Predicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm

Sreekanth Rallapalli
Faculty of computing
Botho University
Gaborone, Botswana
rallapalli.sreekanth@bothouniversity.ac.bw

Suryakanthi T
Faculty of Business
University of Botswana
Gaborone, Botswana
suryakanthi.tangirala@mopipi.ub.bw

***Abstract*—Electronic Health Records (EHR) is growing at an exponential rate that is being stored in enterprise databases or cloud storages. These records have now grown to be called as Big Data. Most of these data are unstructured. The data can be efficiently processed on cloud for lowering the processing costs. Predictive analytics help the physicians, doctors to identify the patient admission to hospital at early stage. To perform predictive analytics various factors with demographic data, hospital parameters, patient past history and various indicators for a specific disease. But identifying the strong indicators for accurate prediction is a challenging task. From the factors being considered for predictive analytics various models and algorithms need to be studied. Classification algorithms like Naive Bayes, Linear Regression; generalized additive model, Random Forest, Logistic Regression, Hidden Markov Models has to be considered for developing a predictive models. In this paper we propose a predictive model using scalable Random forest classification algorithm which can accurately identify the classifier rate for risk of diabetes.**

Keywords—Algorithm, Big Data, Classification, Cloud, EHR, Predictive model, Random Forest.

I. INTRODUCTION

Identifying the patients with high risk of being admitted to the hospital in nearby future will help the physicians, doctors to take decisions accurately. Hospitals can provide better healthcare to the patients by providing the needed infrastructure at right time. This can only be possible by providing the health care providers with accurate data analysis and better predictions using the available Big data sets. But with the availability of huge unstructured EHR data sets and hundreds of patient attributes it is challenging to find how best the indicators help in predicting the accuracy for risk of hospitalization. New programming framework such as Apache Hadoop [1] which implements MapReduce [2] computational paradigm is good for data intensive applications. Machine learning tools such as Apache Mahout [3] works well for classification and clustering. Predictive modeling uses a mathematical model which can predict about the future. Input predictors are identified and a mathematical model is designed. The model provides the response which can be used for faster decision making. The demand for electricity in a state, trading strategies of the market, patient hospitalization, and patient readmission are the few examples where predictive modeling is used. We can use the supervised learning methods such as classification, regression on the data sets provided. This can derive a model for accurate prediction on various issues. We can iterate the process till we find the best model and then integrate the model into applications to find best predictions. The main goal of this paper is to build two predictive models Classification and Regression Tree (CART) and Random Forest for EHR data sets to assess the variables related to predict the risk of diabetes. To assess the risk of diabetes we utilize the scalable Random forest algorithm for Big Data EHR records. The rest of paper is organized as follows. Section II focus on Big Data EHR sources. Section III describes the analytic platform. In Section IV classification algorithms are discussed. Section V focus on building two predictive models for EHR data sets. In Section VI results are provided. Section VII concludes the paper.

978-1-5090-2576-3/16/$31.00 ©2016 IEEE

II. BIG DATA EHR SOURCES

Electronic Health Records are growing at an alarming rate with high volume, speed and variety of data being generated through various sources. These data has to be aggregated for predictive analytics [4]. EHR helps the patient to get better healthcare, doctors to get the timely information to treat patients. The performance of a medical system can be improved with EHR. In this section we study the various sources where EHR can be generated. Structured EHR data from various hospital management systems, unstructured EHR data from clinical notes, medical devices, and images from various patient lab records are all the sources of health data. Fig.1 represents the various sources of data in healthcare system which forms the Big data Electronic health records.

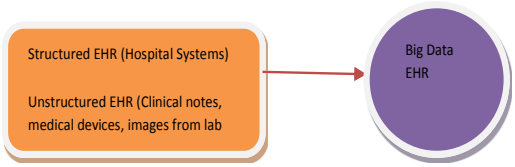

Fig 1: Big Data EHR sources

III. BIG DATA EHR ANALYTIC PLATFORM

In this section the platform for big data analytics is presented. Analytics is a process of inspecting, cleaning and transforming the data for predicting new information and which supports in efficient decision making process [5]. With EHR data doctors can able to take decision whether the patient needs an immediate specialized care or patient needs hospitalization in near future. The information has to be extracted using the extraction tools, selection of proper data need to be identified. Predictive models have to be designed in order to exactly predict the patient status [6,7]. The analytics platform for the healthcare data is shown in Fig.2.

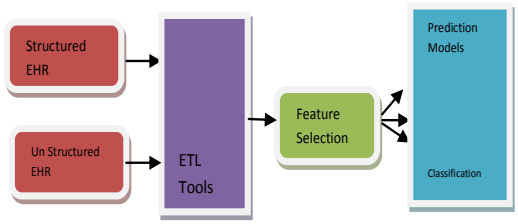

Fig 2: EHR Analytics platform

IV CLASSIFICATION ALGORITHMS

In this section we discuss the two most popular supervised learning classification algorithms. The first one Naïve bayes classification which are successful algorithm to classify the text documents. Here we use unstructured EHR clinical notes documents which are usually the text documents.

Naïve Bayes Algorithm: This algorithm is based on Bayes Theorem which is independent among the predictors. Naïve Bayes classifier assumes that in a class if any particular feature is available the presence of a particular feature in a class is unrelated to the presence of any other feature. This model is useful with large data sets. In our example we use large EHR data sets to classify the clinical notes of the physicians. The Bayes rule is given by

$$P(C|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

P(C|x) is the posterior probability of the class

P(c) is the prior probability of the class

P(x|c) is the likelihood which the probability of predictor given class

P(x) is the prior probability of predictor

But the limitation of the Naive Bayes algorithm is that it is impossible to get a set of predictors which are independent.

Linear Regression: This model is used to predict the clinical cost [11,12] and estimation of medical inspection. It can be expressed in the mathematical formulae as

$$y_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j \quad (2)$$

α is the intercept and β is the coefficient vector

The parameter estimation is viewed as the loss function minimization over the training data set in supervised learning.

The limitation of the linear regression model for data analysis is the output is not linear for all the inputs. The data is not sufficient to determine the coefficient of the model.

Generalized additive model: **T**his model is used when we need continuous outcomes in regression which is a combination of smooth functions [13]. The mathematical expression can be given as

$$y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \cdots \ldots \ldots + f_p(x_{ip}) \quad (3)$$

The limitation of this model is that when the smoothed variables have the values outside of range of trained data set it will lose the predictability.

Logistic Regression: For clinical prediction tasks logistic regression is one of the popular methods for binary classification [14,15]. The formulation of logistic regression for a two class scenario with N samples can be given as

$$\log \frac{P_r(y_i = 1|X_i)}{P_r(y_i = 0|X_i)} = \sum_{k=0}^{p} x_{ik}\,\beta_k = x_i\,\beta. \quad (4)$$

Multiclass Logistic Regression: in a multiclass logistic regression [16], the condition on one specific individual X_i, the probability that is observed output $y_i = j$ is

$$P_r(y_i = j|X_i) = \frac{\exp(X_i\beta_j)}{\sum_{k\neq j}\exp(X_i\beta_k)} \quad (5)$$

Random Forest algorithm: Large amount of data can be efficiently classified by using the Random Forest algorithm [8]. This algorithm is combined by tree predictors where each tree depends on the values of a random vector sampled independently. Weak learners in the group can come together to form a strong learner. Single decision trees often have high variance or high bias.

Classification trees are grown from Random Forest which is stated as below

a. Consider the data set with N sample cases at random – the original data is replaced which derives the training set for growing tree.

b. For a P input variable, by using the split node a number mP is specified. In each node, m variables are selected at random out of the P. During the forest growing the value of m remains constant.

c. With no pruning each tree is grown to largest extent possible.

The random forest model for the EHR data set is shown in the figure 3.

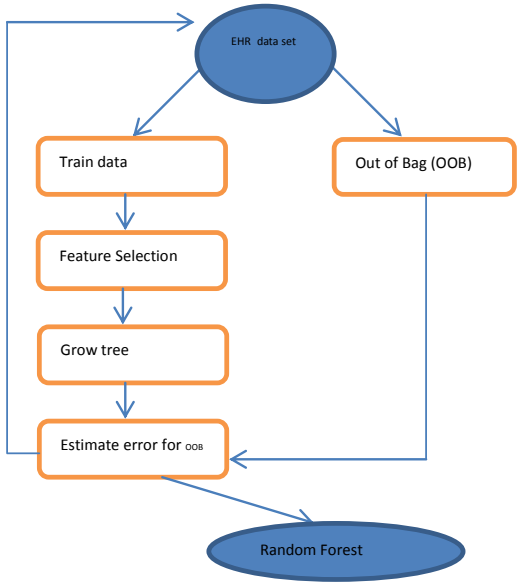

Fig 3: Random forest for EHR data set

As we deal with Big data Electronic health records with massive datasets the traditional algorithms [9] will not fit for these datasets. We need to build a distributed MapReduce based algorithm to classify such data sets. The Scalable Random Forest algorithm based on MapReduce is proposed in [9]. In this section we use this algorithm to test the Area Under Curve (AUC) of the diabetes dataset.

The massive data set of patients with various variables is loaded to the distributed environment. We first are being to configure the parameters of the jobs. After this the job is submitted to hadoop system. The data set with patient records is converted into vectors and stores them in list. As we know that in MapReduce programming [10] the map function is used to map the data set values with similarity variables, and reduce function will reduce the variables. Map function here will generate the decision tree under each of execution. The reduce function here will collect the information from the map function decision tree and then produce the random forest object.

V. PROPOSED PREDICTIVE MODEL FOR EHR DATA SET

In the medical terminology there are too many applications for decision tree models which can result in diagnosis and identification of the treatment protocols. The EHR dataset which we will be considering in this paper contains the information related to patients who have been diagnosed with diabetes. The main goal of this section is to build two predictive models Classification and Regression Tree (CART) and random forest to assess the variables relating to prediction of risk in diabetes which may further lead to chronic diabetes.

The variables considered for the dataset is as shown in Table1. The dataset contains 1500 observations and contains various variables such as Plasma level, Blood pressure, Insulin, BMI, Age and previous family history of diabetes.

TABLE 1: DATA SET FOR DIABETES

Plasma Level	Blood Pressure	Insulin	BMI	Age	Family History
150	160	20	34.6	50	1
85	90	25	27.5	40	0
190	185	29	24.5	32	1
90	90	0	25.5	21	1
140	140	160	28.1	33	0
120	120	88	45.3	30	1
80	80	0	35.5	26	1
120	120	100	30.5	29	0
200	200	150	37.8	53	1
125	130	500	40	55	1
110	120	40	45.8	31	1

The next step is to build the CART model for the dataset. This can be done after the data is split into training and test set. Pruning should not be done on the data set. Performance is to be evaluated for the test set. Figure 4 shows the various levels of depth of the classification tree for diabetes.

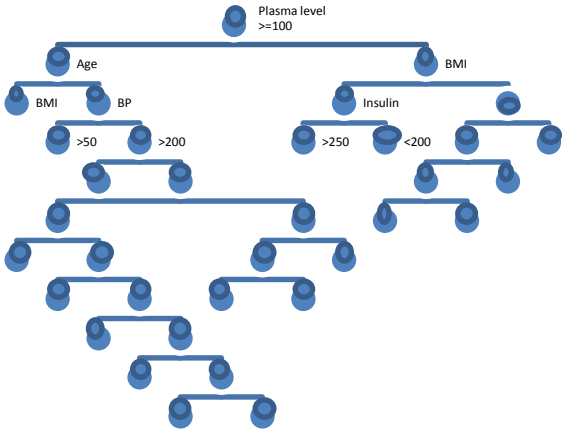

Fig 4. Classification tree for diabetes

Random Forest model

In this section we build the random forest model with R package[17] by using 10 variables defined at each split. We build 800 separate decision trees. The Out of Bag (OOB) estimate of the error rate is 25.04%. This model has identified that the plasma levels, BMI, Age, and family history of diabetes are the most important factors for diabetes.

VI. RESULTS

The performance of the classification model can be represented by the confusion matrix on the set of EHR data set where the true values are known. By using the CART model the classifier made a total of 400 predictions. The confusion matrix for the model is shown in Table 2.

TABLE 2: CONFUSION MATRIX FOR CART MODEL

N=400	Predicted		
Actual		No Diabetes	Yes Diabetes
	No Diabetes	40	20
	Yes Diabetes	120	220

The accuracy of the classifier can be found by the following formula

(True Positive + True Negative)/Total = Accuracy Value

Where True Positive is the case where the disease is predicted and disease exists. True negative is the case where the disease is not predicted and they do not have the disease.

(220 + 40)/400 = 0.65

Now we use the Random forest model for predicting the disease. The confusion matrix is given in Table 3

TABLE 3: CONFUSION MATRIX FOR RANDOM FOREST MODEL

N=400	Predicted		
Actual		No Diabetes	Yes Diabetes
	No Diabetes	50	20
	Yes Diabetes	80	250

The accuracy of the classifier is (250 +50)/400 = 0.75

Now we use the Scalable Random Forest algorithm to prove the accuracy of the results. This is shown in Table 4.

TABLE 4: CONFUSION MATRIX FOR SCALABLE RANDOM FOREST MODEL

N=400	Predicted		
Actual		No Diabetes	Yes Diabetes
	No Diabetes	50	10
	Yes Diabetes	40	300

The accuracy of the classifier is (300 +50)/400 = 0.875

Hadoop cluster distributed environment can be used to run the scalable Random Forest Algorithm and the results are calculated. The Scalable Random forest Algorithm provides the accurate results for the diabetes data set.

VII. CONCLUSION

In a distributed computing environment processing the massive data is done based on MapReduce model. In order to find the accuracy of the patient data the classification model will be helpful. This paper is aimed to find the nearest accuracy of the classifier by using the scalable random forest algorithm. The CART model and Random forest is built for the data set and the accuracy of the classifier is found. Results shows that by using the Scalable Random forest algorithm we can get the nearest accuracy of the prediction.

REFERENCES

[1] Borthakur, D. The Hadoop Distributed File System: Architecture and Desgin, 2007.

[2] J Dean and S Ghemawat, "MapReduce: Simplified data processing on large clusters," Commun ACM, Vol 51, no 1, pp 107-113, 2008.

[3] https://mahout.apache.org/users/basics/algorithms.html

[4] Linda A Winters Miner, "Seven ways predictive analytics can improve healthcare", Elsevier Connect

[5] https://en.wikipedia.org/wiki/Data_analysis

[6] F. Randy Vogenberg, "Predictive and Prognostic Models: Implications for Healthcare Decision-Making in a Modern Recession", Am Health Drug Benefits. 2009 Sep-Oct; 2(6): 218–222.

[7] Kenney Nga, Amol Ghotinga, Steven R. Steinhublb, c, Walter F. Stewartd, Bradley Maline, f, Jimeng Suna PARAMO: A PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records, Journal of Biomedical Informatics, Volume 48, April 2014, Pages 160–170.

[8] J. Breiman, L, Random forests, Machine Learning 45(1), 5-32 (2001).

[9] Jiawei Han, Yanheng Liu, Xin sun, "A Scalable Random Forest algorithm based on MapReduce", PP 849-852, 2013 IEEE, 978-1-4673-5000-6/13.

[10] Jerry Zhao, Jelena Pjesivac-Grbovic, MapReduce: The programming model and practice, SIGMETRICS (2009).

[11] Edwin Rietveld, Hendrik CC de Jonge, Johan J Polder, Yvonne Vergouwe, Henk J. Veeze, Henriette A. Moll and Ewout W.Steyerberg. Anticipated costs of hospitalization for respiratory syncytical virus infection in young children at risk. The Pediatric infectious Disease journal, 23(6):523-529, 2004.

[12] Michael A. Cucciare and William O'Donohue, Predicting future healthcare costs: how well does risk-adjustment Work? Journal of Health Organization and Management, 20(2):150-162,2006.

[13] Trevor Hastic, Robert Tibshirani, Generalized additive models, Statistical Science, 1(3), 297-310, 1986.

[14] Daryl pregibon, Logistic Regression diagnostics, The Annals of Statistics, 9:705-724,1981.

[15] Leo Breiman and Jerome H Friedman, Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association, 80(391):580-598, 1985.

[16] Kevin P Murphy, Machine Learning: A probabilistic Perspective, The MIT Press, 2012.

[17] https://cran.r-project.org/bin/windows/base/