Intelligent System for Disease prediction using Deep Learning Techniques

Sam V George
PG Scholar
Department of Computer Science and Engineering
Mangalam College of Engineering, Ettumanoor, Kottayam

Abstract: The origin of BigData technologies can enhances the value of various industrial sectors including HealthCare. The large data generated is needed to be managed, stored and analyzed for developing forecast and recommendation intelligent systems. Diabetes is one of the most common and dangerous life style disease .The diabetes diagnosis is a lengthy process and the diagnosis play a major role in controlling this condition by altering the life style by the victims. In this paper an intelligent system is proposed for prediction of Diabetes using supervised learning methodology with Deep learning techniques. The model shows promising results so that further Deep learning models can be used for future researches.

***Keywords*: Big Data, Diabetes, Intelligent Systems, Deep learning.**

1. Introduction

Over the last few years the healthcare has generated enormous amount of data. It includes clinical reports, doctor's prescription, diagnostic reports, medical images, pharmacy information, health insurance related data, data from social media and other sources too. All these information collectively forms the Big Data This data can be used for analysis which will helps for forming prediction and recommendation systems which helps to improve the healthcare, with proper treatment at the early stages of disease. The Big Data is associated with characteristics like Volume, Velocity, and Variety. The collection and storage of health related data over the time its self affects the Volume of Data. Accessing this data in the real time causes the Velocity. Various features are subjected for the analysis in the healthcare domain which is said by the variety. The healthcare domain is currently moving to a data driven domain. Big Data and its Technology holds tremendous role in changing the Healthcare domain as we known to a much more Intelligent and sophisticated.

Diabetics also called as diabetes mellitus, which is a common condition in which the victims have high blood sugar content over a long period of time that is basically caused by a group of metabolic activities. The symptoms of diabetics are high blood sugar, increased thirst, frequent urination and weight loss. The long time persistence of these symptoms leads to failures in heart, eyes, kidneys and nerves systems. As per the studies conducted by WHO 1 in 3 adults have diabetics. And 1.6 million people die every year due to diabetics [1]. As per the survey done by International Diabetes Federation, in 2015, it was estimated that 415 million individuals are affected by diabetes around the world. And the number is estimated to rise to 642 million individuals by the year 2040.The diabetics are

mainly of two types they are called as Type I and Type II diabetics. The type I diabetics is caused due to the failure of pancreas to no longer produce adequate insulin for the functioning of body. While the Type II diabetics occur due to the increased insulin resistance of the cells. People with insulin resistance often have a group of conditions including high blood glucose, extra fat around the waist, high blood pressure, and high cholesterol. For some extend the Type II diabetics occurs due to lifestyles like getting no exercise, Stress, Smoking and Irregular sleeping patterns.

Diabetics is diagnosed by conducting Oral glucose tolerance test (OGTT) The OGTT measures blood glucose after you fast for at least 8 hours. This is a highly time-consuming process and there are many other features which need to be considered while attempting to detect whether a patient is diabetic or not. These other factors are: insulin, body mass index, blood pressure and age. The hereditary also influence the chances of being diabetics. Presently there is no non- invasive methods to detect the person is having an affiliation to Type I diabetics. Hence there is the origin of the need of an intelligent system to predict the person is diabetics or not which helps for easy and faster diagnosis which in turn helps to start the treatment at the earliest. The extensiveness of diabetics changes from tribes, villages, continents. Deep learning involves working on a model that is very similar to that of the human brain. It has the capability to decode complex problems much like that of the human brain. Deep neural networks can handle massive sets of data and complex tasks. The basic concept behind deep learning is its ability of learning from the data representations rather than a task specific algorithm. It can perform well with the supervised, unsupervised or semi supervised learning strategies. This ability of Deep learning is extremely useful for faster processing and handling of higher dimensions of data better than that of traditional machine learning algorithms.

2. Literature Survey

J.Sivaranjani, A.NeelaMadheswari proposed a system for theme disclosure in extensive scale time arrangement utilizing Hadoop condition. Continuous information of ECG from UCR dataset is considered for assessment [2]. Theodoros Mavroeidakosy, Nikolaos Tsolis and Dimitrios D. Vergados all together propose Centralized Management of Medical Big Data in Intensive Care Units. The QoS in ICU will be drastically improve since the analysis of the data in ICU assist n decision making in the critical situations. It also helps to supervise the treatments which will identify the errors in diagnosis and treatments [3]. Rezvan Pakdel, John Herbert put forward a cloud based data analysis method in the health care area which is very scalable. The data stored in the cloud is used for various purposes. On analyzing larger data the framework shows scable and effective solutions [4]. For mining of massive document-data of medicine Li Wei,Liu Guangming, Shao yachao,Liu junlong,Zuo you proposed a map reduce based modified Apriori algorithm [5].This solution is applicable for mining associative rules in between the features and reduce the scaning overhead. Ayush anand, Divya shakti discussion on establishing a relationship between diabetes risk likely to be developed from a person's daily lifestyle activities such as his/her eating habits, sleeping habits, physical activity along with other indicators like BMI [6](Body Mass Index), waist circumference. Min Chen, Yixue Hao, Kai Hwang,Lu Wang, and Lin Wang a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm[7] using structured and unstructured data from hospital. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. For simple diseases the intelligent diagnosis works perfectly. Both structured and unstructured data is considered for

analysis. To solve the difficulty of incomplete data, we use a latent factor model to reconstruct the missing data. We experiment on a regional chronic disease of cerebral infarction.

3. Data Collection and Pre-Processing

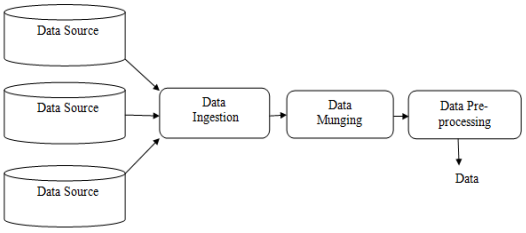

Fig 3.1. Data Engineering process

The raw data is collected over various sources like hospitals, pharmaceuticals and other sources. These data is imported for immediate use. The operation involving of importing necessary data from the data source is called as data ingestion. The raw data is subjected to various transformations and filtering, which enables the collected raw data transform to another format which will be useful and meaningful in different contest.

After this Munging operation the data is subjected to preprocessing, which will perform the steps like data cleaning where the data is filled where they are incomplete, identify the outliers and smooth out noisy data and correct the inconsistent data.

Table 3.1. Features of the dataset

Feature	Description
Pregnancies	Number of times pregnant
Glucose	Result of oral glucose tolerance test
Blood pressure	Diastolic blood pressure
Skin Thickness	Triceps skin fold thickness
Insulin	2 hour serum insulin
BMI	Body mass index
DPF	Diabetes Pedigree Function
Age	Age(years)
Outcome	Class variable(0 or 1)

The data transformations include data normalization where the scaling of attribute fall in a suitable range and data generalization and attribute construction with help of a domain expert. Then preprocessing have another operation Data reduction is another step involved in the data preprocessing which helps to reduce the number of attributes considered for analysis. It also helps to sample the data so that reducing the computational complexity and space complexity during the execution time. After these steps the real world data with the nature of incompletion which is the lacking of attribute values or the lacking of attributes of interests, Noisy and inconsistent are transformed to useful filtered data for the analysis. The data dealing on the system architecture is pre-processed before passing for analysis.

The dataset used for analysis is a simulated version of Pima Indian Dataset. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. It consists of 768 rows of data with 8 features and one outcome. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. There are eight features and one label for identifying the result in the dataset.

4. System Architecture

The Architecture consists of the following:

Data: The raw data collected is preprocessed and stored in the system for the analysis. It can be stored in databases like Cassandra, MongoDB.

Train Data: 67% of the data is split[9] apart from the Pima Indian Dataset and it is kept it as the training data. This data is used to feed into the deep learning algorithm, which accepts this data as an input to train the model for predicting the diabetes.

Test Data: The remaining 33% of the Pima Indian dataset is kept as the test dataset. After training the model with a data, model is tested whether it is working properly by feeding this as the input and trying to predict the result based on the model trained by the training data. This data is passed as a batch for the testing.

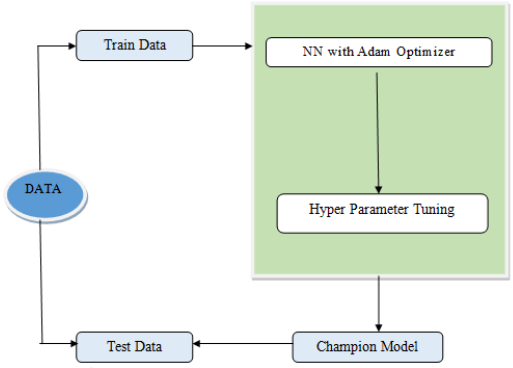

Fig 3.2. System Architecture

ANN Model: The Artificial neural network model is used to simulate the activity of human brain with the help of artificial neurons as the processing element. For the implementation of this model Tensorflow libraries are used. The Tensorflow library is developed by the Google and later it is made open source for the public. It is one of the most famed machine learning libraries. It uses the computational graph concept, automatic differentiation which helps the users to solve real life problems faster and easier. The Tensorflow utilize not only the CPU but also GPUs for the processing of computational graphs. It convert a problem into a computational graph and feed in data through the placeholders, calculate the output of the computational graph, compare the output to the desired output with a loss function, modify the model variables according to the automatic back propagation, and finally repeat the process until a stopping criteria is met.

Hyper Parameter tuning: The created model is subjected to tuning. The model is subjected to change some parameters like step size, batch size, epoch and these parameters varies the quality of the model. The batch size is

the number of sub samples given to the network after which parameter update happens. How much data is feed to the Neural Network model as input is depending on this factor. Number of epochs is the number of times the whole training data is shown to the network while training. Increase the number of epochs[8] until the validation accuracy starts decreasing even when training accuracy is increasing. It means over fitting. Other parameter used for the tuning is the learning rate. The learning rate defines how quickly a network updates its parameters. Low learning rate slows down the learning process but converges smoothly. Larger learning rate speeds up the learning but may not converge. Step size is another factor used for the perfection of model improvement. The step size determines the ability of algorithm to reach optimal solution rather than faster executions.

Champion model: The resultant model after the improvement of performance is given to the champion model module. The champion model works better for the test data than that of the model with lower accuracy.

4. Result Analysis

The ANN model is trained and tested against the simulated Pima Indian Dataset. The dataset is simulated into 175,104 attributes. The 67% is taken as training dataset and the remaining 33% of the data is taken for testing the model. The accuracy of the system is analyzed using confusion matrix and ROC curves.

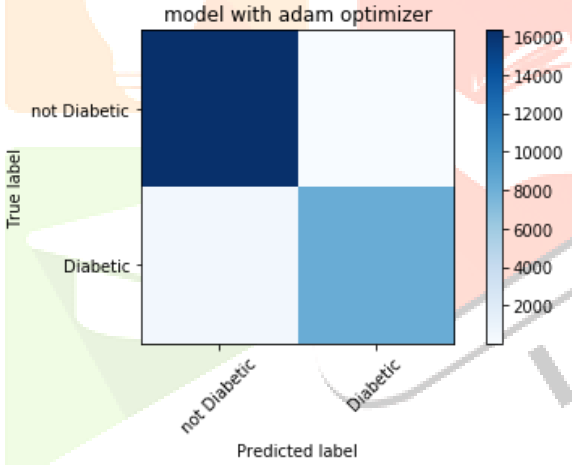

Fig 5.1. Confusion Matrix for the ANN model with Adam optimizer

The model above is illustrated through confusion matrix. The machine learning algorithms which deals with the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one are used. The problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one.

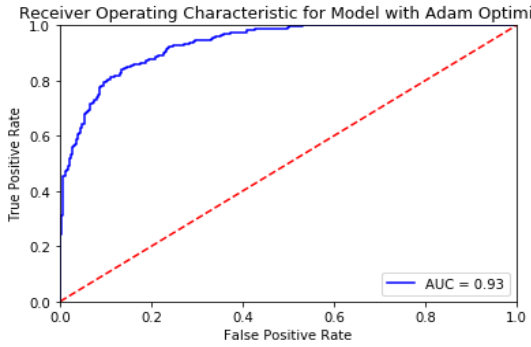

Fig 5.2. ROC curve for the ANN model with Adam optimizer

The true positive rate is plotted against the false positives in this visualization technique called ROC curve. The slope of the tangent line at a cut point gives the likelihood ratio (LR) for that value of the test. The AUC value which is called as area under the curve shows the efficiency of the model. The model shows an AUC of 93%. The diabetic people who are not diabetic are predicted in the range of 14000-16000 .Similarly the prediction for diabetic are in the range of 6000-10000.

6. Conclusion and Future Scope

The large scale of data in the Medicare field forces to adapt BigData technologies for the management and analysis. Introduction of this technologies can helps to improve the quality of life by using intelligent systems to predict diseases and reduce the overhead of doctors for the disease diagnosis. The analysis and prediction of the diabetes using Deep learning techniques shows promising results. This model can be implemented in the hospitals for easy determination of diabetes. Even though the model shows promising results more studies and research had to done in this area. The application of various Artificial Neural Network models and its effects to be determined on the collective data. Future studies can be directed to weight assignment based on the feature impact on outcomes.

7. References

[1] http://www.who.int/diabetes/en/

[2] J.Sivaranjani, A.Neela Madheswari, " A Novel Technique of Motif Discovery for Medical Big data using Hadoop", Proc. IEEE Conference on Emerging Devices and Smart Systems (ICEDSS 2017),3-4 March 2017, Mahendra Engineering College, Tamilnadu, India.

[3] Theodoros Mavroeidakosy, Nikolaos Tsolis and Dimitrios D. Vergados," Centralized Management of Medical Big Data in Intensive Care Unit: A Security Analysis" IEEE 978-1-5090-4476-4/16/$31.00 ©2016

[4] Rezvan Pakdel, John Herbert," Scalable Cloud-based Analysis Framework for Medical Big-data ", 40th Annual Computer Software and Applications Conference 2016 IEEE pg 647-652

[5] Li Wei,Liu Guangming, Shao yachao,Liu junlong,Zuo you " Optimization and Application in Medical Big Document-Data of Apriori Algorithm based on MapReduce" International Conference on Computer Communication and Informatics (ICCCI -2016), Jan. 07 – 09, 2016, Coimbatore, INDIA

[6] Ayush Anand,Divya Shakti Prediction Of Diabetes Based On Personal Lifestyle Indicators " 1st International Conference on Next Generation Computing Technologies (NGCT-2015) Dehradun, India, 4-5 September 2015

[7] MIN CHEN1, YIXUE HAO1, KAI HWANG, LU WANG1, AND LIN WANG," Disease Prediction by Machine Learning Over Big Data From Healthcare Communities" IEEE access VOLUME 5, 2017 pg 8870- 8879

[8] www.tensorflow.org

[9]www.sci-kitlearn.org