IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

Predictive Analysis to Prognosis of Mellitus Caused by the Diabetics

K. Arun[1], M. Ananda Kumar[2]

Student Member, Department of Computer Science and Engineering, Arasu Engineering College, Kumbakonam, Tamil Nadu, India[1]

Assistant Professor, Department of Computer Science and Engineering, Arasu Engineering College, Kumbakonam, Tamil Nadu, India[2]

Abstract: In the modern digitalized world, the health care industries are moving towards the processing of those health records to analyze it for getting the immediate remedy for the complexities due to the mellitus. The growing of unstructured data has the nature if Big Data. So, we have to emphasis its size in nominal value with possible solution done by converting it as structured data. It is necessary because the health care industries have to face the difficulties over analyzing the Health data's. The Diabetic Ailment (DA) was the one of the NCD, which is the major health hazard in the Developing countries such as India. The DA was associated with huge number of long term complications and health disorder. In this paper, the use of predictive analysis algorithm based on the ANN in MapReduce of Hadoop Environment have been able to predict the disorders associated with the DA and what type of treatment has been provided. By the analysis, this system provides an efficient way to take care of patients who has affected by the diabetics and help them to cure from it.

Keywords: Hadoop/MapReduce, Predictive analysis (Artificial Neural Networks), Big Data, Health care Industries, Diabetic Analysis (Non-Communicable Diseases);

I. INTRODUCTION

Diabetes Ailment (DA) is a metabolic disorder characterized by chronic hyper glycaemia with disturbances of carbohydrate, fat and protein metabolism. There are three main types of diabetes ailment (DA). Type 1 DA results from the body's failure to produce insulin, and presently requires the person to inject insulin or wear an insulin pump. This form was previously referred to as "insulin-dependent diabetes Ailment" (IDDA) or "juvenile diabetes". Type 2 DA 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 ailment (NIDDA) or "adult-onset diabetes". 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. As of 2000 it was estimated that 171 million people globally suffered from diabetes or 2.8% of the population. Type-2 diabetes is the most common type worldwide. Figures for the year 2007 show that the 5 countries with the largest amount of people diagnosed with diabetes were India (40.9 million), China (38.9 million), US (19.2 million), Russia (9.6 million), and Germany (7.4 million)[1]. 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 data are 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 [6]. 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.

II. RELATED WORKS

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

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

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 [11]. The classical neural network model is used for prediction, on the pre-processed dataset.

K. Rajesh, V. Sangeetha[4] have applied data mining techniques to classify Diabetes Clinical data and predict the likelihood of a patient being affected with Diabetes or not. The training dataset used for data mining classification was the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases from UCI Machine Learning Repository. The dataset contains 768 record samples, each having 8 attributes. We used this dataset for our classification exercise using C4.5, as the data is complete with no missing values.

A Genetic Algorithm (GA) based model was developed by Sabibullah [8] in patients particularly diabetic for accessing and forecasting the prone risk of heart attack and stroke. To screen the people who are suffering from these diseases, the application model would provide a possibility of risk behind the heart attack or other neurodeficit diseases. An efficient soft computing based algorithm would enable a probable model in the classification process so as to reduce the mortality rate in the population data. It is a valuable option so as to share the knowledge and experiences towards the model screening of diabetic patients by the proposed technique.

An artificial neural network (ANN), often just called a "Neural network" (NN), is a mathematical model or computational model based on biological neural network. Neural networks process information in a similar way the human brain does. The network is composed of a large number of highly interconnected processing elements (neurons) working in parallel to solve a specific problem. In medicine, ANNs have been used to analyze blood and urine samples, track glucose levels in diabetics, determine ion levels in body fluids and detect pathological conditions. Artificial Neural networks are well suited to tackle problems that people are good at solving, like prediction and pattern recognition. Neural networks have been applied within the medical domain for clinical diagnosis, image analysis and interpretation, signal analysis and interpretation and drug development [12].

Various big data technology stack and research over health care combined with efficiency. Cost savings, etc., are explained in better healthcare [2]. 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 [3].

All the above researchers have been successful in analyzing 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 the ailments caused by diabetics. This system provides efficient way to care and cure the patients at low cost.

III. PURPOSED SYSTEM

The architecture of Purposed system includes various phases like Query Processing, Map Reduce, and predictive analysis. Figure 1 shows the complete architecture of proposed method.

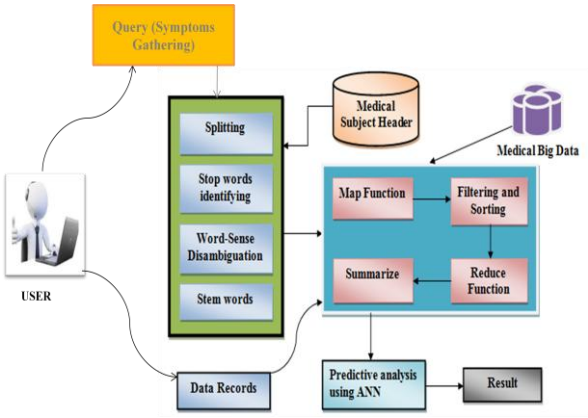

Figure 1: Purposed System Architecture

A. QUERY PROCESSING

The actual symptoms of the client and their details are gathered and processed in this phase. For the more effective process the symptoms which are given by clients are analyzed by as follows

1) GATHERING USER INFORMATION

On this part, the input query has been gathered from the user based on their symptom first then the details of the user attributes were as Sex, Diastolic B.P, Plasma glucose, Skin fold thick, BMI, Diabetes Pedigree type, No. of times Pregnant, 2 hr Serum Insulin.

2) WORD SPLITTING

The symptom gathered as the user query has been in the form of sentence/paragraph. This has to be splitted into separated words for identification process.

3) WORD DISAMBIGUATION

Word disambiguation is the process of removing the similar words from the input of the user. This process is done before the process of identification of keywords for efficient analysis of medical data.

4) IDENTIFYING THE KEYWORDS

In this phase the stop words/stem words are collected from the splitted query of the user and the conjunctions are removed from the query. The stop words are matched with the MeSH Medical data dictionary to know about the related terms with the actual symptoms based on the input query.

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

5) DATA RECORD
The information about the user knowned details, are gathered from this part. Commonly the details like Sex, Diastolic B.P, Plasma glucose, Skin fold thick, BMI, Diabetes Pedigree type, No. of times Pregnant, 2 hr Serum Insulin are gathered and for processing.

B. MAP-REDUCE PROCESS
As we already know that the medical big data has to be pre-processed to analysis it. The Map-Reduce functions are used to pre-process the medical data. Here the diabetics related data's are separated from the general medical data sets. The further detailed processes are discussed below:

i) MAP FUNCTION
The Basic process of mapping function is to do the Filtering and Sorting process over the medical data sets. The Algorithm filtering the data set has done with the help of the user keywords and the data's provided by the user input. The algorithm for the filtering process is shown in the table 1.

TABLE 1: ALGORITHM FOR FILTERING

Algorithm: FILTERING
Input: Medical Big Data and user keyword
Output: Diabetics Data
Steps
1: CC ←connected components in {e ∈ E | re ≤ t}
2: let h : [n] → [nδ4] be a universal hash function
3: map each edge (u, v) to machine h(u) and h(v)
4: map the assignment of node u to its connected component
CC(u), to machine h(u)
5: on each reducer rename all instances of u to CC(u)
6: map each edge (u, v) to machine h(u) + h(v)

ii) REDUCE FUNCTION
The reduce function is to performs a summary operation. The output produced by map function is taken as input and the summarized data has been generated as the output.

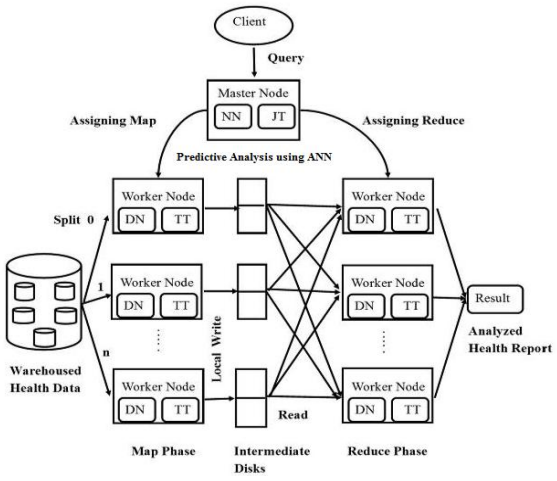

Figure 2: Map-Reduce flow

Here the filtered data's are summarized. The conceptual figure of Map-Reduce process is shown in the Figure2.

C) 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 DA, complications associated with it and the type of treatment to be provided. The predictive analysis can be done using the one of the pattern matching technique (Machine Learning) called as Artificial Neural Networks (ANN). The ANN follows feed-forward or back-propagation algorithmic method to compute their input and produce the output. The Back-Propagation is more efficient than the feed-forward algorithm. It has better reliability as well as the consistency of the algorithm proved it.

BACK-PROPAGATION ALGORITHM
Back propagation, an abbreviation for "backward propagation of errors", is a common method of training artificial neural networks used in conjunction with an optimization method such as gradient descent. The method calculates the gradient of a loss function with respect to all the weights in the network, so that the gradient is fed to the optimization method which in turn uses it to update the weights, in an attempt to minimize the loss function.

Back propagation requires a known, desired output for each input value in order to calculate the loss function gradient. It is therefore usually considered to be a supervised learning method, although it is also used in some unsupervised networks such as auto encoders. It is a generalization of the delta rule to multi-layered feed forward networks, made possible by using the chain rule to iteratively compute gradients for each layer. Back propagation requires that the activation function used by the artificial neurons (or "nodes") be differentiable. The algorithm for back propagation is as shown in the table2.

TABLE 2: BACK PROPAGATION ALGORITHM

The back propagation learning algorithm can be divided into two phases
Phase 1: Propagation
Each propagation involves the following steps:
1. Forward propagation of a training pattern's input through the neural network in order to generate the propagation's output activations.
2. Backward propagation of the propagation's output activations through the neural network using the training pattern target in order to generate the deltas (the difference between the targeted and actual output values) of all output and hidden neurons.

DOI 10.17148/IJARCCE.2016.51138

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

Phase 2: Weight update

For each weight-synapse follow the following steps:

1. Multiply its output delta and input activation to get the gradient of the weight.
2. Subtract a ratio (percentage) from the gradient of the weight.

IV. RESULT ANALYSIS

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 healthcares. It is the platform for intelligence and knowledge prediction in real time handling of large volume of data. The analysis over the different data sets and their false positive ratio has been showed in the figure3.

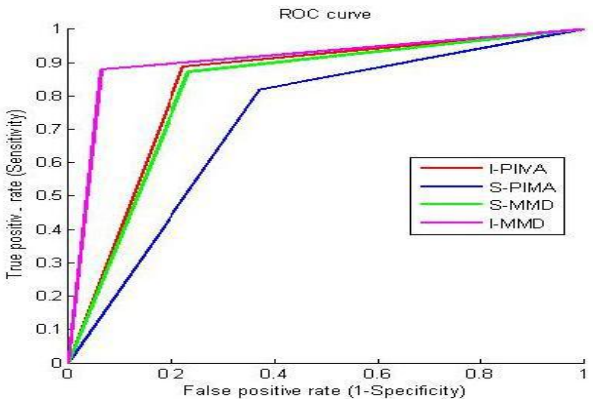

Figure 2: False-Positive and True-Negative data's over different Datasets

V. CONCLUSION

This work proposes a new approach for preprocessing real-time medical big data, used for predictive analyses of diabetes. It can able to predict the disorders that are related with DA. In future this framework can able to work with Multiple Ailment like Cancer, heart diseases, and other major Non-Communicable diseases. The experimental results show that it will be more efficient when works with the real-time data.

REFERENCES

[1] 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

[2] Andre. Kushniruk, "Predictive Analytics and Forecasting in Health Care: Integrating Analytics with Electronic Health Records", SAS Institute Inc, 2008.

[3] 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.

[4] 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.

[5] Muni kumar, Manjula,"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.

[6] Nishchol Mishra, Dr.Sanjay Silakari, "Predictive Analytics: A Survey, Trends, Applications, & OppurtunitiesChallenges", International Journal of Computer Science and Information Technologies, vol. 3(3), 4434- 4438 4434, 2012.

[7] P. Padmaja, "Characteristic evaluation of diabetes data using clustering techniques", IJCSNS International Journal of Computer Science and Network Security,VOL.8 No.11, November 2008.

[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.

[9] Sadhana, Savitha Shetty, "Analysis of Diabetic Data Set Using Hive and R", International Journal of Emerging Technology and Advanced Engineering, vol 4(7), 2014.

[10] Shantakumar B.Patil, Kumaraswamy Y S,"Intelligent and effective heart attack prediction system using determining and artificial neural netwok", European Journal of Scientific Research, Vol. 31 No.4, 642-656, 2009.

[11] 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.

[12] 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.