International Journal of Engineering & Technology, 7 (1.3) (2018) 21-23

International Journal of Engineering & Technology

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

Research paper

Prediction of diabetes with hybrid prediction model using big data in health care

E. Rama Kalaivani [1]*, E. Ramesh Marivendhan [2], N. Suma [3]

[1]Asst. Professor/CSE, Karpagam College of Engineering, Coimbatore, India
[2]Asst. Prof./ECE, Dhanalakshmi Srinivasa College of Engineering, Coimbatore, India
[3]Professor/ECE, Dhanalakshmi Srinivasa College of Engineering, Coimbatore, India
**Corresponding author E-mail: ramakalaivani.e@gmail.com*

Abstract

Technology is advancing in healthcare to comply with unique regulatory guidelines designed to support public safety. The realistic way of Big Data can unify all patient related data to get a 360-degree view of the patient to analyze and predict outcomes. Big Data is a buzzword which is reigning the innovation market from quiet sometime and trends which can give birth to new line of treatment of diseases and provide high quality healthcare at lower cost to all. These issues include benefits of Big Data, its applications and opportunities in medical areas and health care. This paper includes the basics of big data, clinical prediction model for predicting whether the diagnosed patient suffers from diabetes. Hybrid prediction model is chosen and it is used to predict the disease.

Keywords: *Diabetes, Hybrid, Big Data, K-means*

1. Introduction

Big Data is a current need for human beings to generate data in explosive fashion. Big Data is usually very complex, focusing most of the areas like business, data mining and analysis. Big data is transforming business landscapes have been generated primarily from the e-commerce and communities. Big data analytics are horizontally scaled and focused on entire life cycle. It is mainly worked on real time data action and needs more automation. Big data analytics design method are based on agile based analytics technology and Big data is the data which exceeds the processing capacity of conventional database systems. Healthcare costs are much higher than they should be, and they have been rising for the past 20 years. Clearly, there is a need of some smart, data-driven thinking in this area. And current incentives are changing as well. Many insurance companies are switching from fee-for-service plans (which reward using expensive and sometimes unnecessary treatments and treating large amounts of patients quickly) to plans that prioritize patient outcomes. Big Data can unify all patient related data to get an overall view of the patient to analyze and identify the outcomes. It improves, new medicinal development and health care financing process It offers benefits such as early disease detection, and good healthcare with better efficiency. This paper introduces the Big Data concept and characteristics, major issues of Big Data and its applications and opportunities in medical areas and health care. The healthcare department has produced huge amount of data which is measured in petabyte /exabyte scale. The objective of healthcare department is to analyze the big quantity of data for unknown and useful facts, patterns, associations with help of various algorithms, which an give birth to new line of treatment of diseases. The goal is to provide better quality healthcare at lower cost to all. There are 5 V's of big data with relevant to healthcare are volume, veracity, variety, velocity and value. Volume is the quantity of the healthcare data in exabyte scale. The varied health care data are categorized in to three i.e. structured, semi structured and unstructured. Velocity refers to the speed at which every new data is generated and moves around. The sensor devices collect real time physiological data of patients at a medium velocity. Veracity means the trustworthiness of data. value means the valuable information that adds to creating knowledge in data. Big data in healthcare refers to electronic health data set which is so complex and difficult to manage with traditional management methods Data in health care is again categorized as Genomic Data which refers to genotyping. Clinical Data and Clinical notes. Structured data srefers to laboratory data,. Unstructured data refers to patient testing reports, patient discharge summaries. Semi-structured data (e.g., copy-paste from other related documents).Behavior Data and Patient Sentiment Data Web and social media data ,datas from Search engines, Internet consumer use and networking sites (Facebook, Twitter, Linkedin, blog, health plan websites and smart phone, etc.) Mobility sensor data or streamed data :datas from telehealth, sensor-based wireless and smart devices Administrative, Business and External Data Insurance claims and other related billing data Biometric data: Fingerprints, handwriting etc; In healthcare, we do have large volumes of data coming in. EMRs alone collect huge amounts of data. Most of that data is collected for recreational purposes . But neither the volume nor the velocity of data in healthcare is truly high enough to require big data today[4]. Our work with health systems shows that only a small fraction of the tables in an EMR database (perhaps 400 to 600 tables out of 1000s) are relevant to the current practice of medicine and its corresponding analytics use cases. So, the vast majority of the data collection in healthcare today could be considered recreational. Although that data may have value down the road as the number of use cases expands, there aren't many real use cases for much of that data today. The biggest difference between big data and relational databases is that big data doesn't have the traditional table-and-column structure that relational databases have.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In classic relational databases, a schema for the data is required (for example, demographic data is housed in one table joined to other tables by a shared identifier like a patient identifier). Every piece of data exists in its well-defined place. In contrast, big data has hardly any structure at all. Data is extracted from source systems in its raw form stored in a massive, somewhat chaotic distributed file system. The Hadoop Distributed File System (HDFS) stores data across multiple data nodes in a simple hierarchical form of directories of files. Conventionally, data is stored in 64MB chunks (files) in the data nodes with a high degree of compression[5], [7]. This is main purpose for approaching big data in health care.

2. Predicative analysis

Prediction of the presence of disease (diagnosis) or an event in the future course of disease (prognosis) becomes more and more important in the current era of personalized medicine. Predictive analytics uses various methods like statistical or machine learning method to identify a prediction about future It consists of text mining for unstructured data, and give path to perform the next step. It uses both historical and present data to predict future regarding activity, behaviour and trends[1]. To do this it makes use of statistical analysis techniques, analytical queries and automated machine learning algorithms. It doesn't expect anything about data but it allows the data lead the way. It uses statics, machine learning, neural computing, robotics, computational mathematics and artificial intelligence to explore all data and find meaningful relationships and patterns. Predictive analytics is a set of business intelligence (BI) technologies that uncover relationships and patterns within large volumes of data that can be used to predict behavior and events.

Predictive analytics require experts to build predictive models. It uses historical and present data to predict future regarding activity, behavior and trends. Supervised learning is a process of creating predictive models using a set of historical data and produce predictive results[1]. Examples are classification, regression and time-series analysis where as in Unsupervised learning does not use the previously known result to train its models. It uses descriptive statics. It identifies clusters or groups. The four phases of predictive analytics(taxonomy of predictive analysis) are

- Prediction.
- Monitoring.
- Analytics.
- Reporting.

Prediction is predicting what might happen in future. Presently what is happening is monitoring. To analyze why the result is happened is analyzing. Reporting is generating a report on what happened. Predictive analysis also includes various steps like defining the project, exploration of the project, preparation of data, building of models, deployment and model management. The following table gives an overview of what type of model is used in which situation.

Table 1: Usage of Predictive Model

Predictive Model	**Purpose**
Clustering algorithm.	Segmentation
Classification	Developing recommender system
Decision tree	Linear decision boundary
Regression algorithms	Predicting next outcome boundary and Predict continuous values
Machine Learning	Classifying text problems with ensemble model sometimes

3. Clinical prediction model

A clinical prediction model can be applied to various clinical scenarios: screening high-risk individuals for asymptomatic disease, predicting future events such as disease or death, and assisting medical decision-making and health education[3]. Clinical prediction models intimate diagnosed patients and their physicians or other healthcare providers of the patient's probability of having or developing a particular disease and help them with associated decision-making (e.g., facilitating patient-doctor communication based on more objective information). Applying a model to a real world problem helps with detection or screening in undiagnosed high-risk subjects that improves the capability to prevent developing diseases with early interventions. Furthermore, in some instances, certain models can predict the possibility of having future disease or provide a prognosis for disease (e.g., complication or mortality). The aim of prediction modeling is to develop an accurate and useful clinical prediction model with multiple variables using comprehensive datasets[1]. Development of clinical prediction analysis include the following steps

1. To perform initial data inspection
2. To apply coding of predictors
3. Specification of models
4. Estimating the models
5. Performing analysis in evaluating the model
6. Internal validation.
7. Presenting the model.

4. Hybrid prediction model

Diabetes is a chronic disease in which blood sugar levels are too high. Sugar from the foods where people eat comes, and insulin is the hormone that helps the entry of sugar into the cells to give the energy. In diabetes type I the body does not secrete insulin, and Type II diabetes, which is the type most prevalent, the body does not make insulin and cannot use them properly; it is an insufficient amount of insulin and sugar stays in the blood is present[2],[8] . Over time can result in the presence of too much sugar in the blood to the emergence of serious problems as it can damage the eyes, kidneys and nerves .Diabetes can lead to heart disease and stroke, and even to complications requiring amputation of one of the parties. To avoid all these serious complications, the blood glucose level should be under control. One of the strong indictors that measure the diabetic, is the accumulative blood glucose level for previous 3 months .This diabetes can be predicted for patients using hybrid clinical prediction model. In healthcare/medical field, large amount of information about patients' medical histories, symptomatology, diagnoses and responses to treatments and therapies is collected. Data mining techniques can be implemented to derive knowledge from this data in order to either identify new interesting patterns in infection control data or to examine reporting practices Earlier days patient history was maintained manually in hospital, but nowadays patient history is automated. With the advantage of that the patient datasets are collected. To improve the accuracy of classification, the data preprocessing is required. The data may be noisy, invalid, incomplete and inconsistent. In the beginning the dataset is preprocessed to remove noise points and missing values and then the data is normalized using z-score normalization [9]. The missing values of the data set that are considered for the experiment are denoted with the value zero. Calculate the difference between each attribute value and the mean value of the attribute and dividing by the standard deviation of the attribute. This is called z-score normalization. In any such situations, data type transformations are required[10]. It is performed since during classification or clustering the attributes may be scaled to fall within the given range of values and to generalize their values.

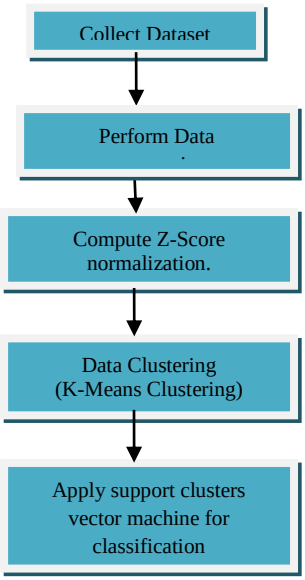

Fig. 1: Steps in Hybrid Prediction Model.

In Data clustering the clustering technique is k-means clustering to remove the outliers[7]. The k-means algorithm can be divided into two phases: the initialization phase and the iteration phase. The following are the steps used in K-means clustering.

Step1:Begin with a decision on the value of K=number of clusters.

Step 2:Put any initial partition that classifies data in to K clusters .you may assign the training samples randomly or systematically as follows,

1) Take the first K training samples as single element clusters.

2) Assign each of the remaining (N-K) training samples to the cluster with the nearest centroid .After each assignment; recompute the centroid of the gaining cluster

Step 3:Take each sample in sequence and compute the distance from the centroids of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.

Step 4:Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments[6]. These observations were analyzed when implemented K-means in data set .Performing Support Vector Machine on clusters: Feature selection plays a very significant role for the success of the system in fields like pattern recognition and data mining. Feature selection provides a smaller but more distinguishing subset compared to the starting data, selecting the distinguishing features from a set of features and eliminating the irrelevant ones. Our goal is to reduce the dimension of the data by finding a small set of important features that can give good classification performance. This results in both reduced processing time and increased classification accuracy. Feature selection algorithms are grouped into randomized, exponential and sequential algorithms[9]. The idea is to consider the feature important if it significantly influences the width of the margin of the resulting hyper-plane.

5. Conclusion

A hybrid model has been developed to predict whether the diagnosed patient may develop diabetes within 5 years or not. The procedure involves to preprocess the dataset, then compute Z-score values of features,, then k-means algorithm is used in data clustering to select feature subset finally Support Vector Machine is used for classification. This hybrid model has achieved higher accuracy. It can be implemented using patient data sets and WEKA tool.

References

[1] N. Jayanthi, B. Vijaya Babu1 and N. Sambasiva Rao "Survey on clinical prediction models for diabetes prediction",Journal of Big Data, DOI 10.1186/s40537-017-0082-7.

[2] Daniel Hartono Sutanto and Mohd. Khanapi Abd. Ghani, "Improving Classification Performance Of K-Nearest Neighbor By Hybrid Clustering And Feature Selection For Non-Communicable Disease Prediction", ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 16, September 2015. ISSN 1819-6608.

[3] Bellini, P., M. d. Claudio, P. Nesi and N. Rauch, "Tassonomy and Review of Big Data Solutions Navigation. In: Big Data Computing, Akerkar, R. (Ed.), Chapman and Hall/CRC, pp: 57-101.

[4] Tarig Mohamed Ahmed,"Developing A Predicted Model For Diabetes Type 2 Treatment Plans By Using Data Mining", Journal Of Theoretical And Applied Information Technology, 31st August 2016. Vol.90. No. E-ISSN: 1817-3195.

[5] Bottles, K. and E. Begoli, 2014. Understanding the pros and cons of big data analytics. Physician Exec., 40: 6-12.

[6] Lidong Wang and Cheryl Ann Alexander, " Big Data in Medical Applications and Health Care", American Medical Journal, DOI: 10.3844/amjsp.2015.1.8.

[7] J. Han, M. Kamber, and J. Pei. 2001. Data Mining Concepts and Techniques. 40(6).

[8] Dr.N.Suma, Sudharshan, S Manoj Kumar, P Suresh Kumar, S Naveen Kumar. 'Setting Up a LAN Connection with Port Security', International Journal of Advanced Research in Biology Engineering Science and Technology, Vol.2 (10), 2016 pp. 1579-1583.

[9] Dr.N.Suma and Dr.T.Purusothaman. "Design and Development of Reliable Energy based Efficient Protocol for Improving Fault Tolerance in MANET'. Asian Journal of Research in Social Sciences and Humanities, Vol.6 (11),2016 pp. 15-25.

[10] Swathi, E.Rameshm, Marivendhan, P.Rajasekar,"Low detector with less timing misalignment and improved signal quality",International Journal of Current Trends in Engineering & Research (IJCTER),e-ISSN 2455-1392 Volume 3 Issue 5,May 2017 pp.96-102.