Integrated Prediction Model for Huge\Big Healthcare Database

Samaher Al_Janabi
Faculty of Information Technology (IT), University of Babylon, Iraq
samaher@itnet.uobabylon.edu.iq

Hayder Fatlawi
Faculty of Information Technology (IT), University of Babylon, Iraq
hayder.alnajfi@gmail.com

Abstract

Prediction techniques represent an effective tool for knowledge discovery in huge and complex dataset in many fields including healthcare. The problem of healthcare is managing available medical resources and preparing plans for the future needs aiming to enhance medical services. This research provides an integrated prediction model to solve the problem above by analyzing medical data records and predicating the duration of future patient's hospitalization. The proposed model consists of three major stages; starting with preprocessing the data; applying prediction algorithm; and ending with evaluating the model based on real data. Our model used Gradient Boosting Machine (GBM) algorithm which reduce training error by building a sequence of decision trees. GBM is characterized by updating values of target feature after the construction of each decision tree. In this research, we tried to discover the effect of reducing the update process on terminal nodes that have lowest percentage of error, the results showed the ineffectiveness of reduction compared to the original. The research tried to determine the best measure for choosing splitter feature during the building of the decision tree, and the results showed that standard deviation is better than mean. The research also studied effect of changing values of GBM algorithm parameters in behavior of training process.

Keywords: Prediction Techniques, Healthcare, Data mining, Gradient Boosting Machine.

الخلاصة

تمثل تقنيات التنبؤ اداة فعالة لاكتشاف المعرفة من قواعد البيانات الضخمة و المعقدة في مجالات عدة منها العناية الصحية. تتركز مشكلة العناية الصحية في ادارة المصادر الطبية المتاحة و اعداد خطط للاحتياجات المستقبلية لتحسين مستوى الخدمات الطبية للمرضى. يقدم هذا البحث مودل تنبؤ متكامل نسبياً يهدف الى حل هذه المشكلة من خلال تحليل بيانات المرضى الطبية و مراجعاتهم السابقة و التنبؤ بعدد ايام المراجعة المحتمله للمستشفى مستقبلا. يتكون المودل المقترح من ثلاث مراحل رئيسية تبدا بمعالجة مسبقة للبيانات ثم تطبيق خوارزمية التنبؤ على البيانات الناتجة و تنتهي بتقييم المودل اعتمادا على بيانات حقيقية. يستخدم المودل خوارزمية (ماكنة زيادة الانحدار Gradient Boosting Machine GBM) والتي تقلص خطا التنبؤ من خلال بناء سلسلة من اشجار القرار اثناء عملية التدريب. تتميز خوارزمية GBM بعملية تحديث قيم المتغير المطلوب التنبؤ به بعد بناء كل شجرة قرار. في هذا البحث حاولنا اكتشاف تاثير تقليص عملية التحديث على العقد النهائية ذات اقل نسبة خطا حيث اظهرت النتائج عدم فعالية التقليص مقارنه بالطريقة الاصلية. كما حاول البحث تحديد المقياس الافضل لاختيار متغير التقسيم اثناء عملية بناء شجرة القرار و اظهرت النتائج افضلية استخدام standard deviation على استخدام mean . ركز البحث ايضا على دراسة تاثير تغيير قيم معاملات خوارزمية GBM في سلوك عملية تدريب المودل.

الكلمات المفتاحية: تقنيات التنبؤ، العناية الصحية، تنقيب البيانات، ماكنة زيادة الانحدار.

1. Introduction

Healthcare is concerned with applying all necessary medical procedures to restore health of people or prevent aggravation of health problems [raym13]. The field of health care and medical services in Iraq faces an acute shortage of medical resources in terms of the number of hospitals and medical community and primary health care and emergency services, as well as appliances and medical laboratories, and because of a period of economic blockade and war. Addressing this deficiency requires the preparation of plans

for manage the available resources and to provide additional resources in the future. Many of tools are used to deal with challenges of healthcare problems such as data mining techniques.

Data mining is gaining useful knowledge from data. It has three main phases; the data is prepared to mining process then applying machine learning techniques for knowledge extraction and finally the results will be processed in understandable form to be helpful for decision making. Mainly, data mining is used for two purposes; discriminate hidden patterns in data or predicate unknown values of interested features form known values of other features [Pang05, Jiaw06].

Predictive analysis could be defined as the task of data analysis to predict unknown values of prediction target feature. It includes classification task for class label prediction and numerical prediction where the task is to predicate continuous values or ordered values. Many statistical methodologies are used for numerical prediction and regression analysis is most often used [Jiaw13].

Regression techniques are various according to the complexity of data of prediction process. The relation between target variable and other interested variables specifies how predication is hard, with linear relation we can use technique like linear regression and with complex relation more advanced techniques are needed. Type of target variable specifies if the problem is classification with binary values or numerical prediction with continuous values.

Decision tree (DT) is easy and effective tool for binary class problems, it uses recursive binary splitting to build prediction model. It is developed to handle regression problem when multiple values for target variable are founded. The decision made by a single regression binary tree may be not precise especially with high complex data [Step11].

Boosting is a machine learning method for building strong predictor from set of weak predictors to overcome problem of bias in prediction model. Gradient boosting machine (GBM) is one of the best implementations of boosting regression trees which depends on the concept of reduce training error by building sequence of trees [Elit08]. It is used and developed in this thesis for solving healthcare problem which mentioned in section 2.1. Figure (1) summarizes types of prediction techniques and boosting methods.

The statement of healthcare problem could be described as management of medical resources needs future sight of patient requirements. Accurate prediction of which patient will admit in hospital in next year could be useful in this management.. The problem is development prediction model can deal with those datasets and produce high accuracy.

The idea of this research is to introduce prediction model for predicate the future hospitalization of patients using ensemble machine learning and produce regression model.

The remained of this paper layout is as follows: Section 2 discussed review of related work; Section 3 explained the challenges and objectivity of research; Section 4 showed the main stages of building proposed model that is called a Modern Prediction Model for HealthCare Problem (MPM-HCP). We can show the main results of implementation MPM-HCP with all details in Section 5. Finally, Section 6 presents the conclusion of this research.

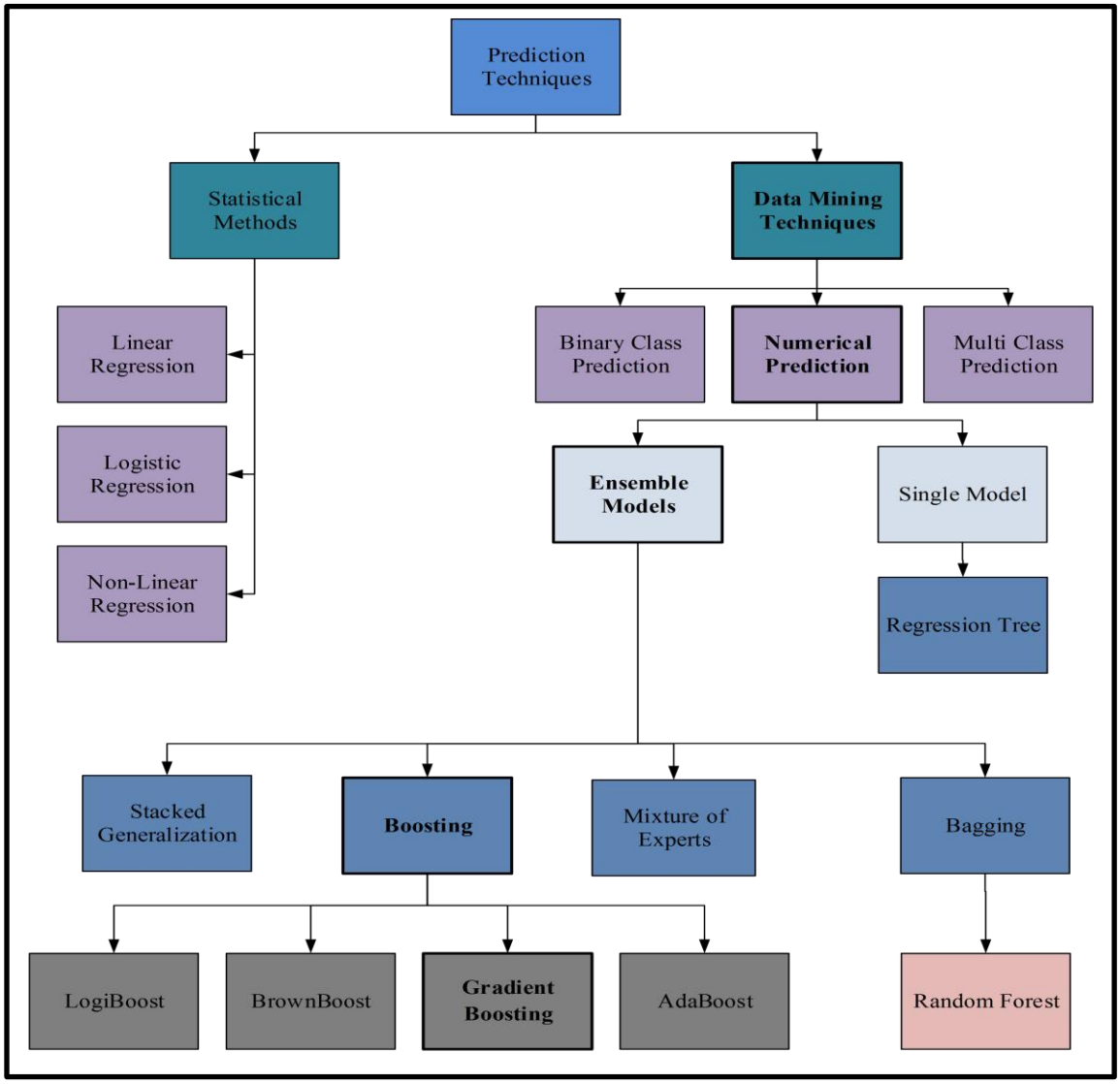

Figure (1) Prediction Techniques where bold text indicate to the type of our model

2. Related Works

Data mining techniques have been applied in many researches to solve the problem of healthcare. Related works in this field are various between recommended systems and prediction model for unnecessary hospitalizations. We will review those works in this section and explain the points of similarity and dissimilarity with our work.

Dimitris et. al., 2008 [Dimi08] utilized modern data mining methods for healthcare resources problem by aggregate many sources of claims data and classify them using classification and clustering with some performance measures like hit ratio, penalty error and R^2. Dimitris's work focused on data preprocessing and used typical data mining techniques for domain of healthcare, while our paper focus on improving a prediction model for the same domain in addition to data preprocessing.

Duana et. al., 2010 [Duan11] created a recommended nursing clinical system to help make right decision and improve clinical quality control by using association rules to find patterns in item sets of community hospital in the Midwest dataset and use support, confidence, and lift as utility evaluation measurements. Duana's work was a descriptive task for dealing healthcare problem while our paper focuses on a predictive task to solve this problem.

Xiang et. al., 2011 [Xian11] used machine learning algorithms to reduce unnecessary hospitalizations by predicate how long the patient will stay in hospital in the next year according to his record in last year ,researchers use support vector machine (SVM), random forest, regression tree and boosting ensemble with HPN 2011 Dataset. Xiang dialed with healthcare dataset using multi-class and regression techniques, our paper focuses on development prediction based on regression technique with same dataset.

Rashedur and Fazle, 2011 [Rash11] Used and compared different decision tree classification techniques to classify admitted patients according to their critical condition and develop application to diagnose and measure the criticality of the newly arrived patient to mining hospital surveillance unit using C4.5 Decision Tree Classifier and evaluated result using False positive rate (FP) ,Recall and Precision. We are similar with this work in the domain of healthcare but we vary in using different dataset, evaluation measures and regression rather than classification.

Jufen et. al., 2013 [Jufe13] constructed a Chi-Squared Automatic Interaction Detection (CHAID) classification tree with 10-fold cross-validation to predict probability of death or hospitalization for heart failure and compared the result with logistic regression (LR) models using ROC curve analysis based on TEN-HMS Dataset , they found CHAID tree performed better than the LR-model for predicting the composite outcome. Our thesis is similar to Jufen's work by utilizing same task of data mining that is the prediction for healthcare while varies in using different dataset and different technique of prediction.

Nannan, 2014 [Nann14] compared performance of three prediction techniques (i.e., linear regression, random forest and gradient boosting) on hospitalization dataset to discover which technique is the best. By experiments, Nannan found that the random forest technique provides the best prediction of patient hospitalization. Nannan's work used typical prediction techniques while our thesis aims to develop a new prediction model for same dataset.

3. Challenges and Objectivity of Research

The data sets of healthcare have many challenges like huge amount of data, high percentage of zeros values in target variable because most of patients don't stay in hospital for long time. Another challenge is low correlation among interest features and target feature. GBM algorithm also has a challenge to detect suitable parameters for building prediction model. Those challenges make prediction processes more complex.

The main objective of this work is to design and implement predication model for providing future information about patient's hospitalization. We proposed a Modern Prediction Model for HealthCare Problem (MPM-HCP) that overcomes the weakness of a single regression predictor by utilizing boosting concept. MPM-HCP integrates preprocessing of medical data records with data mining technique to produce predicted information having less prediction error and less execution time.

The final result of MPM-HCP could be utilized by many beneficiaries to enhance the medical services. Ministry of health can use predicted information for preparing future plans about the requirement of the health sector such as number of hospitals, number of medical staff and amount of drugs then provide the required financial and human resources to cover these needs. Hospitals can employ result of prediction model to improve the management of available medical resources according to the needs of patients. Medical behavior could be discovered using this result which supports avoiding unexpected decline of health status of patient and pre-estimates the future requirement of healthcare for that patient.

4. Design of MPM-HCP

In this work, a boosted prediction model is presented to solve healthcare problem by training boosted regression tree based on real and complex dataset to get lowest prediction error and less time. Proposed includes three stages; the first stage is dataset preprocessing that includes features transformation, features construction, data row aggregation and tables combination. The second stage involves building prediction model including parameters detection and regression trees building, final stage is the evaluation of the result of the previous stage .Those stage are declared in Algorithms (1) and illustrated in Figure (2).

Algorithm (1): MPM-HCP

Input: *Tr*: Healthcare Training Dataset, *Ts*:Healthcare Testing Dataset, Tmax: maximum number of trees, *Tnmax*: maximum number of terminal nodes, *Smin*: minimum number of samples in terminal node, *RC*: training row count, *Shr*: shrinkage rate, Tmodel: regression trees model.

Output: Number of days in hospital for each patient in next year.

Step 1: Call Pre-Processing Procedure for Raw Healthcare Dataset.

Step 2: Split Healthcare Dataset into Training dataset *Tr* and Testing Dataset *Ts*.

Step 3: For I=1 to *Tmax*, do the following for each tree *ti* :

- T*i* = Tree_Building (*Tr* , *Tnmax* , *Smin*, *Rc*)
- Er*i* = Training_Error_Calculation (*Ti* , *Tr*)
- For J =1 to *Number_of_terminal_nodes(Ti)*
- For K=1 to *Number_of_Records_of_Node J*
  - Update Target values *Dt* of *D* as follow :
  - Trt [i+1,j,k] = Trt [0,j,k] + (Shr × mean(Trt[i,t,j)

  End for

End for

End For

Step 4 : Testing Tmodel on Test dataset : Testing_Trees_model (Tmodel , *Ts*)

End

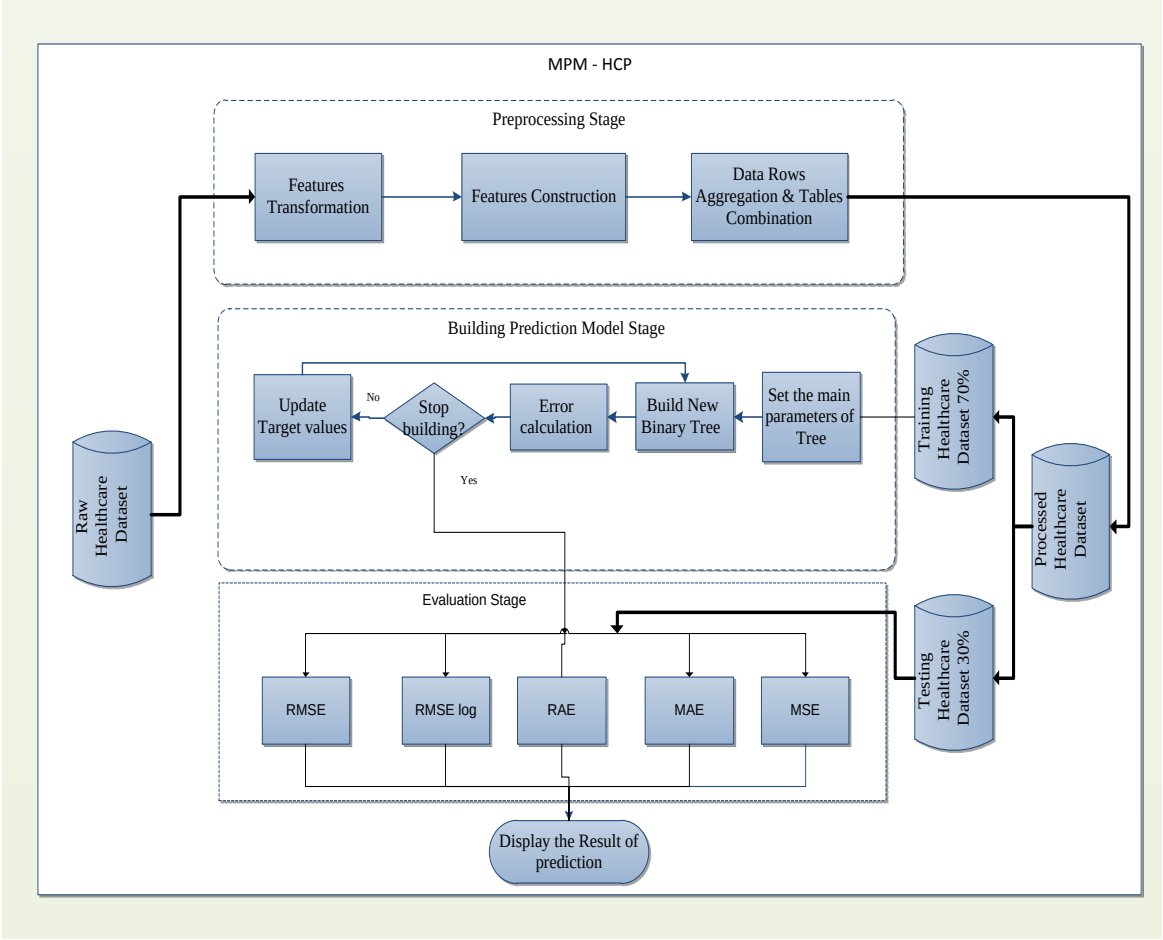

Figure (2) Block Diagram of Proposed Model

4.1 Preprocessing Stage

What is the reason for focusing on preprocessing in this research? Real datasets like health care dataset usually have some inappropriate characteristics that need preprocessing before use it. Preprocessing may deal with missing values, noisy data, features selection, normalization, discretization and data reduction. MPM-HCP performs new preprocessing methods on healthcare dataset as explained in **step 4** and **step 5** of ***Procedure of Preprocessing***:

i. Features Transformation: Some of numerical features of health care dataset (such as age) are discretized to categorical features by replacing numerical values with intervals. It reduces computational operations of choosing best split point during building of decision tree.

ii. Features Construction: New set of features are constructed to extend the original features set which reduce error of prediction. ***(Hint: There are two reasons for doing this step: (1) high prediction error when use the original features set (2) Aggregate all claims of specific year to be single record require separate the sub categories to be new features)***. This step performs by calculate statistical information such as maximum value, minimum value, mean and standard deviation and retrieve sub features. Claims of a patient could be one or more than one in a specific year and each claim may have different value of features, like Specialty which may be

Surgery in the first claim and Diagnostic Imaging in the next one. Aggregations of all claims in a year to be one record, which perform in the next step, require making new set of features by splitting some features into its sub features. The value of new features would be a counter indicates the number of presence of sub feature in that year. Section 5.1 clarifies construction of new features from values of original features.

iii. Data Rows Aggregation: All data rows, which mean claims on healthcare related to a patient in one year are aggregated to be a single row. Structure of healthcare data set consists of target table which contain the number of days in the hospital and related with other medical and personal information by patient identifier, this structure is considered the main challenge of finding the number of hospitalization days in a next year which enforced us to make this aggregation.

iv. Tables Combination: Tables of healthcare dataset are combined to collect all Members, Claims of hospitalization, drugs and labs information together to be one table depending on patient identifier . The reason of this step is to reduce the complexity of training processing which can be more efficient in the process of tree building than dealing with separated tables.

Procedure of Preprocessing

Input: D: Raw Healthcare Dataset, *Fnum* : Number of features.

Output: Processed Healthcare Dataset.

Step1: For all numerical features D, discretize them features by replacing numerical values with intervals.

Step2: For J=1 to *Fnum* in D, construct a new features as follow :

- ***For each categorical feature xj in features set of D, construct a new features form sub features of xj.***
- For each numerical feature *xj* in features set of D, construct a new features form statistical operations on *xj* values *(max, min, average, STD).*

Step3: For all data rows in D, do the following:

- Find all data rows that related to same patient identifier and same year.
- Aggregate those data rows to be single record.

Step4: Merge all tables in D in one table depending on patient identifier.

End

4.2 Building Prediction Model Stage

This stage could be considered the core of GBM, it starts with detection of some important parameters which are needed in boosted regression then, in each iteration, binary tress is built to reduce the error of previous one until stop condition satisfied. Both mean and standard deviations are used in the model as splitting criteria of choosing best splitter feature inside binary tree procedure. The steps of this stage are clarified in Figure (2) and could be explained as follow:

i. Parameters detection

The first step in building reliable prediction model is to choose the parameters of the algorithm carefully. Gradient Boosted regression has four main parameters which should be selected [Elit08]:

- Maximum number of trees in the model that controls the execution of the algorithm.
- Maximum number of terminal nodes in every single binary tree that controls the number of rules.
- Minimum number of samples in each terminal node that effects the coverage of the rule of that node.
- Shrinkage that represents the learning rate in the training process.

ii. Build Binary Regression Tree

While current number of trees in the model is still less than maximum number of trees, new binary tree is created. Each tree works in the same set of features except the values of

target of prediction which is the error of previous tree and it must be minimized.

The algorithm of build the regression tree consists of recursion procedure to as follows :

A. Best Splitter Feature Choosing: The most important step in binary tree building is to choose splitter feature which has more relation with target according to specific criteria. Binary tree for classification problem depends on information gain, gain ratio and Gini index for choosing best feature while regression problem uses mean and standard deviations. GBM use mean and standard deviation to find the quality of each feature for splitting. For every level in binary tree building, goodness of all features will be tested by calculating standard deviation (or mean) of target values before and after splitting data. The Feature that minimizes standard deviation value after splitting will be chosen as a splitter because the data are more homogeneous when standard deviation closest to zero. Figure (3) clarify evaluation of goodness of features for splitting data in regression binary tree.

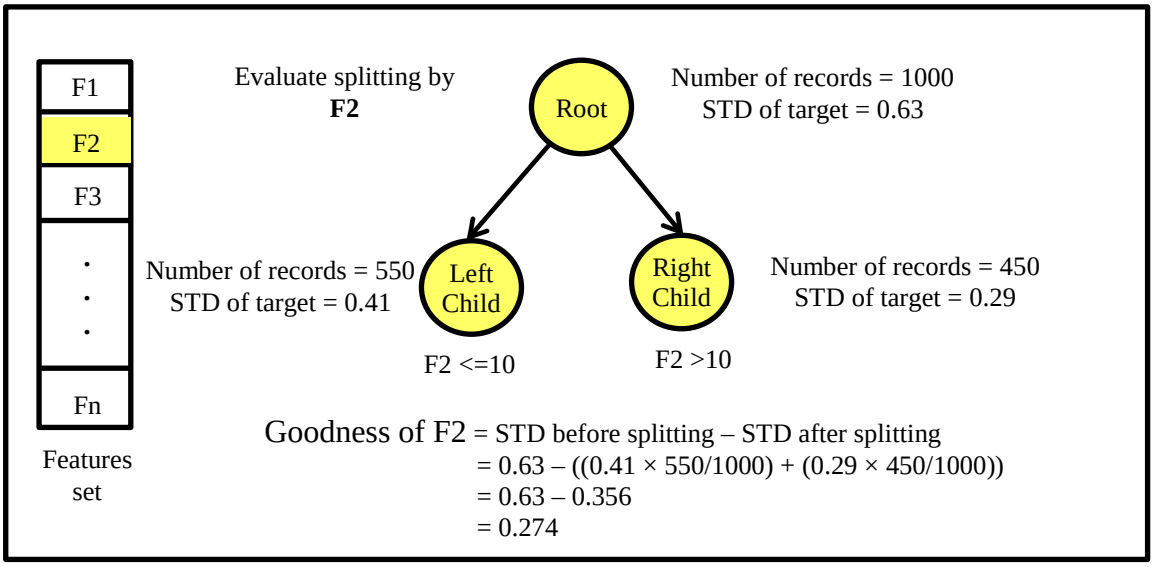

Figure (3) Evaluation of goodness of features for splitting data in regression binary tree

B. Best Split Point Choosing: After best feature is chosen, best split point from all possible points should be selected. The typical method is to try all possible split point and evaluate the quality of each point, this operation is costly computational task because the effort increases as the values of feature increases for example with continues feature which has k values there are k-1 possible split point.

C. Data Splitting: In this stage, current data are divided into two parts according to the condition of best split point which is detected in previous section. The data rows that have value of splitter feature less than split point become the data of new left child node and those have value equal or larger than split point become the data of new right child node.

D. Prediction Value in Terminal Nodes: Each path from the root of tree to a terminal node is a rule which consists of many of conditions. The rule should lead to predicate a specific value for target of prediction. In the terminal node, mostly there are more than one data row and each one has a target value, the prediction value of specific node simply could be the mean of all values of data rows of such node.

iii. Training Error Calculation

Gradient Boosted regression depends on the concept of minimizing gradient of error [Frei01]. The error of every data row is calculated by subtracting the original target value of that data row from predicted value of it. Average training error for specific tree could be founded by dividing the summation of the error of all data rows by the number of them. Algorithm 3 explain required operations for this step.

iv. Target Values Updating

The value of target in each data row is updating by adding prediction of new tree multiplying with shrinkage rate. This step considers the most important in Gradient Boosted regression and it is differentiate this technique from other prediction techniques. It could be explained as looking for the best point which close to all target value as possible. Updating of target values during training process is clarified in Example (1) which showed how predicted value is closing to desired value during training of boosted model.

Example1:

Desired target value of row0 = 1, learning rate = 0.008
Initial target guess of row0 = 0
Predicated target value of row0 in tree0 = 0.23
Target value of row0 for tree1 = 0 + (0.008 × 0.23) = 0.0018
Target value of row0 for treeN= 0 + (0.008 × target value of row0 in treeN-1) = 0.61

4.3. Evaluation Model Stage

In this stage, all trees model that build by Gradient Boosted regression are evaluated based on test dataset that never seen before. MAE, MSE, RAE, RMSE, RMSE Log and measures are used in this step to evaluate the prediction error of every tree then find the total error for combination of all trees model. The equations of measures that mentioned above could be described as follow:

(Where n: number of data rows, yi: actual target value of record i, $y'i$: predicted target value of record i):

- Mean Absolute Error : $MAE = \frac{\sum_{i=1}^{n} |yi - y'i|}{n}$ …… (1)
- Mean Squared Error : $MSE = \frac{\sum_{i=1}^{n} (yi - y'i)^2}{n}$ …… (2)
- Relative Absolute Error : $RAE = \frac{\sum_{i=1}^{n} |yi - y'|}{\sum_{i=1}^{n} |yi - y"|}$ …… (3)
- Root Mean Squared Error : $RMSE = \sqrt{\frac{\sum_{i=1}^{n} (yi - y'i)^2}{n}}$ …… (4)
- Root Mean Squared Error with Log : $RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\log(yi+1) - \log(y'i+1))^2}{n}}$ …… (5)

5. Experimental Results

A real and a huge dataset have been used as implementation example to discover the behavior of the proposed model. Heritage Provider Network (HPN) provided healthcare dataset for researchers and data miners aiming to reduce the costs by predicting hospitalization of patients. It contains data of more than (113000) patients in nine tables which linked by patient identifier and describe in Table 1. Each patient has one or more than one claims on a year, the data of claim included information about the condition that causing hospitalization and the medical procedure which required for patient treatment. Features of claims table that used in our model are explained in Table2.

Table (1) Description of Healthcare dataset

Table Name	Description	Number of Features
Claims	The main table of healthcare dataset and contains information of medical case and patient.	14
Members	Contain personal information such as age.	3
DaysInHospital_Y2	Contain the number of days which spend by patient in second year, it used in training stage.	3
DaysInHospital_Y3	Contain the number of days which spend by patient in third year, it used in training stage.	3
DrugCount	Number of drugs that consumed by patient.	4
LabCount	Number of Labs tests that consumed by patient.	4
PrimaryConditionGroup	Describe of Primary Condition Group coding.	2
LookupProcedureGroup	Describe of medical Procedure Group coding.	2

The values of days in hospital are between 0 and 15, the duration more than 15 days is rounded to 15 in this dataset. Those integer values made the prediction process consider as a regression problem and the result may contain real value while grouping those values in multiple ranges could make the process as a multi class classification problem by grouping target values in specific ranges. Our model treats this dataset as a regression model to get more precise and reliable results which can reduce the total error of prediction.

5.1 Apply Preprocessing on HPN

All claims of a year of specific patient are aggregated which make process of choosing best feature as a splitter in building tree step more efficient. This aggregation reduces the overall complexity of prediction problem with respect to the knowledge in HPN dataset. The number of data rows of claims table would reduce from more than one million to 147473 records after aggregating. Features construction which explained in section 4.1 is performed on PrimaryConditionGroup, Specialty, ProcedureGroup and PlaceSvc features of HPN dataset, Table (3) describe features construction on Specialty and ProcedureGroup features.

Before starting training process, we need to combine All tables of HPN dataset which are explained in Table 2 in one table that contains (136) features including the target of predication and (147473) data rows. This dataset are considered as an input of training process to build prediction model. HPN dataset is used to build a boosted set of regression tree in which the training error should be optimized. Dataset has been split into two parts; 70% for training stage (103231 records) and 30% for testing stage (44242 records).

Table (2) Description of Claims table

Feature Name	Description	Number of Categories
Specialty	General specialty of patient's condition.	13
PlaceSvc	General place of healthcare service.	9
PayDelay	Number of days delay between the date of service and the date of payment.	163
LengthOfStay	Number of days of hospitalization that mean the delay between discharge date and admission date.	10
DSFS	Days since first claim.	13
PrimaryConditionGroup	Code of medical diagnostic which describe in PrimaryConditionGroup table.	46
CharlsonIndex	A value of affect diseases.	4
ProcedureGroup	Code of procedure diagnostic which describe in ProcedureGroup table.	18
SupLOS	Binary value of suppression of claim.	2

Table (3) ProcedureGroup and PlaceSvc Features Construction

ProcedureGroup		Specialty	
Original value	New feature	Original value	New feature
EM	Prgr_EM	Internal	Spc_Int
PL	Prgr_PL	Laboratory	Spc_Lab
MED	Prgr_MED	General Practice	Spc_Gen
SCS	Prgr_SCS	Surgery	Spc_Surg
RAD	Prgr_RAD	Diagnostic Imaging	Spc_Daig
SDS	Prgr_SDS	Emergency	Spc_Emg
SIS	Prgr_SIS	Pediatrics	Spc_Ped
SMS	Prgr_SMS	Rehabilitation	Spc_Reh
ANES	Prgr_ANES	Obstetrics and Gynecology	Spc_Obst
SGS	Prgr_SGS		
SEOA	Prgr_SEOA	Anesthesiology	Spc_Anes
SRS	Prgr_SRS	Pathology	Spc_Path
SNS	Prgr_SNS	Other	Spc_Other
SAS	Prgr_SAS		
SUS	Prgr_SUS		

The benefit of preprocessing stage is clear on Figure (4), the difference between training error before and after applying steps of preprocessing is distinguishable. Gradient boosting machine (GBM) had 0.668 error measure when applying on HPN dataset without preprocessing, the error reduced after 80 iteration to reach 0.49, while preprocessed dataset produced error starting from 0.515 and ending with 0.48 according to RMSE log measure.

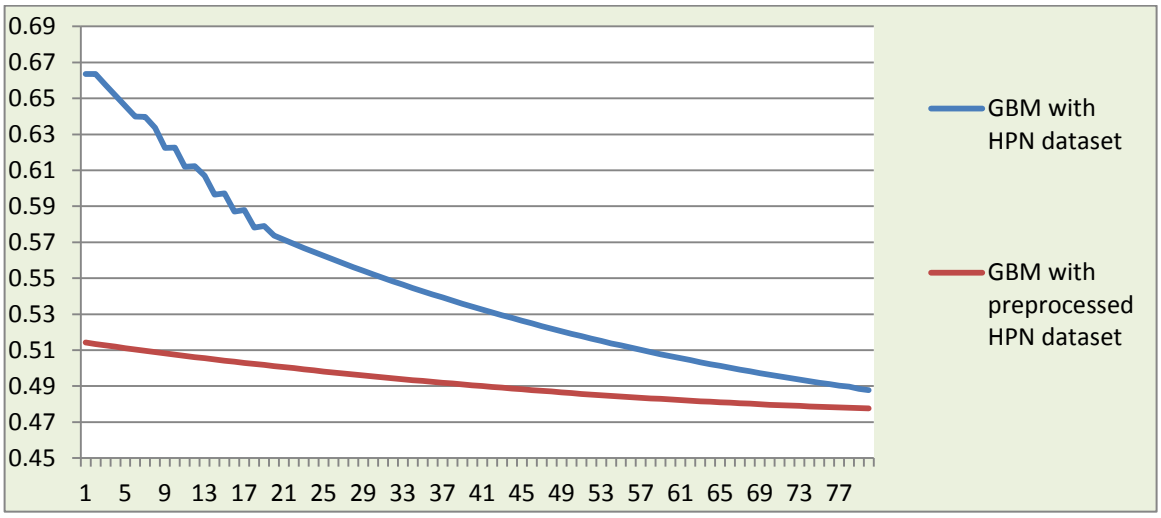

Figure (4) Comparison between Training Error of GBM before and after preprocessing HPN dataset with RMSE log measure

5.2 GBM Parameters Values Selection

Choosing of parameters roles the behavior of training process and affects the result of this process. Multiple values of Gradient Boosted Machine parameters have been used with HPN dataset aiming to find optimal values for them. Shrinkage parameter is tested with the rang of values (0.1 – 0.001) and the best result was with the range (0.005- 0.008) with respect to other parameter. Another parameter should be selected for GBM is the maximum number of terminal nodes which related with complexity of trees and number of rules. The popular depth of regression tree in GBM is between 2 and 4 levels which mean the maximum number of terminal nodes is in the range of (4-16) nodes.

In this implementation, depth of three ,four and five levels are tested and the stable result was getting using five levels which means the maximum number of terminal nodes with HPN dataset was (32) nodes that are considered medium complexity for dataset with (135) features. Third parameter in GBM is the minimum number of samples in terminal nodes which mean if the number of data rows in the node is equal or less than this parameter, the node would be terminal node and the split operation should stop. This parameter was set to (20) samples with HPN dataset which represents (0.000017) ration from total number of data rows.

The number of trees is the most important parameter in this process which needs to be chosen carefully, with HPN dataset, the rang of (80 – 160) tree is used. The effect of increasing the number of trees could be seen in Table (3). In the original GBM that used standard deviation the training error is reduced and reached to the lowest error (0.466) after 104 iterations of training then it was increased again. In the mean based GBM, the lowest training error (0.475) has been reached after 99 iterations.

Table (3) Comparison between training error of original GBM based on STD and GBM based on Mean according to the number of trees in training process

No. of Trees	Original GBM-STD		Original GBM-Mean	
	Best Tree	Training Error RMSE Log	Best Tree	Training Error RMSE Log
80	80	0.4693976	80	0.4775828
120	104	0.4669166	99	0.4759887
160	104	0.4669166	99	0.4759887

5.3 Criteria of Choosing Best Splitter Feature: Mean vs. Standard Deviations

The quality of each feature for splitting process is evaluated by Mean or Standard deviations. Two observations could be seen after building (160) regression trees in both Mean and STD:

- **First:** observe that GBM based on standard deviation has training error less than GBM based on Mean.
- **Second:** the gradient of error was increasing after building (100) tress in both Mean and STD. Both observations are shown in Figure (5).

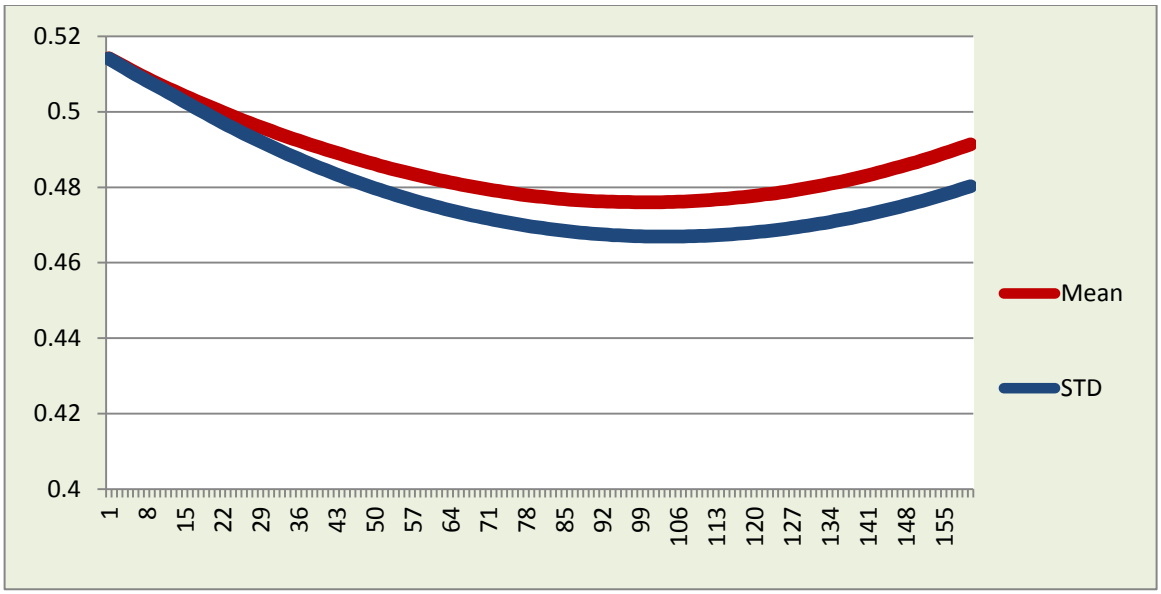

Figure (5) Comparison between Training Error GBM based on mean and GBM based on STD with RMSE log measure

5.4 Target Values Updating: All Nodes vs. Best Node and Worst Node

The core of Gradient Boosting Machine is updating the values of target feature during sequence of trees building. This process applies to all data records in all terminal nodes. We inspired two methods from pruning concept to reduce the updating process on some nodes according to the prediction error. The first one updates data rows in the best node only which is chosen depending on the ratio of prediction error and number of data rows in this node. This method reduces the effect of each regression tree in the best node only, and the final model would be a combination of best nodes. Second method depends on updating data rows in all nodes except the worst node in every tree. The comparison among original GBM and these two methods observed that the training error would increase after reduce updating target values. Both first method and second method have training error more than original GBM as shown in Figure (6).

5.5 Evaluation Model Stage

Real and trusted evaluation for data mining techniques should be based on test data which have never seen before. HPN dataset is spitted into two parts in which the test data are 30% from original data that means the number of test data rows is (44242) records. Training stage is finished with building (160) regression trees that needs to be evaluated.

There are two possible methods to test the final model; first method is to take the final tree for testing only. The final tree has accumulative model from all previous trees, in this method original GBM based on mean had (0.502) training error in RMSE log measure while original GBM based on standard deviation had training error of (0.482). Table 4 shows comparison of this method using five error measures.

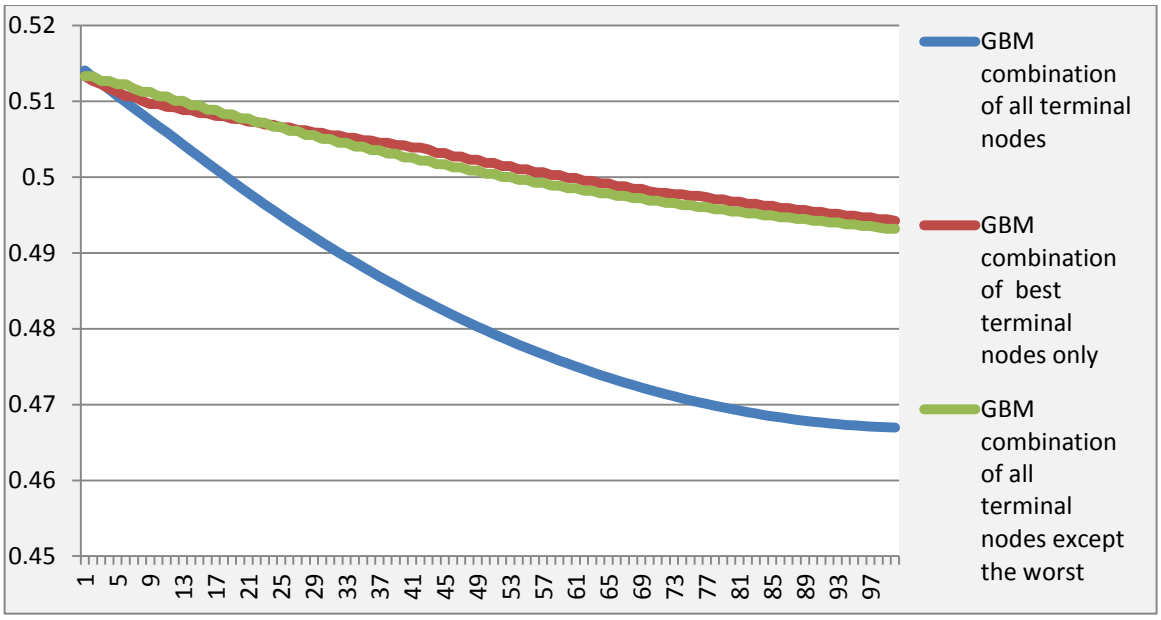

Figure (6) Comparison among Training Error of GBM with update all nodes, update best node only, and update all nodes except worst node with RMSE log measure

The second method is to use all trees which build in training process for evaluating the model and control contribution of each tree by learning rate parameter. This method includes testing every record on (160) tree sequentially and multiplying the result of each tree by shrinkage learning rate. In this method original GBM based on mean had (0.502) training error in RMSE log measure, original GBM based on standard deviation had training error of (0.482). Table 5 shows comparison of this method using five error measures.

Table (4) Comparison of Testing Error evaluation among GBM based on Mean, GBM based on STD with single last tree

	MAE	RAE	MSE	RSSE	RMSE Log
GBM based on Mean	0.920810	1.197997	2.530408	1.006323	0.588864
GBM based on STD	0.862924	1.122686	2.444053	0.989003	0.566087

Table (5) Comparison of Testing Error evaluation between GBM based in Mean, GBM based on STD with all trees combination

	MAE	RAE	MSE	RSSE	RMSE Log
GBM based on Mean	0.694361	0.903381	2.473407	0.994924	0.491753
GBM based on STD	0.662535	0.861974	2.444053	0.982422	0.481393

6. Conclusions

Healthcare dataset is considered complex challenge for prediction techniques because a huge amount of medical data records and imbalanced distribution of target feature values. Gradient Boosting Machine presents a reliable solution for healthcare challenge through building a prediction model consists of sequence of regression trees. GBM use Mean or STD for choosing spitting feature, the experimental result showed that STD had less training error than Mean. The results discovered a major effect of parameters of GBM algorithms in training process. The best result was gained by using maximum number of trees around (100) trees and we observed that training error was increasing after this range. Another important observation was the effect of reducing update target values on best node for each tree, the training error was worse than original method which means pruning any rules from regression tree would cause increasing the training error.

References

Duana, L.; Streeta W.N. and Xu , E. "Healthcare information systems: data mining methods in the creation of a clinical recommender system", Enterprise Information Systems Vol. 5, No. 2, 2011.

Elith, J.; Leathwick, J.R. and Hastie, T. "A working guide to boosted regression trees", Journal of Animal Ecology, vol (77), p.p 802–813, 2008.

Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", ISBN 10: 1-55860-901-6, Elsevier, 2006.

Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining: Concepts and Techniques", 3RD Edition, , ISBN: 978-0-12-381479-1, Elsevier, 2013.

Jufen Zhang, Kevin M. Goode, Alan Rigby, Aggie H.M.M. Balk and John G. Cleland, "Identifying patients at risk of death or hospitalization due to worsening heart failure using decision tree analysis: Evidence from the Trans-European Network-Home-Care Management System (TEN-HMS) Study", International Journal of Cardiology, vol (163), p.p 149-156, Elsevier, 2013.

Nannan He, "Data Mining for Improving Health-Care Resource Deployment", Master thesis, University of California Santa Cruz, 2014.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, "Introduction to Data Mining", (First Edition), Addison-Wesley Longman Publishing Co, USA, ISBN: 0321321367, 2005.

Rashedur M., Fazle Rabbi, "Using and comparing different decision tree classification techniques for mining ICDDR, B Hospital Surveillance data", Expert Systems with Applications vol(38),p.p 11421–11436, 2011.

Raymond L. Goldensteel, Karen Goldensteel, "U.S. Healthcare system", 7th Edition, ISBN: 978–0-8261–0930-9, Springer Publishing, 2013.

Stéphane Tufféry, "Data Mining and Statistics for Decision Making", First Edition, ISBN: 978-0-470-68829-8, John Wiley & Sons, Ltd, 2011.

Xiang Peng, Wentao and Wu Jia Xu, "Leveraging Machine Learning in Improving Healthcare", Association for the Advancement of Artificial Intelligence, 2011.