Disease Prediction using Data Mining Techniques

Sharad Mathur[1], Dr. Bhavesh Joshi[2]

[1]Research Scholar, Faculty of Computer Science, PAHER University, Udaipur (Rajastha) -India

[2]Research Guide, Faculty of Computer Science, PAHER University, Udaipur, (Rajastha) -India

ABSTRACT

The progress in computers science provided large amount of data. The task is to study the input data and acquire the necessary information which can be done by various data mining techniques. Data mining is one of the important areas of research that is popular in health organizations. Data mining has an active role for discovering new trends and patterns in healthcare organization which is useful for all the parties related with this field. Medical dataset has heterogeneous data in the form of numbers, text and images that can be mined to deliver variety of useful information for the physicians. The patterns gained from the medical data can be useful for the physicians to discover diseases, predict the survivability of the patients after disease, severity of diseases etc. The focus of this paper is to analyze the application of data mining in medical domain and some of the techniques used in disease prediction. We thoroughly reviewed many research paper of data mining related to disease prediction.

Keywords: Data Mining, Disease Prediction, Medical dataset.

1. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used for industrial, medical and scientific purposes. As such the process of data mining involves sorting through large amounts of data and discovering patterns in the data [1]. Medical, Agricultural and biological research studies have used various techniques of data analysis including, natural trees, statistical machine learning and other analysis methods [2]. In healthcare, data mining is becoming increasingly popular. Healthcare industry today generates large amount of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices, etc. The large amount of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining provides a set of tools and techniques that can be applied to this processed data to discover hidden patterns and also provides healthcare professionals an additional source of knowledge for making decisions.Medical reports always gives useful information for diagnosis and also facilitates therapeutic improvement. The medical knowledge management is shown as cycle among clinical research, guidelines, quality indicators, performance measures, outcomes and concepts [3]. Thus huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Medical data mining is used in the knowledge acquisition and analyses the information obtained from research reports, medical reports, flow charts, evidence tables, and transform these mounds of data into useful information for decision making[4].

2. RELATED WORK

In the paper proposed by Dhanya P Varghese and Tintu P B [5], the data mining classification techniques used on medical system and also the various papers presented on medical data mining using classification techniques are discussed. They have also explained the importance of data mining in healthcare domain.

Parvathiand and Rautaray proposed a hybrid approach by combining association rule and classification tree technique. They explained applications of data mining in medical fields with advantages and disadvantages of data mining in medical field. They discussed many algorithms of data mining used in disease predication [6].

Ranitha.S and Vydehi has examined need and usefulness of data mining technique for healthcare dataset. They have also discussed many algorithms generally used for data mining of medical data [7].

3. IMPORTANCE OF DATA MINING FOR MEDICAL DATA MINING

Generally all the healthcare organizations across the world stored the healthcare data in electronic format. Healthcare data mainly contains all the information regarding patients as well as the parties involved in healthcare industries. The storage of such type of data is increased at a very rapidly rate. Due to continuous increasing the size of electronic healthcare data a type of complexity exist in it. In other words, we can say that healthcare data becomes very complex [8].

Medical science is another field where large amount of data is generated using different clinical reports and other patient symptoms. Data mining can also be used heavily for the same purpose in medical datasets also. These explored hidden patterns in medical datasets can be used for clinical diagnosis. However, medical datasets are widely dispersed, heterogeneous, and huge in nature. These datasets need to be organized and integrated with the hospital management systems. If any disease attacks a person so instantly that it hardly gets any time to get treated with. So diagnosing patients correctly on timely basis is the most challenging task for the medical. A wrong diagnosis by the hospital leads to earn a bad name and loosing reputation. At the same time treatment of the said disease is quite high and not affordable by most of the patients particularly in India [9].

Disease prediction plays an important role in data mining. Finding of a disease requires the performance of a number of tests on the patient. However, use of data mining techniques, can reduce the number of tests. This reduced test set plays significant role in performance and time. Health care data mining is an important task because it allows doctors to see which attributes are more important for diagnosis such as age, weight, symptoms etc. This will help the doctors diagnose the disease more efficiently. Image mining techniques improved the disease prediction task and makes health care decision making easier [10]. Main aim of this paper is to study application of data mining algorithms in the area of medicine.

4. KIDNEY DISEASE PREDICTION

Kidneys are the organs that filter waste products from the blood. They are also involved in regulating blood pressure, electrolyte balance, and red blood cell production in the body. When blood flows to the kidney, sensors within specialized kidney cells regulate how much water to excrete as urine, along with what concentration of electrolytes [11].Due to Chronic Kidney Disease (CKD) millions of people die each year because they do not undergo proper treatment. The Global Burden of Disease study held in 2010 determines that chronic kidney disease ranked 27th position in list of causing total number of deaths worldwide around 1990 and can be rises to 18th position in 2010 [12].

K-Means Clustering Algorithm along with a single mean vector of centroids have been formulated to classify the clusters of varying probability of likeliness suffers from CKD. The results are obtained from a real case dataset (UCI Repository) to show the probability of disease causing factors [13].Three different neural network models have been implemented for chronic kidney disease prediction which includes back propagation neural network, generalized feed forward neural network and modular neural network. This research shows that all these models influences genetic algorithm in to their respective neural factor. All three models give better accuracy more than 85%. Compared to other models, back propagation neural network has the highest accuracy[14].Lamboder Jena, Narendra Ku. Kamila [15] analysed a chronic kidney disease dataset from UCI machine learning repository. They have used algorithms such as Naïve Bayes, Multilayer Perceptron, SVM, J48, Conjunctive rule and Decision tree for comparing the classification accuracy, presented that multilayer perceptron algorithm gives better classification accuracy and prediction performance to predict chronic kidney diseases.

Machine learning algorithms like AD Trees, J48, KStar, Naive Bayes, Random Forest algorithms are used to predict the kidney disease. The performance result of the Naive Bayes shows better accuracy rate compared to other algorithms [16].Totally, 400 records are used for training set to perform prediction. Among these, Random Forest outperforms well. Real time collected from Apollo Hospital dataset involves many imbalanced data. To address this issue a rebalancing algorithm called SMOTE has been presented to predict kidney disease of patients [17].Kidney illnesses are anticipated and looked at by the assistance of Support Vector Machine and Artificial Neural Network calculation in light of the exactness and execution time. Result demonstrates that ANN beats with diminished execution time [18].

5. LIVER DISEASE

The liver plays an important role in many bodily functions from protein production and blood clotting to cholesterol, glucose (sugar), and iron metabolism. It has a range of functions, including removing toxins from the body, and is crucial to survival. The loss of those functions can cause significant damage to the body. When liver is infected with a virus, injured by chemicals, or under attack from own immune system, the basic danger is the same – that liver will become so damaged that it can no longer work to keep a person alive. Liver disease caused by hepatotrophic viruses imposes a substantial burden on health care resources.

Persistent infections from hepatitis B virus (HBV), hepatitis C virus, and hepatitis delta virus result in chronic liver disease [19].Chronic liver disease occurs throughout the world irrespective of age, sex, region or race. Cirrhosis is an end result of a variety of liver diseases characterized by fibrosis and architectural distortion of the liver with the formation of regenerative nodules and can have varied clinical manifestations and complications. According to WHO, about 46% of global diseases and 59% of the mortality is because of chronic diseases and almost 35 million people in the world die of chronic diseases. Hence, it is estimate that more than fifty million people in the world, taking the adult population, would be affected with chronic liver disease [20].

P. Sindhujaand R. Jemina Priyadarshini, [21] in their paper, described classification techniques for analyzing liver disorder. Advantages and disadvantages of algorithms such as C 4.5, Naïve Bayes, Decision Tree, Support Vector Machine, Back propagation and Classification and Regression Tree are compared. They have presented that C 4.5 gives better performance than other algorithms. The Prasad et al. [22] Implementation of Partitional Clustering to Predict Liver Disorders by using ILPD dataset and compares the performance of supervised learning classifiers; such as Naïve Bayes, C4.5 decision tree, and k- Nearest Neighbor; to find the best classifier in liver disorders diseases. The experiment results show that NB classifier produces highest accuracy 69%.

The Tapas et al. [23] Analysis of Data Mining Techniques for Healthcare Decision Support System, by Using Liver Disorder Dataset and it compares the classification algorithms such as ANN, J48, NB, IBK, ZeroR, and VFI. The ANN classifier produces better accuracy 71.59% when compared with other classifiers. Karthik et.al [24] were applied a soft computing technique for intelligent diagnosis of liver disease. They have implemented classification and its type detection in three phases. In the first phase, they classified liver disease using Artificial Neural Network (ANN) classification algorithm. In the second phase, they generated the classification rules with rough set rule induction using Learn by Example (LEM) algorithm. In the third phase fuzzy rules were applied to identify the types of the liver disease

P.Rajeswari, G.Sophia Reena et al., has proposed the data classification is based on liver disorder. The training dataset is developed by collecting data from UCI repository consists of 345 instances with 7 different attributes. This paper deals with results in the field of data classification obtained with Naïve Bayes algorithms .FT tree algorithms, and KStar algorithms and on the whole performance made know FT Tree algorithm when tested on liver disease datasets, time taken to run the data for result is fast when compare to other algorithm with accuracy of 97.10%Based on the experimental results the classification accuracy is found to be better using FT Tree algorithm compare to other algorithms [25].

6. HEART DISEASE

Cardiovascular diseases (CVDs) are disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. Four out of five CVD deaths are due to heart attacks and strokes. Individuals at risk of CVD may demonstrate raised blood pressure, glucose, and lipids as well as overweight and obesity [26].

Chaitrali S. Dangare et.al [27] has analyzed prediction systems for Heart disease using more number of input attributes. The data mining classification techniques, namely Decision Trees, Naive Bayes, and Neural Networks are analyzed on Heart disease database. The performances of these techniques are compared, based on accuracy. Authors' analysis shows that out of these three classification models Neural Networks has predicted the heart disease with highest accuracy. According to Randa et al. Data sets dealing with the same medical problems like Coronary artery disease (CAD) may show different results when applying the same machine learning technique. The classification accuracy results and the selected important features are based mainly on the efficiency of the medical diagnosis and analysis. The aim of his work is to apply an integration of the results of the machine learning analysis applied on different data sets targeting the CAD disease. The results show that the classification accuracy of the collected dataset is 78.06% higher than the average of the classification accuracy of all separate datasets which is 75.48%.

T. Revathi and S. Jeevitha [29] analysed the data mining algorithms on prediction of heart disease. The clinical data related to heart disease is used for analysis. The results of Neural Network, Naïve Bayes, and Decision Tree algorithms are compared,Neural Network achieved good accuracy. An Intelligent Heart Disease Prediction System (IHDPS) is developed by using data mining techniques Naive Bayes, Neural Network, and Decision Trees was proposed by Sellappan Palaniappan et al.[30]. Each method has its own strength to get appropriate results. To build this system hidden patterns and relationship between them is used. It is web-based, user friendly & expandable. Kiyong Noh et al. [31] uses a classification method for the extraction of multiparametric features by assessing HRV (Heart Rate Variability) from ECG, data pre-processing and heart disease pattern. The dataset consisting of 670 peoples, distributed into two groups, namely normal people and patients with heart disease were applied to carry out the experiment for the associative classifier.

7. ASTHMA DISEASE

There is a continuous need to identify factors associated with asthma because the prevalence of asthma is rising. Various factors, such as affluence, sedentary life style , environmental tobacco smoke (ETS), childhood viral infections and air pollution, have been suggested to be important in the pathogenesis of asthma[32]. Peyman Rezaei and Hachesu study is to prediction of asthma control levels by applying data mining algorithms. Samples consist of 600 referred patient with asthma disease. Data were collected based on the study's inclusion criteria. Preprocessing was performed and various algorithms include Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN) and Naïve Bayesian was assessed. Finally results were evaluated by confusion matrix.19 Features of 24 was chosen as the most effective asthma control features. Cough has the highest Info Gain, Relief-F and Gain Ratio comparing with other features. Results shows KNN and Naïve Bayes have the highest accuracy near to 98% [33]. Pochini et al. used data from the 2013 Behavioral Risk Factor Surveil-lance System to examine the effect of various demographic, social and behavior factors on asthma prevalence in a representative sample of U.S. adults nationwide. They applied various data mining techniques, including logistic regression, neural networks, gradient boosting, and decision tree, in order to choose the best model that predicts asthma prevalence among US population [34].

Kasturi and Prasanna have focused on managing asthma in children. The approach used is C4.5 algorithm. predictive model will help in categorizing of asthma and also suggesting the best possible treatment. The choice of treatment is dependent on severity of the disease. Classification method is designed to learn from the past successes and failures and then predict the outcome [35]. Poorani and Brindha has uses data mining techniques and employed to find the correlation between indoor air quality measures and asthma symptoms and trigger. The main trigger considered in this research is the concentration of nitrogen dioxide. First the classifier tests performed on 4 subcategories which are Bayes, Function, Trees, and Rules, selected Navie Bayes Simple. RBF Network, AD Tree, and NNge as the best classifier algorithms yielding the highest classification correctness in project [36].

8. CONCLUSION

The main objective of this paper is to investigate application of data mining in healthcare for disease prediction. To achieve this various research related to field are studied. It is noticed that results may vary for different disease diagnosis based on the tools and techniques used. Data mining gives satisfactory results in disease diagnosis when proper tools and techniques used. So data mining is the promising field for healthcare predictions.

REFERENCES

[1] Witten,I. and Eibe,F. Data mining practical machine learning tools and techniques.2nded,Sanfrancisco:Morgan Kaufmann series in data management systems.,2005..

[2] Cunningham,S.J. and Holmes, G. Developing innovative applications in agriculture using data mining.In the proceedings of the south east Asia, Regional computer confederation conference., Newzealand,1999.

[3] McCourt,B.,Harrington,R.A.,Fox,K.,Hamilton,C.D.,Booher,K.,Hammond,W.E.,Walden,A. and Nahm,M.Data Standards: At the Intersection of Sites, Clinical Research Networks, and Standards Development Initia-tives. Drug Information Journal.,2007,41(3): 393-404.

[4] Wang,X.S.,Nayda,L. and Dettinger,R. Infrastructure for a Clinical Decision–Intelligence System. IBM Systems Journal.,2007,46(1), pp. 151-169.

[5] Dhanya P Varghese, Tintu P B, ―A Survey on Health Data using Data Mining Techniques‖ , International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 07, e-ISSN: 2395-0056, p-ISSN: 2395-0072, Oct-2015.

[6] Parvathi I, SiddharthRautaray, ―Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain‖ , International Journal of Computer Science and Information Technologies, Vol. 5 (1), 838-846, ISSN: 0975-9646, 2014.

[7] Ranitha.S, Vydehi.S (2017), Data Mining In HealthCare Datasets, International Journal of Engineering Development and Research, Volume 5, Issue 4 | ISSN: 2321-9939.

[8] Parvez Ahmad, SaqibQamar, Syed QasimAfserRizvi (2015), Techniques of Data Mining In Healthcare: A Review, International Journal of Computer Applications, Volume 120 – No.15, June 2015, (0975 – 8887).

[9] B. Umadevi1, M. Snehapriya (2017), International Journal of Science and Research,A Survey on Prediction of Heart Disease Using Data Mining Techniques,Vol 6, Issue 4, 2319-7064

[10] SarangamKodati α & Dr. R. Vivekanandam, (2018), Analysis of Heart Disease using in Data Mining Tools Orange and Weka, Global Journal of Computer Science and Technology: Software & Data Engineering , Volume 18 Issue 1 Version 1.0 Year 2018, 0975-4172.

[11] https://www.medicinenet.com/kidney_failure/article.htm

[12] S.DilliArasu, R.Thirumalaiselvi,Review of Chronic Kidney Disease based on Data Mining Techniques, International Journal of Applied Engineering Research, Volume 12, Number 23 (2017) pp. 13498-13505, ISSN 0973-4562.

[13] AbhinandanDubey. (2015). A Classification of CKD Cases Using Multivariate K-Means Clustering. International Journal of Scientific and Research Publications, 5(8), 1-5

[14] Ruey Kei Chiu and Renee Yu-Jing, Constructing Models for Chronic Kidney Disease Detection and Risk Estimation. Proceedings of 22nd IEEE International Symposium on Intelligent Control, Singapore, pp. 166-171, 2011.

[15] Lambodar Jena and NarendraKu.Kamila, Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease.International Journal of Emerging Research in Management &Technology, 4(11), pp.110-118, 2015.

[16] Swathi Baby, P. &Panduranga, T. (2015).Vital, Statistical Analysis and Predicting Kidney Disease Using Machine Learning Algorithms. International Journal of Engineering Research and Technology, 4(07), 206-210

[17] Sai Prasad Potharaju&Sreedevi M. (2016). An Improved Prediction of Kidney Disease using SMOTE. Indian Journal of Science and Technology, 9(31), 1-7.

[18] Vijayarani, S. &Dhayanand, S. (2015). Kidney disease Prediction Using SVM and ANN Algorithms. International Journal of Computing and Business Research, 6(2), 1-12.

[19] NazmunNahar, FerdousAra, (2018), liver disease prediction by using different decision tree techniques, International Journal of Data Mining & Knowledge Management Process, Vol.8, No.2, March 2018, 2231007x

[20]http://www.worldgastroenterology.org/publications/e-wgn/e-wgn-expert-point-of-view-articles-collection/global-burden-of-liver-disease-a-true-burden-on-health-sciences-and-economies

[21] D.Sindhuja, R. JeminaPriyadarsini, “A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder”, International Journal of Computer Science and Mobile Computing, Vol.5, Issue.5, ISSN 2320–088X, May 2016.

[22] B. MS Prasad, M. Ramjee, SomeshKatta, and K. Swapna. "Implementation of partitional clustering on ILPD dataset to predict liver disorders." In Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on, pp. 1094-1097. IEEE, 2016.

[23] B. Tapas Ranjan, and Subhendu Kumar Pani."Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset."Procedia Computer Science 85 (2016): 862-870.

[24] Karthik. S, Priyadarishini. A, Anuradha.J and Tripathi. B. K, Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types, Advances in Applied Science Research, 2011, 2 (3): page no 334-345

[25] P. Rajeswari ,G. Sophia Reena , Analysis of Liver Disorder Using Data Mining Algorithm,Global

Journal of Computer Science and Technology,2010.

[26] https://www.who.int/cardiovascular_diseases/en/

[27]Chaitrali S. Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 – 888), Volume 47– No.10, June 2012, page no 44-48

[28] E. Randa, Mostafa A. Salamay, Omar H. Karam, and M. EssamKhalifa. "Feature analysis of coronary artery heart disease data sets." Procedia Computer Science 65 (2015): 459-468.

[29] T. Revathi, S. Jeevitha, ―Comparative Study on Heart Disease Prediction System Using Data Mining Techniques, Volume 4 Issue 7, ISSN (Online): 2319-7064, July 2015.

[30]SellappanPalaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008

[31] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee, and KeunHoRyu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer, Vol:345, pp: 721-727, 2006.

[32] SheetuSingh,BharatBhushan Sharma, S. K. Sharma, Mohammed Sabir, Virendra Singh,Prevalence and severity of asthma among Indian school children aged between 6 and 14 years: Associations with parental smoking and traffic pollution, Journal of Asthma, September 2015, 1532-4303, DOI: 10.3109/02770903.2015.1087558

[33] PeymanRezaei-Hachesu, TahaSamad-Soltani, RuhollahKhara, Mehdi Gheibi, NazilaMoftian, PREDICTION OF ASTHMA CONTROL LEVELS USING DATA MINING METHODS: AN EVIDENCE-BASED APPROACH, 5th International Society for Evidence-Based Healthcare Congress, Kish Island, IranOral., February 2017.

[34] Alma Pochini, Ben M. Williams, Hasanboy M. Isomitdinov, Gongzhu Hu, A Data Mining Analysis of Asthma Risk Factors, 3rd International Conference on Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence, July 2015 , 978-1-4673-9642-4/15, DOI 10.1109/ACIT-CSI.2015.101.

[35] K.Kasturi, Dr.S.Prasanna, Data Mining Approach of Classification for Managing Asthma in Children, International Journal of Emerging Technologies in Engineering Research (IJETER), Volume 4, Issue 7, July (2016), 2454-6410.

[36] S. Poorani, P. Gokila Brindha, An overview of data mining techniques in asthma prediction, International Journal of Multidisciplinary Research and Development, Volume 3; Issue 3; March 2016; Page No. 111-114; (Special Issue), Online ISSN: 2349-4182 Print ISSN: 2349-5979