Building Prediction Based System For Future Length Of Stay Using Gradient Boosting Machine

[1]Er. Spinder Kaur,[2]Dr. Sandeep Kautish
Department of Computer Engineering
Guru Kashi University
Talwandi Sabo Bathinda,Punjab India.

Department of Computer Engineering
Guru Kashi University
Talwandi Sabo Bathinda,Punjab India.

ABSTRACT: Data Mining is the emerging field in the world. There are various applications which are directly and indirectly be producing large amount of data. This data will be very difficult to process manually. For extracting the useful facts there requires automatic processing machines. These machines can process the data and extract the useful facts. Similarly in case of health field prediction system is required which can predict the patient future length of stay in the hospital. This prediction is based on past data of the patient or various health parameters of the patient. It identifies the patient future length, so that patient can know the expected expenses that are going to be occurred for full medical procedure. Random forest and the Gradient Boosting machine are two basic techniques. In existing research paper Random forest based technique was used. In proposed research for enhancing the prediction accuracy and reduction in the error rate Gradient Boosting machine based technique is used. Its performance in both the perspective is better compared to the random forest.

KEYTERMS: FLOP, Random Forest, Gradient Boosting Machine.

I. INTRODUCTION

In recent time various types of services are emerging in the society. These services are related to the different fields of the society. Out of those fields major field is medical. India is a large country where large population resides. Various types of organized and unorganized medical facilities are available in the country. But due to the population explosion each facility remain scarce. To overcome and catalyst the growth in this part of the applications. Various researchers are involved which are growing with different researches so that the problem of scarcity of the resources can be catered without increasing the much cost.

In this researches cloud is one of the major thrust area. That can solve the problem of this medical mismanagement. According to this research paper there will be a cloud of the medical data. That is consisting of integration of various small city level data. Any company will provides processing ability which can process this large integration of the data. So that any patient data can be provided at the required place. His case history or we can say medical history can be recorded at each step.

These data are mainly stored and isolated in disparate local systems, and are underutilized in terms of data analysis and knowledge discovery. Cloud computing techniques have made computational infrastructure capable of handling such enormous information burst in a cost-effective way. Up to the present, most works focus on migrating healthcare IT system and data storage to the cloud platform rather than taking advantage of information hidden in the data. A typical healthcare cloud computing system has a hierarchical structure including system layer, control layer, and service layer. System layer constructs fundamental storage and computing environment using distributed computing resources, storage resources, and network resources. In control layer, system administrators control the load balancing, monitor system performance, and build programming environment. Finally service layer is responsible for providing large-scale healthcare services via real time management, privacy protection, and data analysis.

1.1 BIG DATA: The amount of data being generated inside and outside each enterprise has exploded. The increasing volume and detail of information, the rise of multimedia and social media, and the Internet of Things are expected to fuel continued exponential data growth for the foreseeable future.

1.2 RANDOM FOREST: Random forest is the supervised classification technique. It creates the forest of the various values lies into the data for prediction purpose. Later on make it random. The forest will generates better results as the number of values grows. That means when number of tree grows. More and more tree will leads to the betterment of the results. So we can say larger the number of trees in the forest and better will be the results. Random forest can be used for both classification and regression tasks. Over

fitting is one critical problem that may make the result worse.

1.2.1 Pseudo Code:

1. Randomly select the k number of features. The feature set consists of M Features. That means M Consist of super set of values {1…….n}. k is always less than the M (k<M)

2. Select the best split point from the set of points of k size. This point d is the super fit point around which left sub-tree and right sup-tree will be builded.

3. Split the whole dataset values into two parts. One is on the left side of the tree and other is on the right hand side of the tree.

4. Repeat the steps 1-3 till all the members lies into the super set will be converted to tree format.

5. repeat the whole procedure for n times till all member of M set will be converted to tree format.

1.3 GRADIENT BOOSTING MACHINE

Gradient Boosting machine is a learning based machine based on classification and regression type of problems. It produces a prediction model in the form of an ensemble of weak prediction model. In this the tree will be builded in stage wise manner. At each stage the weakness of the previous stage will be removed. At each stage gradient boosting performs the optimization of an arbitrary function. This function reduces the loss function.

The idea of gradient boosting originated in the observation by Leo Breiman. That Boosting can be interrelated as an optimization algorithm on a suitable cost function, explicit regression gradient boosting algorithms were subsequently developed by Joeome.

II. LITERATURE SURVEY

[1] A.R. PonPeriasamy and M .Chand-amona (2017): big-data revolution is under approach in health care and begin with the immensely enlarged offer of health data. Which push us to apply these new technologies to induce off their benefits and improve the medical sector. This paper can show the importance of applying predictive analytics techniques in medical platforms, and provides architecture design which mixes big data analysis, data-mining and also the mobile health care for self-monitoring. this method are going to be ready to exploit the attention data through an intelligent method analysis and big data processing.

[2] Peng Zhang(2016) et al: in this paper the research has been undertaken for creating cloud of the healthcare data. So that it can be integrated from different place. With the help of processing ability and cloud services different types of patient data be processed to generate prediction based scenario.

[3] Nima Jafari Navimipour(2016) et al: in current research Replica selection requires information about the capabilities and performance characteristics of a storage system. It is based on the user demand and failure occurs during response time. In data cloud, the selection of replica is an important issue for users and to access a data file. There research is mainly focused on replica selection mechanism in order to achieve the best performance. This research proposes new replica selection base on ant colony optimization to improve average access time.

[4] Abhinav Raj(2016) et al: A framework for secure sensitive data sharing on data mining platform, that includes secured delivery, usage, storage, and data destruction for semi-trusted data mining sharing platform. We propose a proxy re-encryption algorithm based heterogeneous cipher text transformation and user protection method based on virtual machine monitor, which provides for realization of system functions. The framework protects security of users sensitive data effectively and shares data safely. At same time, data owners retain complete control of their own data sound environment for modern Internet information security.

[5] Yurong Zhong(2016) et al: Spatial data is different from the general data, it not only contains some kind of property information of space feature, but also has the spatial feature of space or location. The spatial clustering analysis can be divided into two broad categories. In order to solve this kind of geographical position and property feature of double meaning, spatial data mining based on K - means spatial clustering algorithm combining geographic location and property feature, practice unified entity properties of spatial proximity and similarity. Considering the sharpening increasing scale of spatial data, the realization of the K - Means spatial clustering based on graphs parallel algorithm.

[6] Wei-Dong Zhu(2014) et al: All big data use cases require an integrated set of technologies to fully address the business pain they aim to alleviate. Due to this complexity, enterprises need to start small, with a single project, before moving on to other issues and pursuing added value. IBM is unique in having developed an enterprise-class big data platform that allows you to address the full spectrum of related business challenges.

III. PROPOSED SYSTEM

In current research Gradient Boosting Machine based system has been used for calculate the prediction for Future Length of Stay in the hospital. The existing Random Forest based technique was used. This random forest based technique sub divide the total list of dataset values into two parts based on the split point. This split point is the optimal point around which left sub tree and right sub tree stands.

Based on this process whole data values will be sub divided into various sub sets.

But Gradient Boosting based technique is optimization based technique. Where the stage wise reduction on loss function is generated. At each level the optimization of the prediction running from low level to the accurate level is generated. This type of prediction system is accurate because iteration of the enhancement of prediction and reduction in the loss function is performed in stage wise fashion.

IV. ALGORITHM

Input Training set {(xi,yi)}n,i=1, a differentiable loss function L(yF(x)), number of iterations m.

1. Initialize Model with constant value.

$F_o(x)=\arg_v min \sum_{i=1}^{n} L(yi,\gamma$

2. for m =1 to M:

i. Compute so called pseudo-residuals:

$T_{im}=\frac{\partial L(yi,F(xi))}{\partial F(xi)} f'(x) = Fm-1(x) for\ i=1,-----n$

ii. Fit a base learner(e.g. tree) Hm(x) to spesudo-residuals i.e. train it using training set $\{(x_i,r_{im})\}i=1$ to n

iii. Compute multiplier Ym by solving the following one dimensional optimization problem:

$Ym=argv\ min\sum_{i=1}^{n} L(yi,Fm-1(xi)+Yhm(xi))$

iv. update the model:

$F_m(x)=_{fm-1}(x)+Y_m h_m(x)$.

3. Output $F_m(x)$.

V. RESULTS AND ANALYSIS

The proposed work is based on prediction system for future length of stay in the hospital. Technique of Gradient Boosting Machine is used. It provides optimal solution. At each stage of the status the accuracy of the prediction is increased. It is compared to the based technique of Random Forest Based technique. Random forest on the contrary prepares the tree of the dataset value. Identifies the optimal value around which whole data set will be sub divided into smaller segments.

5.1 FLOS Prediction comparison in actual and Random Forest

Fig. 1 shows the comparison of prediction between the actual and the Random Forest based technique. There is a marked difference between the actual and the Random forest based technique.

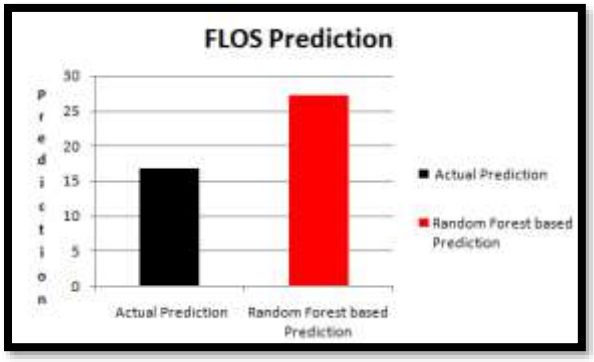

Fig. 1 Actual and R.F.

5.2 FLOS Prediction comparison in actual and Gradient Boosting based

Fig. 2 shows the prediction comparison between Actual and the Gradient Boosting based prediction. The difference between the actual and G.B. based technique has less difference.

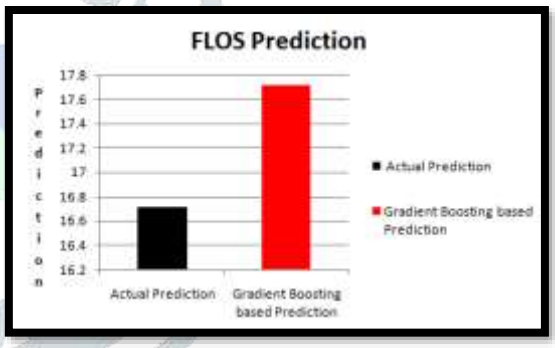

Fig. 2 Actual and G.B.

5.3 FLOS Prediction comparison in actual and Gradient Boosting based

Fig. 3 shows the prediction comparison between Random Forest and Gradient Boosting based technique. The is large difference between the prediction for both the techniques.

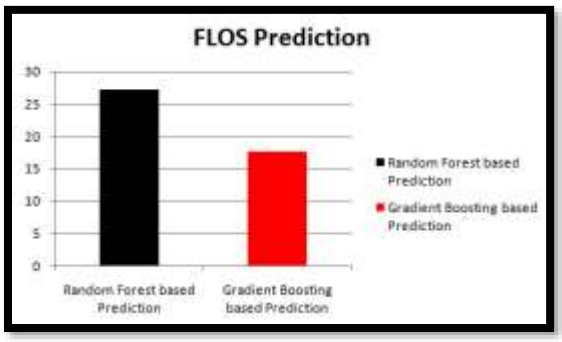

Fig. 3 R.F. and G.B.

5.4 FLOS Prediction comparison in actual and Gradient Boosting based and Random Forest

Fig. 4 shows the prediction comparison between the actual, Random forest and the Gradient Boosting based technique. There is very less difference between the Actual and the Gradient Boosting based technique.

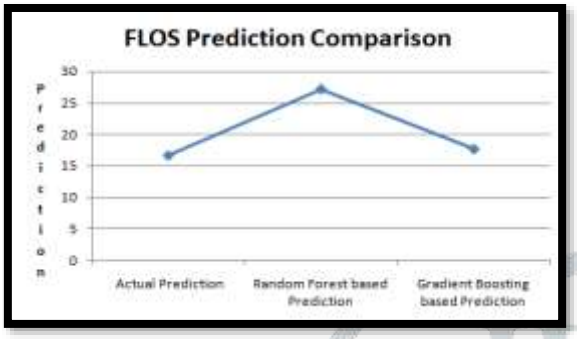

Fig. 4 Actual ,R.F. and G.B.

5.5 FLOS Prediction comparison in Gradient Boosting based and Random Forest

Fig. 5 shows the prediction comparison between the actual and the random forest based technique. There is a difference of around 12 base point.

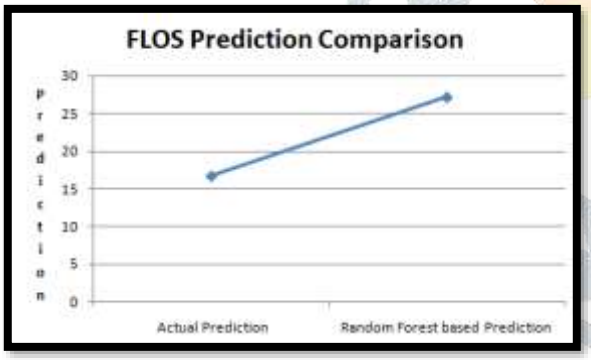

Fig. 5 R.F. and G.B.

5.6 FLOS Prediction comparison in Gradient Boosting based and Actual

Fig. 6 shows the FLOS comparison between the Actual and the Gradient based technique. There is a different of only two basis points.

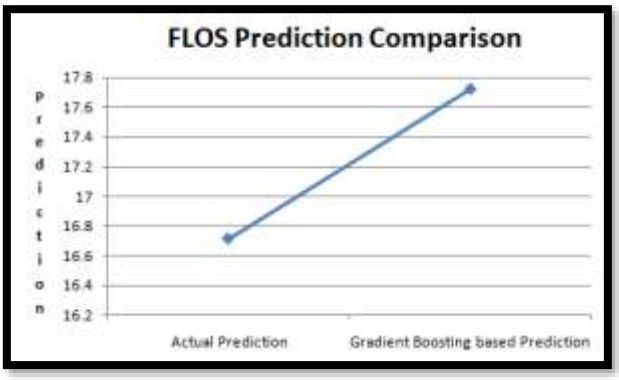

Fig. 6 Actual and G.B.

5.7 Comparison Table

Table 1 shows the prediction comparison between Actual , Random Forest based and the Gradient Boosting based.

Table 1. Prediction Comparison

Technique	Prediction
Actual Prediction	16.71
Random Forest based Prediction	27.19
Gradient Boosting based Prediction	17.72

Table 2 shows the Prediction percentage difference between actual and the random forest and the Gradient Boosting based technique.

Table 2. Percentage Difference

Technique	Percentage
Random Forest based Prediction	38.54%
Gradient Boosting based Prediction	5.68%

VI CONCLUSION

Processing the large data generated from various types of application is today's need. Because the analysis of these large data will give great facts. These analyzed facts can be used for various other efficient resource allocation process. Gradient Boosting based machine is best technique in comparison to the Random forest. It produces the prediction for future length of stay in the hospital based on various parameters values generated after the tests. The difference of prediction between the Actual and the Random forest is 38.54%. On the other hand the difference of the prediction accuracy between the Actual and the Gradient Boosting based technique is only 5.68%. that means the result of

prediction based on Gradient Boosting based technique is better than the Random forest.

VII. FUTURE WORK

In current research Gradient Boosting based technique for prediction for Future length of stay is evaluated. it is optimization based technique. It reduces the loss percentage at each step. Also it optimizes the results at each step. In future this work can be further improved by taking year wise data of the patients. So that the accuracy percentage can be enhanced further.

REFERENCES

[1] Peng Zhang, Shang Hu, Jing He, Yanchun Zhang, Guangyan Huang, Jiekui Zhang ," building cloud-based healthcare data mining Services", IEEE International Conference on Services Computing,Vol. 4,pp:90-98,2016.

[2] Nima Jafari Navimipour, Bahareh Alami Milani "Replica selection in the cloud environments using an ant colony algorithm" ,IEEE,Vol.8,pp:190-196,2016

[3] Abhinav Raj, A. Bender, N. Spring, B. Bhattacharjee and D. Starin, 2009. Persona: An online social network with user-defined privacy. Proceedings of the ACMSIGCOMM ,Vol. 4,pp:67-77,2016.

[4] Yurong Zhong, P.K. and J.H. Weber-Jahnke, Privacy preserving decision tree learning using unrealized data sets. IEEE,vol. 24,pp: 353-364. 2015.

[5] Wei-Dong Zhu, M. Piatek, A. Krishnamurthy and T. Anderson,Privacy-preserving P2P data sharing with One Swarm. Proceedings of the ACM,Vol.10, pp:111-122. 2015.

[6] Alberto Colorni and J. Pato, Preserving Privacy basedon semantic policy tools. Security Privacy IEEE,vol. 8: pp:25-30. 2014.

[7] E.K. Burke,privacy-preserving SVM classifier. Proceedings ofthe IEEE Transactions on Knowledge and DataEngineering, (TKDE' 11), IEEE Xplore Press, pp:1704-1717. DOI: 10.1109/TKDE.2010.193Lu, R., X. Lin and X. Shen, 2012. SPOC:

[8] Khaled Mahar,A secure andprivacy-preserving opportunistic computingframework for mobile-healthcare emergency,IEEE,Vol.3,pp:23-30,2014.

[9] Dipti Srinivasan, Marco Dorigo, Vittorio Maniezzo, "A Genetic Algorithm to Solve theTimetable Problem", Centre for Emergent Computing, Napier University, Edinburgh EH105DT, Vol.98,pp:876-890,2015.

[10] Swapna Borde, D.G.Elliman, R.F.Weare, "The Automation of the Timetabling Process in Higher Education",Vol.6,pp:900-910,2014

[11] Hana Rudova, TianHouSeow, JianXinXu, "Automated Timetable Generation Using Multiple Context Reasoning for University Modules",IEEE,Vol.23,pp:78-88, 2014.

[12] Ashish jain, Dr. Suresh Jain and DR. P.K. Chande," A New Approach to Generate Time Table", International Journal of Engineering Research and Applications,vol.4,pp: 2248-9622,2014.

[13] Yao-Te Wang and Keith Murray, "University Course Timetabling with Soft Constraints".IEEE,Vol.12,pp:78-88,2014.