Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 203 – 208

Procedia Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Predictive Methodology for Diabetic Data Analysis in Big Data

Dr Saravana kumar N M[1] *, Eswari T[2] , Sampath P[3] & Lavanya S[4]

[*1] *Associate Professor, Dept of CSE, Bannari Amman Insitute of Technology,Sathyamangalam – 638401,India.*
[3] *Associate Professor, Dept of CSE, Bannari Amman Institute of Technology, Sathymangalam – 628401,India.*
[2,4] *Assistant Professor, Dept of IT, Sri Krishna College of Engineering & Techechnology,Coimbatore-641008,India.*

Abstract

Modernizing healthcare industry's move towards processing massive health records, and to access those for analysis and put into action will greatly increases the complexities. Due to the growing unstructured nature of Big Data form health industry, it is necessary to structure and emphasis its size into nominal value with possible solution. Healthcare industry faces many challenges that make us to know the importance to develop the data analytics. Diabetic Mellitus (DM) is one of the Non Communicable Diseases (NCD), is a major health hazard in developing countries such as India. The acute nature of DM is associated with long term complications and numerous of health disorders. In this paper, we use 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 provided. Based on the analysis, this system provides an efficient way to cure and care the patients with better outcomes like affordability and availability.

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Keywords: Healthcare industry; Hadoop/Map Reduce; Big Data; Predictive analysis;

1. Introduction

In recent years the healthcare industry has generated large amounts of data. The value based treatment in hospitals and digitization of world likes to have the computerized data rather than hard copy form. The health care data includes Electronic Health Reports (EHR) of patients data, clinical reports, doctor's prescription, diagnostic reports, medical images, pharmacy information, health insurance related data, data from social medias and medicinal journals [13]. All these information collectively forms Big Data in health care. By employing the analysis of big data will produce the predicted results for understanding the trends to improve the health care and life time expectancy, proper treatment at early stages at low cost. The analytics associated with big data is described by four characteristics: volume, velocity, variety and veracity [8]. The accumulation of health-related data continuously,

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
doi:10.1016/j.procs.2015.04.069

resulting in an incredible volume of data; Velocity is accessing those data in real-time at a rapid speed; Variety includes diabetic glucose measurements, blood pressure readings, and various EHRs; Whereas veracity assumes the simultaneous scaling up in performance of the architectures and platforms, algorithms and tools to match the need of big data [8]. The healthcare industry is moving from reporting facts to discovery of insights, toward becoming data-driven healthcare organizations. Big data holds great potential to change the whole healthcare value chain from drug analysis to patients caring quality.

The probability of a 30-70 year old Indian dying from the four main non-communicable diseases - diabetes, cancer, stroke and respiratory diseases - is 26 percent at present, according to the World Health Organization. According to the Global Status Report, Non-Communicable Diseases (NCDs) would claim nearly 52 million lives globally by the year 2030. Nearly 8.5 million people died of NCDs diseases in the WHO's South-East Asia Region in 2012. In India, NCDs are estimated to have accounted for 60 percent of all deaths in 2014, while 26 percent between the ages of 30-70 years had a probability of succumbing to the four diseases.

Diabetic Mellitus (DM) is one of the Non Communicable Diseases (NCD), is a major health hazard in developing countries such as India. The acute nature of DM is associated with long term complications and numerous of health disorders. There are three main types of this disease. Type1 DM results from the body's failure to produce insulin, and presently requires the person to inject insulin. This form is referred as Insulin - Dependent Diabetes Mellitus (IDDM). Type 2 DM results from insulin resistance, a condition in which cells fail to use insulin properly, sometimes combined with an absolute insulin deficiency. This form was previously referred to as Non-Insulin - Dependent Diabetes Mellitus (NIDDM). The third main form, gestational diabetes occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level. It may precede development of type 2 DM. It was estimated that 61.3 million people aged 20-79 years live with diabetes at 2011 in India. This number was expected to increase to 101.2 million by 2030.

Due to the growing unstructured nature of diabetic data form health industry or all other sources, it is necessary to structure and emphasis its size into nominal value with possible solution. With the help of technological developments, it is necessary to combine robust diabetic data sharing and electronic communication systems can facilitate better access to health services at all the levels of patients. So that all patient information needs to be in one repository. Deploying a Health Information Exchange (HIE) can extract clinical information from several disparate repositories and integrate that data within a single patient health record that all care providers can access securely. Predictive analysis is a method, that incorporates a variety of techniques from data mining, statistics, and game theory that uses the current and past data with statistical or other analytical models and methods, to determine or predict certain future events [7]. Significant predictions or decisions can be made by employing big data analytics in health care field. In this paper, we use 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 provided. Based on the analysis, this system provides an efficient way to cure and care the patients with better outcomes like affordability and availability.

2. Existing Work

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. Combination of classification-regression-genetic-neural network, handles the missing and outlier values in the diabetic data set, and also they replaced the missing values with domain of the corresponding attribute [1]. The classical neural network model is used for prediction, on the pre-processed dataset.

In predictive analysis of diabetic treatment using regression based data mining techniques to diabetes data, they discover patterns using SVM algorithm that identify the best mode of treatment for diabetes across different age [2]. They concluded that drug treatment for patients in the young age group can be delayed whereas; patients in the old age group should be prescribed drug treatment immediately. Prediction and classification of various type of diabetes

using C4.5 classification algorithm was carried out in Pima Indians Diabetes Database [3]. A detailed analysis of the Pima diabetic data set was carried out efficiently using of Hive and R. In this analysis we can derive some interesting facts, which can be used to develop the prediction models [4].

The soft computing based prediction model was developed for finding the risks accumulated by the diabetic patients. They have experimented with real time clinical data using Genetic Algorithm [5]. The obtained results pertaining to the level of risk which prone to either heart attack or stroke. The novel pre-processing phase with missing value imputation for both numerical and categorical data. A hybrid combination of Classification and Regression Trees (CART) and Genetic Algorithms to impute missing continuous values and Self Organizing Feature Maps (SOFM) to impute categorical values was improved in [6].

Deploying a health information exchange (HIE) repository promote and integrate the data within a single point of robust data sharing. This sharing of information and electronic communication systems enable access to health services and also promotes additional care over dual eligible patients. It recognizes which patient is requiring more care and attention than others. It gives needed data to determine which strategies should be put in place to maximize positive behavior modification [9].

The predictive analytics works in three areas such as Operations management, Medical management and biomedicine, and System design and planning. Healthcare predictive analytics system can help one of the issues that is to address the cost of patients being repeatedly admitted and readmitted to a hospital for chronic diseases which is similar or multiple. The survey of New England Journal of Medicine tells that one in five patients suffer from preventable readmissions. Therefore,1% of the population accounts for 20% of all US healthcare expenditures almost and 25% for over 80% of all expenditures [10].

Various big data technology stack and research over health care combined with efficiency. cost savings, etc.,are explained in better healthcare [11]. The hadoop usage in health care became more important to process the data and to adopt the large scale data management activities. The analytics on the combined compute and storage can promote the cost effectiveness to be gained using hadoop [12].

All the above researchers have been successful in analysing the diabetic data set and developing good prediction models. In this paper, we use the predictive analysis technique in Hadoop/Map Reduce environment to predict and classify the type of diabetes. This system provides efficient way to care and cure the patients at low cost with better outcomes like affordability and availability.

3. Predictive Analysis System Architecture

The architecture of predictive analysis system includes various phases like data collection, data warehousing, predictive analysis, processing analyzed reports. Figure 1 shows the complete architecture of proposed method.

3.1 Data Collection

The raw diabetic big data or data set is given as input to the system. The unstructured voluminous input data can be obtained from various Electronic Health Record (EHR) / Patient Health Record (PHR), Clinical systems and external sources (government sources, laboratories, pharmacies, insurance companies etc.), in various formats (flat files, .csv, tables, ASCII/text, etc.) and residing at various locations [8].

3.2 Data Warehousing

In this phase massive unstructured data warehoused into single unit in which, data from various sources is cleansed, accumulated and made ready for further processing. Integration of various EHRs can help in identifying the patterns for diabetes prediction system.

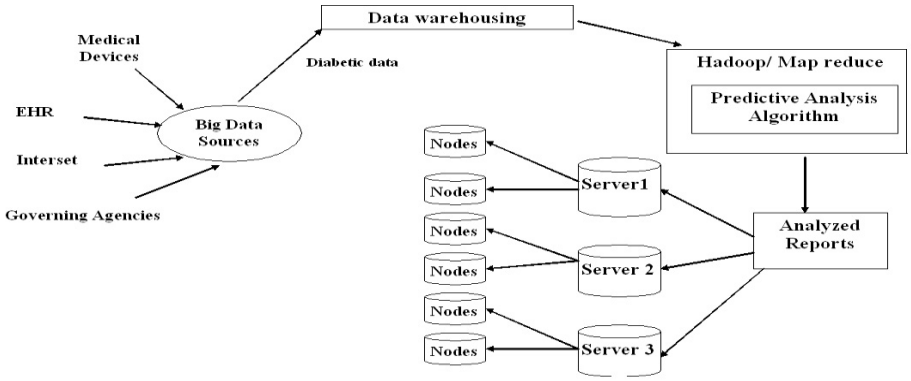

Figure 1. Architecture of the predictive analysis system-Health Care Applications

3.3 Predictive Analysis

Predictive analysis can help healthcare providers accurately expect and respond to the patient needs. It provides the ability to make financial and clinical decisions based on predictions made by the system. This system uses the predictive analysis algorithm in Hadoop/Map Reduce environment to predict and classify the type of DM, complications associated with it and the type of treatment to be provided.

Hadoop:

Hadoop is the open-source distributed data processing platform from Apache. Hadoop can serve the twin roles of data organizer and analytics tool [8]. Hadoop has the potential to process extremely large amounts of health data mainly by allocating partitioned data sets to numerous servers like clusters, each of which solves different parts of the larger problem and then integrates them for the final result. Hadoop uses two main components to do its job: Map/Reduce and Hadoop Distributed File System.

- Map/Reduce: Hadoop's implementation of Map/Reduce is based on programming models to process large data or datasets by dividing them into small blocks of tasks. Map/Reduce uses distributed algorithms, on a group of computers in a cluster, to process large datasets. It consists of two functions:
  - The Map ( ) function which resides on the master node and then divides the input data or task into smaller subtasks, which it then distributes to worker nodes that process the smaller tasks and pass the answers back to the master node. The subtasks are run in parallel on multiple computers.
  - The Reduce ( ) function collects the results of all the subtasks and combines them to produce an aggregated final result — which it returns as the answer to the original big query.
- Hadoop Distributed File System (HDFS): HDFS replicates the data blocks that reside on other computers in the data center (to ensure reliability) and manages the transfer of data to the various parts of the distributed system.

Pattern discovery:

For diabetic treatment it is necessary to test the patterns like, plasma glucose concentration, serum insulin, diastolic blood pressure, diabetes pedigree, Body Mass Index (BMI), age, number of times pregnant.

The pattern discovery of predictive analysis must include the following [14]:

- Association rule mining- Association between diabetic type and pages viewed (e.g. laboratory results)
- Clustering- clustering of similar patterns of usage, etc.
- Classification- Classification of health risk value by the level of patient health condition.
- Usage of statistics
- Application of pre-defined deductive rules across data

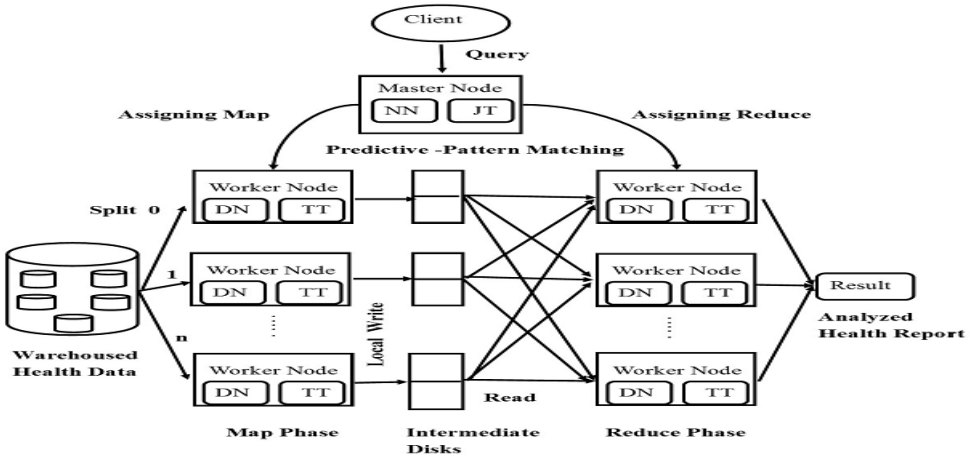

Figure2. Predictive – Pattern Matching System

Predictive Pattern Matching:

Whenever the warehoused dataset was sent to Hadoop system, immediately the map reduce task is performed. In mapping phase, the Master Node splits large data into smaller tasks for numerous Worker Nodes. Figure 2. deploys the exact operation of predictive pattern matching system. The Master node is one consists of Name Node (NN) and Job Tracker (JT), which always employs the map and reduce task. The Worker Node or Slave Node receives the order from the Master Node, process the pattern matching task for diabetes data with the help of Data Node – Same Machine (DN) and Task Tracker (TT). The predictive matching is the process of comparing the analyzed threshold value with the obtained value. If the pattern matching process was completed by all Worker Nodes based on the requirement, it was stored in intermediate disks. This process is known as local write. If the reduce task was initiated by Master Node, all other allocated Worker Nodes will read the processed data from intermediate disks. Based on the query received from Client through Master Node, the reduce task will be performed in Worker Node. The results obtained from the reduce phase will be distributed in various servers.

3.4 Processing analyzed reports

After the analysis of large diabetic data went through Hadoop, the final results are distributed over various server and replicated through several nodes depending on the geographical area. By employing proper electronic communication technology to exchange the information of individual patients among health care centers will leads to get proper treatment at right time in remote locations at low cost.

Benefits of this predictive analysis system

The diabetes may associate with severe diseases such as heart attacks, strokes, eye diseases and kidney diseases, etc. Analyzing the risk value by the level of patient health condition using above results of can be used by the physicians at remote locations to serve the people. Detecting diseases at earlier stages can help to be treated more easily and effectively. In developing countries such as India, it is mandatory to manage specific individual and population health and detecting health care fraud more quickly. The middle-income families can be with the high availability of medical facility at minimum cost. This system leads to the improved focus on every individual patient health. Thereby we can reduce and save our next generation from diabetic mellitus.

Analysis and proofs

This system becomes master in health care management system and drives extreme growth. This system tends to be data centric for most of the multidimensional global health cares. It is the platform for intelligence and knowledge prediction in real time handling of large volume of data.

IV. CONCLUSION

Big Data Analytics 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. By transforming various health records of diabetic patients to useful analyzed result, this analysis will make the patient understand the complications to occur. The goal of this research deals with the study of diabetic treatment in healthcare industry using big data analytics. The design of predictive analysis system of diabetic treatment may give enhanced data and analytics yield the greatest results in healthcare. By employing location aware healthcare service, anyone from rural area can get proper treatment at low cost. This research mainly focused for the patients in the rural area. Treatment can be offered when it is identified in advance.

REFERENCES

1. V. H. Bhat, P. G. Rao, and P. D. Shenoy, "An Efficient Prediction Model for Diabetic Database Using Soft Computing Techniques," Architecture, Springer-Verlag Berlin Heidelberg, pp. 328-335, 2009..
2. Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, "Application of data mining: Diabetes health care in young and old patients", Journal of King Saud University – Computer and Information Sciences, vol. 25, pp. 127–136, 2012.
3. K. Rajesh, V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis" in International Journal of Engineering and Innovative Technology (IJEIT) Vol 2(3), 2012.
4. Sadhana, Savitha Shetty, "Analysis of Diabetic Data Set Using Hive and R", International Journal of Emerging Technology and Advanced Engineering, vol 4(7), 2014.
5. 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.
6. V. H. Bhat, P. G. Rao, S. Krishna, and P. D. Shenoy, "An Efficient Framework for Prediction in Healthcare," Most, Springer-Verlag Berlin Heidelberg , pp. 522-532, 2011.
7. Nishchol Mishra, Dr.Sanjay Silakari, "Predictive Analytics: A Survey, Trends, Applications, Oppurtunities & Challenges", International Journal of Computer Science and Information Technologies, vol. 3(3), 4434- 4438 4434, 2012.
8. Wullianallur Raghupathi, and Viju Raghupathi, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, vol. 2(3) pp. 2-10, 2014.
9. Mansoor Khan," Managing Vulnerable Populations: How Predictive Analytics is a Key Component for Understanding User Behavior and Improving Care Quality", Managed Care Outlook, vol 27(4), February 15, 2014.
10. Andrew Pearson, Qualex Asia, "Predictive Analytics for the Healthcare Industry", Andrew Pearson, Qualex Asia Limited, 2012.
11. http://www.intel.com/content/www/us/en/healthcare-it/bigger-data-better-healthcare-idc-insights-white-paper.html
12. D. Peter Augustine, "Leveraging Big Data analytics and Hadoop in Developing India's Health Care Services", International Journal of Computer Applications, vol 89(16), pp 44-50, 2014.
13. Muni kumar N, Manjula R,"Role of Big Data Analytics in Rural Health Care – A Step Towards Svasth Bharath", International Journal of Computer Science and Information Technologies, vol 5(6), pp 7172-7178, 2014.
14. Andre W. Kushniruk, "Predictive Analytics and Forecasting in Health Care: Integrating Analytics with Electronic Health Records", SAS Institute Inc, 2008.