ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 12, July 2017

Analysis of Diabetic Patient Medical Records using Machine Learning Algorithms and Hadoop

Govindaraju G N [1], B. K. Raghavendra [2]

Research Scholar, VTU Belagavi, Department of Computer Science and Engineering, K S School of Engineering and Management, Bengaluru, India [1]

Professor & Head, Department of Computer Science and Engineering, K S School of Engineering and Management, Bengaluru, India[2]

ABSTRACT: Diabetes is a disease that results in too much sugar in the blood, or high blood glucose. Prevalence is increasing worldwide, particularly in low and middle income countries. Access to quality health care for these people makes diabetes a dangerous disease. The traditional system involves a tedious process of multiple lab assessments and assumptions based on certain guidelines by the doctor. The existing testing techniques do not extensively cover all the required aspects to diagnose the condition of diabetes and are often time consuming. In this paper we are proposing a analysis and prediction of new patient record with the large set of other patients records using Big Data and Machine Learning algorithms like Naive-Bayes and k-means. With the severity of risk the doctors can advise the further course of action.

KEYWORDS: Diabetes, Machine Learning, Big Data, Hadoop, Prediction

I. INTRODUCTION

Big Data is emerging as a solution to the problems associated with large amount of data. The large amount of data generated can now be used in order to provide an inner view of what is really taking place and spot the emerging trends. Big Data can also be used in the field of health care in order to make the system more effective. Big Data refers to the large amount of data which may be structured or unstructured and cannot be processed using a relational database model. Unstructured data refers to the data that cannot be stored in a particular row and column format. Big Data also goes beyond the processing capacity of the conventional database systems [1].

Healthcare sector data is growing beyond the dealing capacity of the health care organisations and is expected to increase in the coming years. Most of the Healthcare data is often unstructured, and resides in imaging systems, medical prescription notes, insurance claims data, Electronic Patient Record etc. Combining structured and unstructured data for advanced analytics is critical to improve healthcare results. Because of data that are isolated in disparate or incompatible formats or due to the lack in processing capability to load and query large datasets in a timely fashion the Healthcare organisations are not in a position to leverage the benefits of the large set of Healthcare data. With the help of advanced computing and numerous Big Data technologies like Cloud Computing , Hadoop, and Machine Learning algorithms it is possible to attain high performance, scalability at a relatively low cost. Big data solutions often come with set of innovative data management solutions and analytical tools, when effectively implemented can transform the healthcare outcomes [2].

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 12, July 2017

The Healthcare data is growing rapidly from an internal as well as an external sources, from mobile devices, wearable sensor devices, Electronic Health Records, Radiology images, Videos, clinical notes, social media, blogs, remote health monitoring devices etc. The other medical data forms like imaging, sensor reading is also fuelling to the need of Big Data solutions to manage these massive data available in the healthcare organisations. The health care industry need to work on prediction, prevention and personalisation to improve their outcomes. Diabetes also known as Diabetes Millet's is a disease that result in too much sugar in the blood, or high blood glucose. Diabetes of all types can lead to complications in many parts of the body and increase the risk of dying prematurely. In 2012 diabetes was the direct cause of 1.5 million deaths globally. A large proportion of diabetes and its complications can be prevented by a healthy diet, regular physical activity, maintaining a normal body weight and avoiding tobacco use. Prevalence is increasing worldwide, particularly in low and middle income countries. Access to quality healthcare for these people makes diabetes a dangerous disease [3].

By using Big Data, it is possible to predict the risk involved for the patient using his/her previous medical history. Healthcare providers are digitising their databases which pave way for the emergence of Big Data analytics. Using algorithms like Naive-Bayes and k-means the prediction of risk involved can be done. The prediction would enable the healthcare providers to quickly assess the patient's situation and also provide an insight into patient's future if the current situation prevails as diabetes is a disease which affects the patient. The risk involved can be assessed by doctors and can base their treatment and also the patient can be advised for lifestyle changes [4].

II. RELATED WORK

The healthcare sector relies on the data obtained from the various methods of diagnosis like physical examination, diagnostic tests and so on. Traditionally the sector has faced a problem of handling the data generated. The modern technology has enabled to digitally collects and stores the complex data generated by the sector. The enormous data generated enables us to apply the concepts of big data in healthcare sector. The healthcare sector can become a major benefactor of application of big data. For instance physicians at Harvard Medical School and Harvard Pilgrim Health Care recently demonstrated the potential for computer algorithms to analyse Electronics Healthcare Record (EHR) data to detect and categorise patients with diabetes for public health surveillance [5].

The researchers used algorithms to review 4 years of HER data. A large multi-site, multi-speciality ambulatory practice, and flag patients suspected of having diabetes based on lab results, diagnosis codes, and prescriptions. The algorithms successfully identified patients who had been clinically diagnosed with diabetes, and were able to distinguish between patients with Type I and Type II diabetes. The algorithms were more effective at identifying patients with diabetes than reviewing diagnosis codes alone[6][7].

III.METHODOLOGY

The modern healthcare provider is equipped with an Electronic Healthcare Records and an automation tool has enabled enormous amounts of data generation. This collected data can be used to implement big data analytics. Apart from basic data being collected modern systems also collect complex data from clinical trials, research and diagnostic tests.

Map Reduce is a programming model for parallel processing of large volume of data. This data can be anything but it is specifically designed to process the list of data. The main theme concept of Map Reduce is to transform list of input data to list of output data. If the input data is not readable then it would be difficult to understand large data input sets. In that case we need a model that can mould input data into readable, understandable output list. In recent years several nontrivial Map Reduce algorithms have emerged, from computing the diameter of a graph to implementing the Expectation–Maximisation (EM) algorithm to cluster massive data sets. Each of these algorithms gives some insights into what can be done in a Map Reduce framework. However, there is a lack of rigorous algorithm analyses of the

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 12, July 2017

issues involved. In this work the authors have presenting a formal model of computation for Map Reduce and compare it to the popular PRAM model [8] [9].

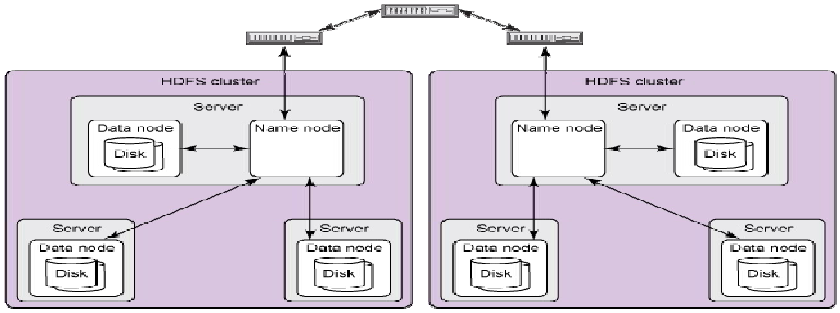

Figure 1: Architecture of Hadoop

The Map Reduce framework maps the <key, value> pairs and produces a set of output pairs of merging all the pairs associated intermediate key, that is, Map Reduce works exclusively on <key, value> pair. The two operations in Map Reduce i.e. map and reduce come from functional programming languages which pass functions as arguments to other functions. Map Reduce framework splits the data into segments. These segments are then passed to different machines/clusters for computation. Map script, which is written by user, runs on each machine to process portion of set of intermediate <key, value> pair. The pairs with same key are grouped together by the Map Reduce library and passed reduce function for aggregation. Reduce script, which is also written by user, takes collection of intermediate key I along with their set values for that keys and merges them according to the user specified script into a smaller set of values. Thus, the reduce function aggregates the values of the collection of intermediate <key, value> pairs having the same key. The diagrammatic representation of work flow of Map Reduce is as shown in Figure 2[10] [11].

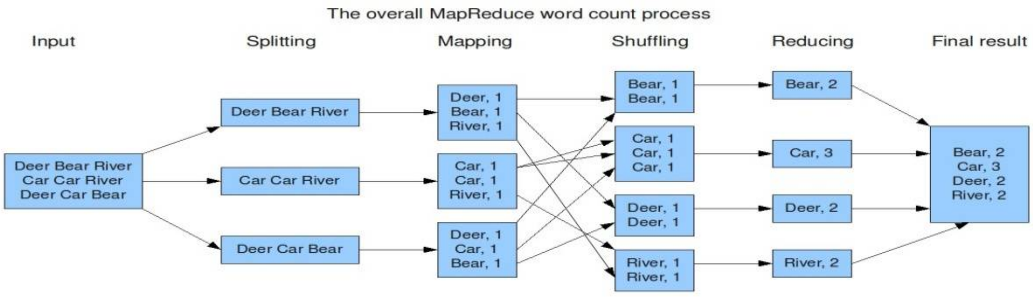

Figure 2: Work flow of Map Reduce

With HDFS and HBase security to make sure that only approved users can operate. Resource Manager employs data locality and server resources to determine optimal computing operations. It provides flexibility to write Procedures virtually in any programming languages. Completes job according to optimised scheduled priority [12].

IV. EXPERIMENTAL RESULTS

The steps involved in data processing are Information Extraction, Feature Selection and Predictive Modelling Information Extraction: Here patients health care records are collected from various sources in hospitals and then organised as a Structured Electronic Healthcare Record (S-EHR). Feature Selection: From the collected patient's record, the most important features required for diabetic prediction and modelling is extracted. Predictive Modelling: We have constructed a predictive model to predict whether the patient have diabetes or not.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 12, July 2017

Regression model is used to provide the effect and dosage of insulin. Cluster the users into groups based on the similarity, so that it is easy for doctors to analyse and recommend medicine to them.

The Feature Subset Selection reads the training file and calculates the Entropy of each attributes present in the Training File. Sorts the Entropy of all attributes in descending order and selects the specified numbers of attributes with highest entropy value. The Naive Bayes algorithm is used to construct classification models. All Naive Bayes classifiers assume that the value of a particular feature is independent of value of any other feature with given the class variable. Clustering Algorithm Provides input stream, reset input stream, find the nearest points in the input, cluster the points, find the nearest cluster and re-compute the clusters as an average. Clustering is done using the similarity of the patient attributes and it is achieved by calculating the K-means value.

The output obtained after clustering is shown in Figure 3. The numbers at the end denote the class to which the particular patient belongs. This is done using the similarity of attributes. The Figure 4 shows the prediction of diabetes in single patient. The Figure 5 shows the output of the risk involved for the patient. The numbers at the end gives us the severity of the condition of the diabetes. No risk, low, medium and high risks are denoted by 0,1,2,3 respectively .The Figure 6 shows the detection time taken. The time varies directly with the size of the dataset that is given as input.

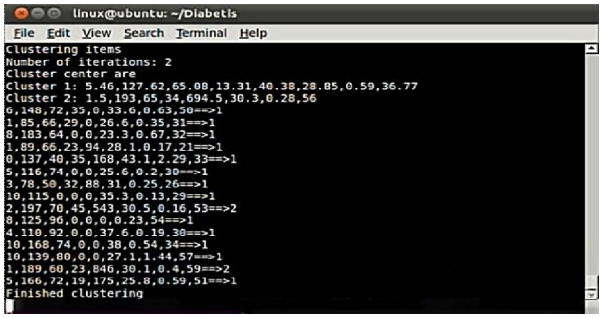

Figure 3: Patient clustering output

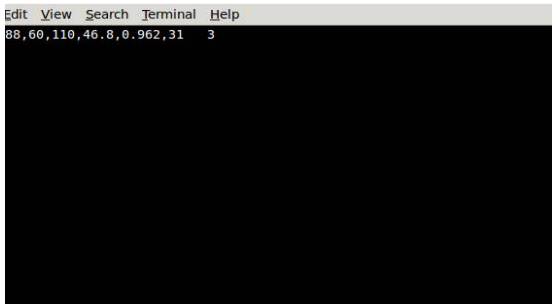

Figure 4: Single patient diabetes prediction output

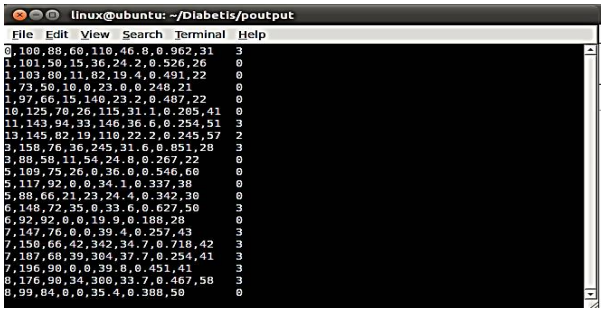

Figure 5: Diabetes Prediction output

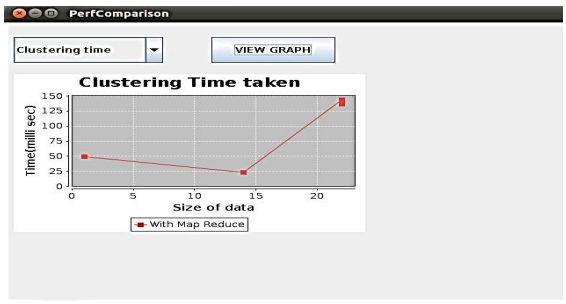

Figure 6: Detection Time Taken

V. CONCLUSION

Big data and Machine Learning Algorithms are used to analyse and predict the severity of risk involved in the diabetic patients using the large set of patient medical records. Hadoop Distributed File System is used to store the large set of data and Map Reduce programming model is used to analyse the massive amount of patient data sets in parallel programming method. Using this analysis it is possible to identify patients likely to suffer from diabetic risk and diagnose the patients with in less time. This enhanced medical data analytics yield the greatest results in healthcare.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 12, July 2017

REFERENCES

[1] N. W. T. Marty Kohn, "Transforming Healthcare Through Big Data," IEEE, vol. 18, no. 6, pp. 21-45, 2015.
[2] P. H. Marco Viceconti, "Big Data, Big Knowledge: Big Data for Personalized Healthcare," IEEE Journal, vol. 16, no. 4, pp. 1209-1215, 2015.
[3] A. K. Chaitanya Kaul, "Comparative Study on Healthcare Prediction Systems using Big Data," IEEE Journal, vol. 5, no. 4, pp. 1-7, 2015.
[4] K. K. a. R. Rani, "Managing Data in Healthcare Information Systems," IEEE Journal, vol. 48, no. 3, pp. 52-59, 2015.
[5] O. Shenyin, "Big Data for Modern Industry: Challenges and Trends," IEEE Journal, vol. 103, no. 2, pp. 143-146, 2015.
[6] G. C. Tung Chee-chen, "A Method for Calculating the probability of Diabetes based on large data," IEEE Journal, vol. 27, no. 9, pp. 13-17, 2014.
[7] S. Saria, "A Trillion Challenge to Computational Scientists Transforming Healthcare Delivery," IEEE Journal, vol. 29, no. 4, pp. 82-87, 2014.
[8] U. Srinivasan, "Anomalies Detection in Healthcare Services," IEEE Journal, vol. 16, no. 6, pp. 12-15, 2014.
[9] S. G. Jeffry Dean, "MapReduce: Simplified Data Processing on Large Clusters," IEEE Journal, vol. 2, no. 3, pp. 75-82, 2014.
[10] B. A. Uma Srinivasan, "Leveraging Big Data Analytics to Reduce Healthcare Costs," IEEE Journal, vol. 15, no. 6, pp. 21-28, 2013.
[11] R. B. Ragunath Nambiar, "A Look at Challenges and Opportunities of Big Data in Healthcare," IEEE Journal, vol. 4, no. 3, pp. 17-22, 2013.
[12] V. S.D, "A study on Role of Hadoop in Information Technology era," IEEE Journal, vol. 2, no. 2, pp. 26-30, 2013.