International Journal of Engineering & Technology, 7 (3.8) (2018) 161-164

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Analyzing Health Data by Automatic Prediction of Patient Health Condition

Miss. Prajakta Sanap[1]*, Prof. Santosh Kumar [2]

[1]Department of Computer Engineering SITRC, Nashik-422213, India
psanap694@gmail.com
[2]Department of Computer Engineering SITRC, Nashik-422213, India
** E-mail: santosh.kumar@sitrc.org*

Abstract

The amount of data generated in various fields is increasing at high speed. Due to increase in the data generation, the problem regarding data storage arises. The increasing digitization in various sector has given rise to digitized data. The data in digital format is very convenient. The healthcare sector also has emerged with such convenient digital data. Digitization of healthcare information is useful for all healthcare stakeholders. The healthcare data once saved in digital format, it will be preserved for long time and also will be available anytime when needed. The health sector includes various different type of data and such varied type of digital data needs to be stored properly. The health digital data can be easily managed and processed by using Big Data concept. Most of the deaths are due to delayed disease detection and its treatment. If the disease is predicted earlier then proper treatment can be given on time to cure that disease. If the automatic disease prediction system is developed then disease can automatically be detected based on the symptoms observed in patient. The proposed system mainly focuses on Automatic prediction of disease. In this paper we use Classification algorithms to classify various disease symptom. We have executed our technique and assessed its execution utilizing different dataset. The results of the experiment shows that this research gives more accuracy of predicted disease.

Keywords: *Big Data; Data Mining; EHR; EMR.*

1. Introduction

The healthcare data includes data of clinical trials of patient, patient past data or patient history, description of disease that patient have, medicines taken by patient, doctors prescriptions, symptoms of other disease, medical treatment given or surgery done, patient insurance cover etc. This healthcare data is increasing at large speed day by day. This huge amount of data cannot be processed efficiently by using relational traditional system. The healthcare data is in different format like structured, unstructured, semi structured. The traditional databases are unable to manage structured, semi-structured, unstructured data by traditional system. Big Data tools can easily manage huge digital healthcare data. Because of all convenience and ease of use, if healthcare is combined with Big Data concept it will be more efficient for all concerned with healthcare. Big Data techniques can be used for healthcare data storage, data retrieval and data processing. By using Big Data concept for healthcare all the healthcare data will be available at one place for the healthcare service providers. Advantage of using Big Data concept is, it can handle efficiently all type of data which can be structured, unstructured or semi-structured.

Digital healthcare has emerged and proves to be beneficial in all aspects for healthcare. Due to digital

healthcare concept the healthcare has becomes more cheaper as all records are available and no need of repetitive medical services that is medical tests, also can be more efficient and convenient way for future healthcare services.

The digital healthcare records are available in form of (EMR) Electronic Medical Records. (MHR) Mobilized Health Records, (PHR) Personal Health Records, (EHR) Electronic Health Records. Most of the health data is now been stored digitally and thus the available digital health data can be stored efficiently and the non-digitized data such as doctor's prescription, nurse notes, treatment past history, patient past history, patient current treatment data in paper format can be converted into digital form.

The use of Big Data technique for healthcare data will be useful for the various healthcare related decision and also for automatic prediction of diseases based on symptoms matching with particular disease symptoms. The patient data can be stored into the system and the system will preserve the data of patient and give the data when needed. Due to this type of data storage all past data will be available to doctors or healthcare providers for better treatment. All the health information of patient will be available at one place and various doctors will be able to use this information for treatment of same patient.

A. Big Data Analytics Can Be Applied In Healthcare Sector In Many Areas Such As:

- Clinical trials: In clinical trials patient is been treated for a particular disease and those clinical trials or treatments are scheduled according to severity of disease. The data of clinical trials can be stored by concept of Big Data and also retrieved whenever needed. This trial data will be preserved for years and years, and can be retrieved anytime whenever patient history is needed.

- Public health: Huge amount of public health data can be stored. This data might be collected by various surveys. All this large amount of data can be used to derive specific result regarding public health. Various types of public needs can be identified, decisions regarding service provided can be taken, detect disease outbreak, preventing from diseases can be done by use of Big Data concept.
- Research and development: If all the data is available at one place in the digital form then various types of research can be done based on largely available digitized health data. By checking the occurrence of specific disease, various types of development can be done in order to provide better healthcare. As all records are available at one place the research and development is more easily done by Big Data concept.

In many hospitals there are systems for patient hospital bill, appointment management, etc. But such a system will store only little information and various patient data will not be stored. If a system for hospital is developed for storage of all patient health information and such a system can be used for disease prediction based on the data stored in it. Such a system can be cost effective as unnecessary tests can be eliminated and treatment can be done by proper way for exact disease. Early prediction of disease is useful in preventing major disease or defect to cause as treatment can be given earlier, before the occurrence of any disease. Such a disease can be predicted using the symptoms observed in the patient. Such a system can be beneficial to reduce medical errors, patient safety, decrease unwanted medical trials and improve patient health services.

2. Review of Literature

In [3] they have presented an intelligent and effective heart attack prediction methods using data mining. The data mining techniques in healthcare are discussed and those include classification data mining technique namely Rule Set classifiers, Decision tree algorithms like CART, ID3 C4.5, Neural Net- work, Neuro-fuzzy, Bayesian Network structure Discoveries. For data preprocessing used (ODANB) and (NCC2) which are extension of nave Bayes.

Classification is one of the data mining technique that classifies unstructured data into the structured class and groups and it helps to user for knowledge discovery and future plan [3].

Developed a prototype intelligent heart disease prediction system by using the three data mining techniques and those are decision tree, Nave Bayes and Neural Network. [5] Developed a system which can discover and extract hidden knowledge from the available data and historical heart disease databases. This system uses data mining extension (DMX), a SQL-style query language for data mining is used. It has 15 medical attributes which are obtained from Cleveland Heart Disease database. It includes total 909 records and those records are split into two equal dataset those are training dataset and testing dataset.

Medical field consist of large information but lack in knowledge and hence there is a need to develop a smart system which will give knowledge from available information. The system for prediction of heart disease is made by [10] and which may lead to control of occurrence of heart disease. They used C4.5 Algorithm, K-means algorithm, MAFIA algorithm. Various types of attributes are taken for examining the patient and comparing with disease symptoms.

By using various techniques it takes less time for prediction of disease with more accuracy. The Fuzzy Intelligent Techniques have improved the accuracy of heart disease prediction system. Use of Classification, Clustering, Prediction etc. benefit by early detection of various different diseases which can be predicted earlier and controlled by giving proper treatment.

Advantages of developing automated disease prediction system is that it prepare history database, can be used for training purpose, Detect if patient is having any disease, less time consuming.

The Feature selection field prove to be efficient for other sectors like data mining and also useful for the pattern recognition for machine learning. Feature selection is applied in following fields-image retrieval for various different medical images like scan images and text categorization for medical textual data. Feature selection includes selecting only the useful features from the given data. In feature selection there is removing of irrelevant features form the data set. The feature selection can be applied more efficiently by the use of clustering approach. According to the requirement of time and space various features can be selected from the set of all features. The idea of using feature selection is to minimize or cut down the unnecessary processing of unwanted features. This will definitely reflect in the outcome as processing is done only for the useful data. For Big data it is useful as the data is very large and out of that few features can only be useful and that features can be filtered using feature selection. By the use of feature selection there is automatic reduction in computational time required. The feature selection can also be done using the graphical clustering method. This approach gets most relevant features selected from the data given. [16]

The basic idea of the classification is to put related data into that section where the common data features are seen and that section have all data which is similar by some orientation. Classification can be done by either supervised or unsupervised technique. By the use of classification decision making process is boosted. The classification is mainly having two phases, first phase is learning process phase where a large amount of data is provides for training of classifier and this is the phase where classifier creates rules and patterns. In the second phase there is testing of that classifier which is trained in previous phase. [17]

3. Key Contribution

In the our system we are using EDT algorithm that is Ensemble decision tree algorithm for classification which gives more accurate results. Existing classification system used NB, SVM, LR, KNN algorithms. For the considered health datasets EDT algorithm gives best results. System uses ensemble learning approach by considering the algorithms giving better results. The results are compared based on accuracy of different algorithms. In our system we considers numeric data as well as medical images. In this system we used Adaptive rule base technique which gives classification rules. For medical images key points are extracted using SIFT algorithm.

4. Proposed Framework

The proposed system uses various algorithms for health data analysis.

The given figure represents the proposed system. It contains following phases- Storage, Feature Selection, Classification, Analysis, Searching, Decision.

Storage: In the proposed system the huge amount of healthcare data is taken to analyze health data. Such data is stored for further processing. This data is in different format. The stored data may contain doctors prescriptions, clinical test data, images, etc.

Feature Selection: From the available data only the useful features can be extracted, in order to continue further process. The extracted features are only taken into account for classification phase. Feature selection can be done using PCA algorithm that is Principle component analysis algorithm.

Classification: Based on some classification techniques such as KNN, SVM, Naive bayes the system will classify data and give outcome of classification. The classification is done on the extracted features and classified according the training data of symptoms. Classification can be done using SVM algorithm, KNN algorithm, etc. For Classifying health data we used ensemble learning ap-

proach which considers the results of multiple algorithms. The EDT algorithm gives the best results of classification.

Analysis: In this step the classified data is analyzed. Decision tree algorithm is used for analysis of the data. This analyzed data is then used for final decision making.

Clinical Decision Support (CDS): This is the final step, in this step final clinical decisions are given. Decisions such as whether patient is having particular disease or risk of having disease in future.

5. Methodology of Evaluation

To evaluate the flow of implemented system the experiments conducted are applied to dataset for disease prediction which will give decision for patient having particular disease. In our task we mainly concentrate to data classification using various algorithms and ensemble learning with high accuracy.

A. Dataset

Implemented system can work on numeric data, text data as well as medical image data. Various type of different disease have different symptoms and by considering these symptoms the patient data can be classified and analyzed so as to give decision regarding patient health condition. Dataset used in this system includes

1) Heart disease dataset: It contains 76 attributes but only 14 of them are used. Link- http://archive.ics.uci.edu/ml/datasets/heart+disease

2) Breast cancer dataset: It contains 32 attributes. Link- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

3) Appendicitis dataset: It contains 7 attributes. Link- http://sci2s.ugr.es/keel/dataset.php?cod=183

B. Evaluation parameter

In our experimental evaluation, we analyze the system performance under varying algorithms accuracy, for same data and attributes. Accuracy is measured for following algorithms- NB, SVM, LR, KNN, DT and EDT algorithm. That is Naive Bayes, Support Vector Machine, Logestic Regression, K- Nearest Neighbor, Decision Tree, Ensemble Decision Tree.

C. Experimental setup

Our experiments were implemented in Java, Eclipse is used as Integrated development environment and carried out on a PC with processor Pentium-IV and 2GB RAM.

Figure 2 represents the results based on accuracy of different algorithms at different number of attributes taken for heart disease dataset. X axis contains number of attributes taken and Y axis contains accuracy of prediction. It gives results by taking attributes ranging from 1 to 10. For 1 attribute SVM gives minimum accuracy and EDT gives maximum accuracy. For 10 attributes taken, DT gives minimum accuracy and EDT gives maximum accuracy. For all number of attribute EDT algorithm gives more accuracy.

Figure 3 represents the results based on accuracy of different algorithms at different number of attributes taken for breast cancer dataset. X axis contains number of attributes taken and Y axis contains accuracy of prediction. It gives results by taking attributes ranging from 1 to 29. For 1 attribute NB gives minimum accuracy and DT and EDT gives maximum accuracy. For more attributes also NB gives minimum accuracy and DT and EDT gives maximum accuracy. For all number of attribute DT and EDT algorithms gives more accuracy.

Figure 4 represents the results based on accuracy of different algorithms at different number of attributes taken for appendicitis dataset. X axis contains number of attributes taken and Y axis contains accuracy of prediction. It gives results by taking attributes ranging from 1 to 7. For 2 attributes DT gives minimum accuracy and EDT,KNN5,SVM gives maximum accuracy. For 7 attributes taken, NB and DT gives minimum accuracy whereas SVM gives maximum accuracy. For all number of attribute SVM algorithm gives more accuracy.

As the algorithm accuracy varies according to number of attributes taken for each dataset, we are using ensemble learning by considering more than one algorithms so as majority of algorithm's decision as taken into consideration

The system gets the decisions regarding the patient health condition for each patient. The system first stores the health- care data received from various health data sources. The stored data is in the form of attributes regarding health information. From the stored data useful attributes are selected with the feature selection phase. Once the feature used for processing is selected the classification of the health data is done. Classification is done only on the selected features. Analysis of the classified data is done to exactly give the predicted output of patient health condition.

In storage phase the health data which is in the form of text, numeric and images can be stored by Big Data techniques. In this we are storing the existing dataset in system. In feature selection phase features required for predicting disease is taken into consideration and rest of features are not considered. In classification phase data is classified for text, numeric and images for health data of patient. Analysis of classified data can be done in analysis phase. Searching module can be used to search for number of patients suffering from particular disease that can be done by observed similar symptoms of that disease. Final decisions for patient health condition is given in Clinical Decision Support phase.

The existing system for healthcare which are used are limited to only storing hospital bills and appointments to the doctor, but the proposed system is capable of storing all health data and also automatic prediction of patient health condition by prediction of disease to patient based on disease symptoms.

The system may also be utilized in other areas such as several types of disease system, as this is general approach for prediction and decision system, various healthcare aspects can be implemented using this system. This may include suggesting different types of exercises to the patient, etc. Along with healthcare this system flow can be used in LIC. As in LIC the insurance covered first checks the severity of diseases and thus by use of system it would be quick decision process and verified patient and diseases. The system can also be used for detecting the IQ of children by providing child history about intellectual data. The system will be able to count IQ of child and give decisions about specific child.

6. Discussion and Conclusion

Big Data plays important role in healthcare sector and so analyzing clinical data is necessary by storing various type of digitized EPR in Big Data. Also by using Big Data concept automated system that can predict the patient health condition can be developed which will reduce healthcare cost and time needed for repetitive unnecessary clinical tests. For training automated system various classification algorithms giver varied accuracy so we have used ensemble learning approach to increase accuracy by considering multiple algorithms. The system is trained based on the multiple instance of that particular disease symptoms. There are certain disease which are very rare and can be found in one patient amongst many. So training dataset is unavailable of that disease. This system can also be extended by training for such a rare disease prediction. System can also be deployed is multiple hospitals and system can be able for learning based on the different patient health condition.

Acknowledgement

I would sincerely like to thank our Professor Santosh Kumar, Department of Computer Engineering, SITRC., Nashik for his guidance, encouragement and the interest shown in this project by

timely suggestions in this work. His expert suggestions and scholarly feedback had greatly enhanced the effectiveness of this work.

References

[1] Asha Rajkumar1, Mrs. G.Sophia Reena,"Diagnosis Of Heart Disease Using Datamining Algorithm ",GJCST Classification, Vol. 10 Issue 10 Ver. 1.0 Sepetember 2010.
[2] Shweta Kharya," USING DATA MINING TECHNIQUES FOR DI- AGNOSIS AND PROGNOSIS OF CANCER EASE ",International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.
[3] Sellappan Palaniappan, Rafiah Awang,,"Intelligent Heart Disease Pre- diction System Using Data Mining Techniques ",IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008.
[4] Asif Adil ,"Analysis of Multi-diseases using Big Data for improvement in Healthcare ",2015 IEEE UP Section Conference on Electrical Computer and Electronics (UPCON) 978-1-4673-8507-7/15/31.00,2015 IEEE.
[5] Shankar Krishnan," Application of Analytics to Big Data in Healthcare ",2016 32nd Southern Biomedical Engineering Conference.
[6] Gemson Andrew Ebenezer J. and Durga S.," BIG DATA ANALYTICS IN HEALTHCARE: A SURVEY ",ARPN Journal of Engineering and Applied Sciences.
[7] Iroju Olaronke," Big Data in Healthcare: Prospects, Challenges and Resolutions",FTC 2016 - Future Technologies Conference 2016.
[8] Min Chen, Yixue Hao, Kai Hwang," Disease Prediction by Machine Learning over Big Data from Healthcare Communities"
[9] K.Srinivas,B.Kavihta Rani Dr.A.Govrdhan," Applications of Data Mining Techniques in Healthcare and Prediction of Heart tacks ",(IJCSE) International Journal on Computer Science and Engineering.
[10] M.A.Nishara Banu, B Gomathy," DISEASE PREDICTING SYSTEM USING DATA MINING TECHNIQUES",International Journal of Technical Research and Applications.
[11] Jirawan Niemsakul," Influencing Factor Analysis for Cost Benefit Shar- ing in Healthcare Supply Chain Collaboration"
[12] Satwik Sabharwal," Insight Of Big Data Analytics In Healthcare Indus- try",International Conference on Computing, Communication and Au- tomation (ICCCA2016) ISBN:978-1-5090-1666-2/16/31.00,2016 IEEE.
[13] Devendra Ratnaparkhi, Tushar Mahajan, ,Vishal Jadhav," Heart Disease Prediction System Using Data Mining Technique.",International Research Journal of Engineering and Technology (IRJET).
[14] V. Krishnaiah," Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A view ",International Journal of Computer Applications.
[15] ABHISHEK TANEJA," Heart disease Prediction System Using data Mining Techniques"
[16] Harshali D. Gangurde ,"Feature Selection using Clustering approach for Big Data ",International Journal of Computer Applications (0975 8887) Innovations and Trends in Computer and Communication Engineering (ITCCE-2014)
[17] PrafulKoturwar, SheetalGirase, Debajyoti Mukhopadhyay ,A Survey of Classification Techniques in the Area of Big Data.