, "Stress Tied to Heart Disease, Especially in People Under 50," The New York Times, 17 04 2019. [Online]. Available: https://www.nytimes.com/2019/04/17/well/mind/stress-tied-to-heart-disease-especially-in-people-under-50.html. [Accessed 23 08 2019].
3. D. M. K. Aravinthan, "A Novel Method For Prediction Of Heart," International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) , vol. Vol. 3, no. Special Issue 20, April 2016.
4. A. K. Sonam Nikhar1, "Prediction of Heart Disease Using Machine," International Journal of Advanced Engineering, Management and Science (IJAEMS), Vols. Vol-2, no. Issue-6, p. 5, 2016.
5. Wikipedia, "Bayes' theorem," Wikipedia, the free encyclopedia, 08 08 2019. [Online]. Available: https://en.wikipedia.org/wiki/Bayes%27_theorem. [Accessed 2019 08 2019].
6. Wikipedia, "Decision tree," Wikipedia, the free encyclopedia, 17 07 2019. [Online]. Available: https://en.wikipedia.org/wiki/Decision_tree. [Accessed 22 08 2019].
7. R. K. Kanika Pahwa, "Prediction of Heart Disease Using Hybrid Technique," in 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), GLA University, Mathura, 2017.