Prognosis Of Ailments Caused By Diabetic Mellitus Using Predictive Analysis

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

[1]*Student Member, Department of Computer Science and Engineering,*

[2]*Assistant Professor, Department of Computer Science and Engineering,*

Arasu Engineering College, Kumbakonam, Tamil Nadu, India

ABSTRACT-

In the digitalized world, the health care industries produce huge numbers of health records in terms of EHR, which is generally in the unstructured format. The growing of unstructured data has the nature of Big Data. So, we have to emphasis its size in nominal value with possible solution done by converting it as structured data. The Diabetic Mellitus (DM) was the one of the NCD, which cause major health problems in Developing countries such as India. The DM was associated with huge number of long term complications and health disorder. This paper uses the Infrastructure as a Service to provide Software as a Service to its users (doctors and patients). This paper envision that predictive analysis based on back propagation algorithm of artificial neural network applied over the data summarized by the hadoop MapReduce through the web applications will be a better and efficient way to simply the risk and complexities in health care data processing to predict the disorder and their treatment associated with the DM. This web application can be used by the both doctor and patients. To provide the security for the data being used by the web application is encrypted with the help of hash functions. 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), Cloud Service;

I. INTRODUCTION

Big data in healthcare is basically a collection of large amount of data generated by patient generated records, patient reports, machine generated data, hospital patient database, prescriptions and methods, etc. Previously, these sets of data were manually collected and since last decade electronic report generation has been adapted which has resulted into huge number of data centers for data handling and storage. With wearable technologies, a large amount of data is generated in real time. This data is scattered across globe and is in different formats. As a result, analysis of this collected differential data with intelligent design and state of the art algorithms is the key to improved healthcare services. Distributed computing has helped solve the problems with data located at different locations. Various tools available in the market have enabled data cleansing and transformation by eliminating unwanted attributes. Whilst many dimensions of big data have issues pertaining to the 5V's of big data namely volume, variety, velocity, veracity, and value, the accuracy, integrity, and semantic interpretation are of utmost importance in any clinical/medico application. These challenges have not deterred but have instead escalated the use and exploration of big data as a proven source of betterment for healthcare. [1]

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

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

In health care services, the consistence is non-debatable. Nothing is more vital than the protection and security of patient information. In any case, in all honesty, there aren't some great, coordinated approaches to oversee security in huge information. In spite of the fact that security is tagging along, it has been a bit of hindsight so far [3].

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.

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. PROPOSED WORK

The Proposed System has the following divisions to perform the analysis over the medical big data for predicting the treatments for the ailments caused by the diabetics.

1. Query Processing
2. Map Reduce
3. Predictive Analysis

1. 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

A) 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.

B) 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.

C) 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.

D) 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.

E) 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. SECURITY IMPLEMENTATION

The details of the user details are more precious and it can be more useful at the same time it will become very dangerous if any intrusions/ modifications on the big data. We have to prevent the modifications on those datasets in the SAAS platform by the intruders, we to encrypt the medical data set. The input of the map-reduce process has get the hashed data set to process it. The Secured Hash Algorithm (SHA-3) has been used to encrypt the data set as showed in Table 1.

TABLE 1: SHA-3 ALGORITHM

ALGORITHM: SHA-3

INPUT: MEDICAL DATA SET
OUTPUT: HASHED MEDICAL DATA SET

```
SHA3[b](A)
{
 for all i in 0…nr-1
   A = Round[b](A, RC[i])
 return A
}

Round[b](A,RC)
{
 θ step
 C[x] = A[x,0] xor… A[x,4] for all x in 0…4  D[x] = C[x-1] xor rot(C[x+1],1), for all x in 0…4 A[x, y] = A[x, y] xor D[x],  for all (x, y) in (0…4,0…4)

 ρ and π steps
 B[y,2*x+3*y] = rot(A[x, y], r[x, y]),  for all (x, y) in (0…4,0…4)
 χ step
 A[x, y] = B[x, y] xor ((not B[x+1,y]) and B[x+2,y]), for all (x, y) in (0…4,0…4)ι step
 A[0,0] = A[0,0] xor RC
 return A
}
```

C. 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 filtering process is shown on the Table 2.

TABLE 2: ALGORITHM 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. The Map-Reduce process has shown in the figure1.

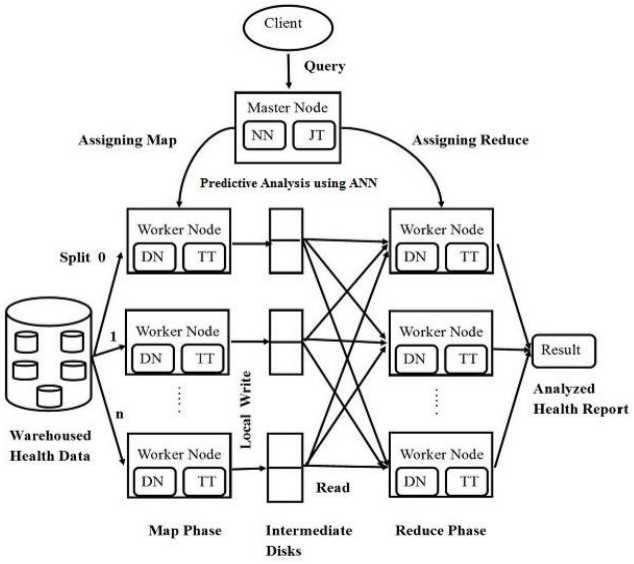

Figure 1: Map-Reduce Flow

D. 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 table3.

TABLE 3: 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.

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 DM to predict the treatment has shown in the figure2.

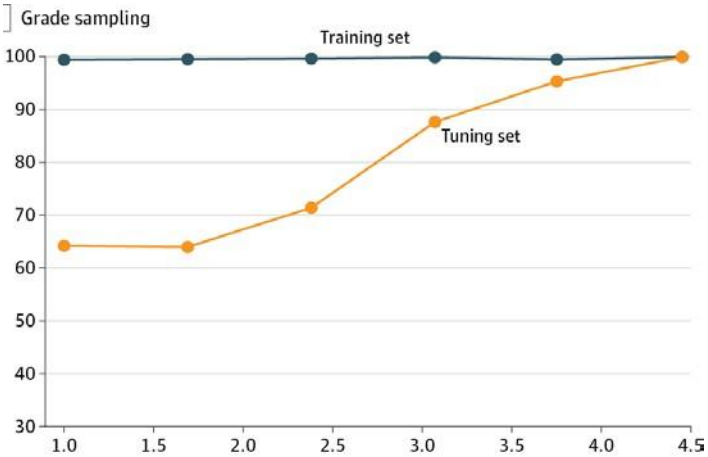

Figure 2: Predicted Treatment for the DM

V. CONCLUSION

This work proposes a new approach for pre-processing real-time medical big data, used for predictive analyses of diabetes. It can able to predict the disorders that are related with DM. 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, (2012), "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.

[2] Bhat V.H, Rao P.G, and Shenoy P.D (2009), "An Efficient Prediction Model for Diabetic Database Using Soft Computing Techniques," Architecture, Springer-Verlag Berlin Heidelberg, vol.2(3) pp. 328-33.

[3] Muni kumar, Manjula (2014), "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.

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

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

[6] Peter Augustine. D (2014), "Leveraging Big Data analytics and Hadoop in Developing India's Health Care Services", International Journal of Computer Applications, vol 89(16), pp 44-50.

[7] Rajesh. K, Sangeetha. V (2012), "Application of Data Mining Methods and Techniques for Diabetes Diagnosis" in International Journal of Engineering and Innovative Technology (IJEIT) Vol 2(3).

[8] Sabibullah. M, Shanmugasundaram. V, Raja Priya K. (2013), "Diabetes Patient's Risk through Soft Computing Model", International Journal of Emerging Trends & Technology in Computer Science, vol 2(6).

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

[10] Shantakumar B. Patil, Kumaraswamy Y.S. (2009),"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] Wullianallur Raghupathi, and Viju Raghupathi (2014), "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, vol. 2(3) pp. 2-10.