Survey on Symptom Based Clinical Document Clustering

I. Mohan
Assistant Professor, Information Technology, Prathyusha Engineering College, India.

Hrithu Sathish
Student, Information Technology, Prathyusha Engineering College, India.

Kruba Gayathri. S
Student, Information Technology, Prathyusha Engineering College, India.

Abstract – This paper helps us to understand the various data mining technologies used in diagnosing heart disease in patients. Datasets of patient details are collected with correspondence to the risk factors with which the medical practitioner can predict the disease before it occurs. The term mining means bringing out patterns which are hidden and previously unidentified for better grasp of the particular problem. Several data mining techniques such as the Classification algorithms like Decision tree, Genetic algorithm, Neural network, Artificial intelligence, Naive Bayes, and Clustering algorithms like SVM, K-means and KNN are the techniques that are utilized in this process. Many models for this prediction system are developed, comparisons are made and the accuracy level of every model is provided.

1. INTRODUCTION

Medical records comprises of vast amount of data with unidentified patterns in them which can be mined. Heart disease is an important disease affecting people all over the world which leads to death since heart is an important organ. The risk factor associated with heart disease are blood pressure, age, tobacco, smoking, alcohol intake, obesity, physical inactivity, family history, poor diet, high cholesterol. These information obtained from doctor's examination helps to segregate the record and produce the result. The diagnosis is performed on of Electrocardiogram (ECG), Echocardiography (ECHO), the previous test result and doctor's experience. Data mining extraction methods helps finding knowledge and unidentified patterns from the available data. Data mining is a process of observing these patterns from the data relevant to the disease so prediction can be performed using these algorithms and techniques.

1.1 BIG DATA ANALTICS AND DATA MINING

Bigdata and Data mining helps in the analysis of large amount of data. In fields of medical industry, the tons of data available are processed using data mining techniques. Early times they used MS excel spread sheet to handle huge data as all other techniques were expensive. Database is also used for handling huge data which was very helpful to collect the exact chunk of data required which is relevant and correct. The clients using this database manager need to prepare queries for information from database through automation. Using Data Mining techniques we can easily cluster and classify data that is collected. They call Data mining as "handler" and Bigdata as an "asset", by combining these two analyses, modelling and diagnosing heart disease is performed. Using both big data and data mining we can perform predicting using various techniques and bring out expected outcome.

1.2 CARDIOVASCULAR DISEASE (CVD) OR HEART DISEASE

The disease which occurs due to blocked or narrowed blood vessels is called as cardiovascular disease. The various diseases, disorders and specific conditions in heart are called as Heart Disease. When considering whole death rate, it is found that the major cause is heart disease. Predicting this disease at an early stage is essential. The risk factors associated with heart are age, family history, high blood pressure, blood sugar level, cholesterol, poor diet, smoking, intake of alcohol, obesity, etc., which are considered for the prediction process.

1.3 HOW DATA MINING IS USEFUL FOR PREDICTING HEART DISEASE

Since these medical data comprises of huge data so these can be analysed and processed using data mining techniques. There are also several tools available for this process. Mining can also be done for this data, modelling and design can be done. Heart disease prediction model is done using data mining techniques in an efficient way and the required outcome is obtained.

2. LITERATURE SURVEY

Many papers are reviewed in which different techniques are used for heart disease prediction.

2.1 CARLOS ORDONEZ (2004) used association rule mining for prediction of heart disease. It actually was about detection

of the disease with the help of the risk factor and also measured heart perfuser along with artery depletion was found [2].

2.2 CARLOS ORDONEZ (2006) in this the imitations in association rule was solved, specified algorithm was designed to search constraint attributes which decreased the set of rules. Bioinformatics significance was based on support and confidence. In this way prediction was performed [3].

2.3 PETER LEIJDEKKERS ET.AL (2006) used heart disease monitoring application embedded in the smart phones and wireless sensor. The process of analysis is done using this application which can monitor the patient's condition to doctor and also provide alarm to the ambulance in case of emergency. The alert message is sent to the nearby health care centre therefore the patient's life can be saved. The model was designed using two methods Microsoft's Window Mobile Pocket PC Platform and .Net Compact Framework extended with OpenNETCF [4].

2.4 HONGMEI YAN ET AL. (2006) use multilayer perception which has 40 inputs and 5 output layers. Back propagation helps train the 352 medical data collected with the help of assessment models like holdout, cross validation and bootstrapping. Multilayer perception is system that has good architecture for neural network. MLP consist of three layers such as input, hidden and output using which these heart diseases were predicted with an accuracy of 90% [10].

2.5 YANWEI X ET AL. (2007) in this data mining algorithm were employed to predicting the survival of coronary heart disease this was done based on 1000 cases. This was done on basis of the observation made on patient for the past 6 months. There were three data mining techniques employed and 10 fold cross validation was performed; they are accurate sensitive and specificity. Then confusion matrix was calculated, then accuracy was calculated (i.e.) 92.1%, 91%, 89.6% was obtained from support vector machine, ANN and DT [5].

2.6 LATHA ET AL. (2007) they used Coactive Neuro Fuzzy Interface which combined two data mining algorithm (i.e.) Genetic Algorithm and Neural Network. Hybrid models were also used for prediction process with help of risk factors associated. Therefore predict the heart disease in the patient [7].

2.7 HEON GYU ET AL. (2007) they implemented multiparametric features like Linear and Nonlinear with high-rate variability of three postures namely supine, left lateral and right lateral position to find HRV indices for finding coronary heart disease. Multiple classification methods were applied like Bayesian classification, SVM and classification based on multiple classification rules and the accuracy was found to be 81%, 85%, 80% respectively. Statistical analysis was performed and their performance was noted [9].

2.8 SELLAPPAN PALANIAPPAN ET AL. (2008) in this they developed an artificial intelligence system based on which prediction was performed they introduced multiple data mining techniques like decision tree, naive Bayes, neural network etc., which gave the advantage of all these techniques put together in one. New patterns were discovered, risk patterns were taken into account there were about 15 attributes therefore the prediction system was developed. The accuracy of Neural network, Naive Bayes, Decision tree are 85.68%, 86.12%, 80.4% respectively [6].

2.9 MINAS A.KARAOLIS ET AL. (2009) used event related risk factor for this process there are two types of event modifiable and non-modifiable. Myocardial infection, percutaneous coronary intervention was also a type of event. Then accuracy was obtained 66%, 70% and 75% for each of the specified event [17].

2.10 K.SIRINIVAS ET AL. (2010) this system is based on "behaviour risk factor surveillance system" the survey was performed in coal mining areas like Singareni collieries company in Andhra Pradesh, India. In these areas it was found that the risk rate of CVD was high compared to other regions. Patient's records were collected and diagnosis was done along with the rest risk factors associated with heart disease was considered ,these records also provided the morbidity rate of that particular area. The evaluation of the system was done based on two key measures such as accuracy and sensitivity [8].

2.11 PETER, T. J. ET AL. (2012) implemented classification based data mining. The input data set consist of intrinsic linear combination this cannot be applied for modelling therefore different classification model was applied to overcome these limitations. Initially the data is cleaned using data mining techniques, Naive Bayes, KNN, Decision tree and neural network were implemented on these data collected their accuracy is noted 83.70%, 76.66% and 75.18% Naive Bayes was found to be efficient among all of these [12].

2.12 CHAITRALI S.ET AL. (2012) they used 13 attributes along with two additional attributes for prediction of heart disease. The data mining methods used are Neural Network, Decision tree, and Naive Bayes. Accuracy for each was found to be 100%, 99.62% and 90.74% respectively which was improved from the techniques confusion matrix is calculated since the accuracy of Neural Network is 100% so it was found to be the most efficient one [20].

2.13 MAI SHOUMAN ET AL. (2012) they proposed a single data mining technique instead of multiple and also used hybrid technique. The specified techniques used are kernel density, automatically defined groups, bagging algorithm and support vector machine was performed. The accuracy found was 84.1% [13].

2.14 SYED UMAR ET AL. (2013) combing two data mining techniques hybrid method was used in which GA optimization benefits are initiated to improve the Neural Network weight.

For learning and training purpose they implemented back propagation techniques. MLP network was used "12 input 10 hidden 2 output nodes". The particular back propagation used is "The Levenberg-Marquardt back propagation algorithm" where bias and weight was recorded and updated using MATLAB R2012a, Global optimization toolbox and neural network toolbox application for this purpose. The specified data mining algorithms used are SVM, Decision tree, multilayer perceptron with an accuracy rate of 82.5%, 82.5%, and 89.7% [14].

2.15 I.S JENZI ET.AL. (2013) used classifier model in data mining. Association rules along with classification techniques were implemented. The interrelationship between patterns was brought out. The GUI used for this purpose is Microsoft .Net platform, with interconnection performed by IKVM interface with java libraries. The result obtained is from receiver operating characteristics (ROC) curves and accuracy is obtained [16].

2.16 SHAMSHER BAHADUR PATEL ET AL. (2013) in this system the 14 attributes is reduced to 6 from 14 then Naive Bayes classifier with the help of clustering and decision tree is used to predict the heart disease then genetic algorithm was applied with the following attributes it was implemented in WEKA tool. The accuracy for Decision tree, Naive Bayes, Classclust is 99.2%, 96.5% and 88.3% respectively [15].

2.17 LOKANATH SARANGI ET AL. (2015) they are using cost effective model with the help of genetic algorithm therefore optimized their weight and fed into the neural network and their corresponding accuracy is obtained. The accuracy for the particular hybrid technique is 90% [18].

2.18 M SATISH ET AL. (2015) implementing Data mining techniques like Rule based Decision Tree, Naive Bayes etc. They also used specialized technique pruning classification rules were used to extract association rules from heart disease warehouses to predict the disease [19].

3. CONCLUSION

In this paper we survey the various techniques used in data mining and big data for prediction of heart disease. Each technique's accuracy is known. Big data is not much used for processing or analysis. We currently propose to use logistics classifier and K-Means for prediction of heart disease. In future using big data many enhancements can be done in the prediction of heart disease process.

REFERENCES

[1] World Health Organization. World Health statistics Annual, Geneva, Switzerland: World Health Organization (2006), available from: http://www.who.int/mediacentre/ factsheets/fs310.pdf.

[2] Carlos Ordonez. Improving heart disease prediction using constrained association rules. Presented in a Seminar at University of Tokyo; 2004.

[3] Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction, Carlos Ordonez IEEE Transactions On Information Technology In Biomedicine, VOL. 10, NO. 2, APRIL 2006.

[4] Personal Heart Monitoring and Rehabilitation System using Smart Phones Peter Leijdekkers, Valérie Gay, Proceedings of the International Conference on Mobile Business (ICMB'06) 0- 7695-2595-4/06 $20.00 © 2006 IEEE.

[5] Yanwei X, Wang J, Zhao Z, Gao Y. Combination data mining models with new medical data to predict outcome of coronary heart disease. Proceedings International Conference on Convergence Information Technology; 2007. p. 868–72.

[6] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International Journal of Computer Science and Network Security (IJCSNS) - Vol.8 No.8, August 2008.

[7] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological and Life Sciences, Vol.3, No.3, pp.157-160, 2007.

[8] Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques, K. Srinivas,G. Raghavendra Rao ; A. Govardhan, in IEEE 5th International Conference on Computer Science and Education (ICCSE), 2010.

[9] Heon Gyu Lee, Ki yong Noh and Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease diagnosis Using Linear and Nonlinear Features of HRV", Springer-Verlag Berlin Heidelberg 2007.

[10] Hongmei Yan, Yingtao Jiang, Jun Zheng, Chenglin Peng and Qinghui Li, "A Multilayer perceptron-based medical decision support system for heart disease diagnosis", ELSEVIER 2006.

[11] Shantakumar B.Patil and Y.S.Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research, Vol.31, No.4, pp.642-656, 2009

[12] An empirical study on prediction of heart disease using classification data mining techniques, Peter, T.J,. Somasundaram, K., 2012 International Conference on Advances in Engineering, Science and Management (ICAESM).

[13] Mai Shouman, Tim Turner, Rob Stocker, "Using data mining techniques in heart disease diagnosis and treatment", IEEE Japan-Egypt Conference on Electronics, Communications and Computers, 2012.

[14] Genetic Neural Network based Data mining in prediction of Heart disease using risk factors. Syed Umar Amin1, Kavita Agarwal2, Dr. Rizwan Beg Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).

[15] Shamsher Bahadur Patel, Pramod Kumar Yadav and Dr. D.P. Shukla, "Predict the Diagnosis of Heart Disease Patients using classification Mining Techniques", IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), 2013.

[16] I.S.Jenzi, P.Priyanka, Dr.P.Alli, "A Reliable Classifier Model Using Data Mining Approach for Heart Disease Prediction", International Journal of Advanced Research in Computer Science and Software Engineering, 2013.

[17] Minas A. Karaolis, Joseph A. moutiris, Dementra Hadjipanayi, "Assessment of the risk factors of coronary heart events based on data mining with decision trees", IEEE transactions on information technology in biomedicine,2010.

[18] Lokanath Sarangi, Mihir Narayan Mohanty, Srikanta Pattnaik, "An Intelligent Decision Support System for Cardiac Disease Detection", IJCTA, International Press 2015.

[19] M.Satish, D Sridhar, "Prediction of Heart Disease in Data Mining Technique", International Journal of Computer Trends & Technology (IJCTT), 2015.

[20] Chaitrali S. Dangare Sulabha S Apte, "Improve study of Heart Disease prediction system using Data Mining Classification techniques", International journal of computer application, 2012.

[21] J. Omana, S. Dhanalakshmi, V. M. Divyalakshmi, S.Mahalakshmi, "Categorization of Drugs Using SVM Classification", International

Journal of Computer Science Trends and Technology (IJCST) – Volume 5 Issue 2, Mar – Apr 2017.
[22] R. Meena, Leyan Francis, "Medical Data Analysis", International Journal of Engineering Science and Computing (IJESE) - Volume 7 Issue 4, Apr 2017
[23] H.Vidhya, R. Meena, "A study of medical big data using lambda architecture", International Journal of Engineering Research and Technology, Vol 3, Issue 4, February 2014
[24] Mohan. I, Ajith Kumar. C, Ajithkumar. B, Bhuvanesh. S, "Relevance Feature Discovery for text mining Using Feature Clustering", International Journal of Scientific Research in Computer Science Engineering and Information Technology, Volume 2 Issue 2 April 2017
[25] Mohan. I "Knowledge Discovery Using Big Data" Journal of Current Computer Science and Technology, Volume 5 issue 5 May 2015.