ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 12, December 2018

Diabetes Mellitus Prediction in Big Data-Using Hadoop / Map Reduce Frame work (Survey)

K.S Praveenkumar[1], Dr. R Gunasundari[2]

Research Scholar, Department of CA, CS & IT, Karpagam Academy of Higher Education, Coimbatore, India[1]

Associate Professor, Department of CA, CS & IT, Karpagam Academy of Higher Education, Coimbatore, India [2]

Abstract: Diabetes Mellitus disease prediction is a growing research in healthcare. More over number of data mining methods have been applied to evaluate the main causes of diabetes, but only small sets of clinical risk factors are considered. So the results generated by such methods may not represent exact diabetes. We have to analyse number of factors such as Hereditary and genetics factors, Stress, Body Mass Index, Increased cholesterol level, High carbohydrate diet, Nutritional deficiency, Nature of Exercises, Tension and worries, High blood pressure, Insulin deficiency, Insulin resistance. Then we evaluate and compare this system using suitable rules and Map Reduce algorithm. The performance of the system is assessed in terms of different parameter like rules used, classification accuracy, and classification error. By considering all these parameters, the system can predict diabetics in a great accuracy. Also this paper surveys about different techniques and tools available in Big Data to predict Diabetes mellitus. Big Data can significantly diabetes research and ultimately improves the quality of health care for diabetics patients.

Keywords: Diabetes Mellitus ; Big Data, Hadoop/Map Reduce; C4.5 algorithm

I. INTRODUCTION

Diabetes mellitus (DM), is commonly known as diabetes. It is a group of metabolic disorders in which there are high blood sugar levels over a long period. Diabetes Mellitus is a condition wherein a person is either incapable of producing insulin or the body is not able to use the insulin present in the body. While many people assess diabetes as a disease with genetic vulnerability. Today it has become one of the most leading lifestyle diseases. There are three types of diabetes: Type 1 DM results from the pancreas's failure to generate sufficient insulin. Type 2 DM starts with insulin resistance, a condition in which cells fail to respond to insulin properly. As the disease continues a shortage of insulin may also develop. The most common cause is immoderate body weight and insufficient exercise. Gestational diabetes is the third main form, and occurs when pregnant women without a previous history of diabetes increase high blood sugar levels. Prohibition and treatment involve maintaining a healthy diet, regular exercise, a normal body weight, and avoiding use of tobacco. Control of blood pressure and maintaining proper foot care are important for people with the disease.

According to the official WHO data, India is in the top position, where the countries with the highest number of diabetics; China, America, Indonesia, Japan, Pakistan, Russia, Brazil, Italy and Bangladesh follow. Also in the year 2030 the number may reach 79.4 million. Definitely, persons with high weight have more chance of falling diabetes. Therefore, apart from eating healthy food, keep regular physical activity is the most important way to overcome lifestyle diseases like diabetes.

II. BIG DATA

Big data is a word that describes the large volume of structured, unstructured and semi-structured data. In the case of big data, amount of data is not important, but what organizations do with the data is important. The proper use of big data can be lead to generate better decisions and predictions. The fields involve Big Data is increasing day by day. In all terms we can say, we have entered the era of Big Data. For the definition of the Big Data [1], there are many different explanations from 3Vs to 4Vs. According to Doug Laney, we can use 3Vs to define big data. They are volume, velocity and variety. The term volume is the size of the data set, velocity indicates the speed of data in and out, and variety defines the range of data types and sources. Some people extend another V according to their special requirements. The fourth V can be value, variability, or virtual. More commonly, Big Data is a collection of very huge data sets with large diversity of types so that it becomes difficult to process by traditional data processing platforms.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 12, December 2018

We have to collect data from different verticals of a person such as, medical history, family history of diabetics, value of body mass index, nature of blood pressure, nature of food habit. All these data together forms our experimental data.

III. HADOOP

Hadoop is one of the open-source distributed data processing platform from Apache[2]. Hadoop has the power to process immensely huge amounts of health data by utilizing distributed data processing platform of hadoop. Hadoop uses two main components to do its job: Map/Reduce and Hadoop Distributed File System.

IV. MAP/REDUCE: HADOOP'S MAP/REDUCE

MAP/REDUCE: HADOOP'S MAP/REDUCE implementation is based on programming models to process huge data or datasets by dividing the data into small blocks of tasks. Map/Reduce uses distributed algorithms, on clusters to process the datasets. It consists of two functions: The Map ( ) function which resides on the master node and then divides the input data into smaller subtasks. This divided data is distributes to worker nodes. The worker nodes process the data and pass the answers back to the master node. The subtasks are run parallelly on multiple computers. The Reduce ( ) function collects the results from distributed nodes and combines them to generate final result. This final result will be the answer to the original problem.

V. Hadoop Distributed File System (HDFS)

HDFS replicates the data blocks which reside on other computers in the data center and manages the data transfer between various parts of the distributed system. The HDFS system then distributes data across a network or it migrate data if necessary.

VI. C4.5

C4.5 is an algorithm used to create decision tree and it is developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees created by C4.5 can be used for classification, because of this reason; C4.5 is used as a statistical classifier. At each node of the tree, C4.5 chooses most suitable data as the node attribute to form the tree. In our study we want to handle different types of data. These data can be obtained from various Electronic Health Record (EHR) / Patient Health Record (PHR), Clinical systems and external sources (individuals, laboratories, pharmacies, insurance companies etc.), in various formats and residing at various locations.

VII. REVIEW OF LITERATURE

A literature review reveals many results on diabetes carried out by different methods and materials of diabetes problem in India. Many people have developed various prediction models using data mining to predict diabetes.

Dr. Saravana Kumar NM, Eswari T, Sampath P, and Lavannya S published a paper about Predictive Methodology for Diabetic Data Analysis in Big Data [3] in 2015. In their study they used the predictive analysis algorithm in Hadoop/Map reduce environment to predict the diabetes types prevalent, complications associated with it and the type of treatment to be proved. The architecture of their predictive analysis system used raw diabetic big data as the input to the system. This system uses the predictive analysis algorithm in Hadoop/ Map Reduce environment to predict and classify the type of diabetic and the type of treatment to be provided. They used different test patterns like, plasma glucose concentration, serum insulin, diastolic blood pressure, diabetes pedigree, Body Mass Index, age, number of times pregnant. They included different tools to correlate different data set. The system used association rule mining for creating association between diabetic types and laboratory results, used clustering to identify similar patterns in medical data, and different classification methods for representing health condition of the patient, and also used some statistical approaches to convert data into values.

Purushottam, Dr. Kanak Saxena, and Richa Sharma [4] designed a system that can predict diabetes based given number of parameter their health. They evaluate and compare the system by using C45 rules and partial tree. The system's performance is evaluated in terms of different factors like rules generated accuracy of classification, errors occurred in classification, problems faced in experimental results. By considering all these facts they could predict diabetes diseases up to 81.27 correctly. In their study they used eight different attributes of patient. The attributes or parameters are, number of times pregnant, concentration of plasma glucose rate, blood pressure(mm Hg), triceps skin fold thickness(mm), serum insulin amount(mu U/ml), body mass index, diabetes pedigree , age in years, also used class variables (0 for tested negative for diabetes and 1 for tested positive for diabetes). They used KEEL (Knowledge

Extraction based Evolutionary Learning) tool to implement this model. Their results shows that C4.5 classifier can correctly classified the diabetes up to 81.27%.

Mr. K. Rajesh and Ms.V Sangeetha published a paper, Application Data Mining Methods and Techniques for Diabetes Diagnosis [5] on IJEIT. In their research work , they have applied data mining techniques to classify diabetes data and predict the chance to being a diabetic patient or not. In their work the primary dataset was Pima Indians Diabetes Database. It contains 768 record sample , each having eight attributes such as, number of times pregnant, concentration of plasma glucose rate, blood pressure(mm Hg), triceps skin fold thickness(mm), serum insulin amount(mu U/ml), body mass index, diabetes pedigree , age in years, also used class variables (0 for tested negative for diabetes and 1 for tested positive for diabetes). They applied number of classification methods to diabetes dataset and the error results obtained. Some they are C-RT, CS-RT, C4.5, SVM, RND TREE. In the used classification algorithms RND TREE gives 100% accuracy but the rule set is huge and it is suffering from over fitting of data. Also the C4.5 gives 91% classification since it is mainly used for medical applications and it is the well-known decision tree induction learning techniques that can apply for medical data processing.

In the paper Diabetic Data Analysis In Big Data With Predictive Method[6], Thanga Parasad S, Sangavi S, Deepa A, Sairabanu F and Ragasudha R, proposed diabetes prediction using big data and Hadoop Distributed File System. They used big data for handling large collection of different types of data related to diabetes. In the work Map Reduce is used as a programming model provided by hadoop that allows to express distributed computations on large amount of data. In their predictive analysis of the system architecture that includes different levels such as predictive analysis, data collection, data processing and different report analysis. Their system uses the predictive investigation algorithms during hadoop/map reduce to guess various categories of diabetes mellitus.

Sadhana, and Savitha Shetty[7], in their paper made an attempt to make analysis of diabetic data set using Hadoop and R. They used eight attributes such as number of times pregnant, plasma glucose concentration, serum insulin, BP, diabetes pedigree, BMI, age and triceps skin fold thickness. A detailes analysis of diabetic dataset was carries out with the help of hive and R.

Sabibullah M, Shanmugasundaram V, & Raja Priya K[8], developed a soft computing based prediction model for finding the risks accumulated by the diabetic patients. They have experimented with real time clinical data using Genetic Algorithm. The obtained results pertaining to the level of risk which prone to either heart attack or stroke.

All the above researchers have been successful in analysing the diabetic data set and they developed good prediction methods. But all these tools or methods are considered only minimum parameters It contains structured and unstructured data or multidimensional data.

VIII. CONCLUSION

Big Data Analysis in Hadoop's implementation provides systematic way for achieving better outcomes like availability and affordability of healthcare service to all population. Non-Communicable Diseases like diabetes, is one of a major health hazard in India. According to the official WHO data, India tops the list of countries with the highest number of diabetics; China, America, Indonesia, Japan, Pakistan, Russia, Brazil, Italy and Bangladesh follow.. The goal of this research is to predict diabetics. The design of this system of diabetic treatment may give enhanced data and analytics yield the greatest results in healthcare. Treatment can be offered when it is identified in advance.

REFERENCES

1. C.L Philip Chen, Chun-Yang Zhang, Data intensive application, challenges, techniques and technologies: A survey on Big Data, Information Sciences 275 (2014) 314–347
2. Big Data (Covers Hadoop 2, MapReduce, Hive, YARN, Pig, R and Data Visualization) Black Book, Authored By DT Editorial Services. Pages 83 – 114
3. Predictive Methodology for Diabetic Data Analysis in Big Data. Dr Saravana kumar N M, Associate Professor, Dept of CSE, Bannari Amman Insitute of Technology,Sathyamangalam. Eswari T , 2,4Assistant Professor, Dept of IT, Sri Krishna College of Engineering&Techechnology,Coimbatore. Sampath P, Associate Professor, Dept of CSE, Bannari Amman Institute of Technology, Sathymangalam. Lavanya S, Assistant Professor, Dept of IT, Sri Krishna College of Engineering & Techechnology,Coimbatore. Procedia Computer Science 50 (2015) 203 – 208
4. Purushottam, 3Amity University, Noida. Dr. Kanak Saxena, S.A.T.I. Vidisha, M.P. Richa Sharma, Amity University, Noida: Diabetes Mellitus Prediction System Evaluation Using C4.5 Rules and Partial Tree. 978-1-4673-7231-2/15/ ©2015 IEEE
5. Application Data Mining Methods and Techniques for Diabetes Diagnosis, International Journal of Engineering and Innovative Technology (IJEIT), Volume 2, Issue 3, September 2012. Mr K Rajesh, M.E in Computer Science and Engineering at Rajalakshmi College of Engineering, Chennai. Ms. V Sangeetha, Asst. Professor in department of IT at Rajalakshmi Institute of Technology, Chennai.

6. Diabetic Data Analysis In Big Data With Predictive Method, Thanga Parasad S, Asst.Professor, Department of CSE PMC tech, India , Sangavi S, Deepa A, Sairabanu F and Ragasudha R, Department of CSE PMC tech, India.
7. Sadhana, Savitha Shetty, "Analysis of Diabetic Data Set Using Hive and R", International Journal of Emerging Technology and Advanced Engineering, vol 4(7), 2014.
8. Sabibullah M, Shanmugasundaram V, Raja Priya K, "Diabetes Patient's Risk through Soft Computing Model", International Journal of Emerging Trends & Technology in Computer Science, vol 2(6), 2013

BIOGRAPHIES

Mr. K.S Praveenkumar, Research Scholar, Department of CA, CS & IT, Karpagam Academy of Higher Education, Coimbatore and Assistant Professor, Department of Computer Science, SNGIST Arts and Science College, Ernakulam.

Dr. R Gunasundari, Associate Professor, and Research Guide, Department of CA, CS & IT, Karpagam Academy of Higher Education, Coimbatore.