Impact Factor (SJIF): 4.542

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 4, Issue 5, May-2017

A SURVEY ON PREDICTING THE WAITING TIME OF PATIENTS IN HOSPITALS

[1]Sanjitha S Laad, [2]Dr.Mohammed Rafi M.E P.hD, [3]Harshaverdana S BE, MTECH

[1]*P.G. STUDENT, Department of Studies in Computer science and engineering, UBDTCE, Davanagere*
[2]*ASSOCIATE PROFESSOR, Department of Studies in Computer science and engineering, UBDTCE, Davanagere*
[3]*ASSISTANT PROFESSOR , Department of Studies in Computer science and engineering, BIET, Hydrabad*

Abstract-Patients wait delay and patient overcrowding is one of the major challenges faced by hospitals. Waiting time increases the frustration on patients. Patient Queue Management and wait time prediction form challenging and complicated job because each patient might require different phases and operations such as check-up and various tests. Based on large scale, realistic dataset, the treatment time of each patient in the queue is predicted. We use an Apache Spark-based cloud implementationand Machine learning techniques to achieve the aforementioned goals.

Keywords- Apache spark, Big data, Machine learning, Hadoop, Random Forest(RF) algorithm., Cloud computing

I. INTRODUCTION

Most of the hospitals are overcrowded and they are inefficient in providing proper queue management. Providing Patient queue management and waiting time prediction is challenging as each patient vary in different operations such as checkup, different tests like X-ray, CT scan, sugar level. Some of the tasks are independent whereas some tasks are waiting to complete other dependent tasks. Most patients must have to wait in different queues for different treatments. In order to complete required treatment in a shortest duration of time waiting time of each task is predicted in real time. Different learning algorithm are proposed for calculating the waiting time, Patient Treatment Time Prediction(PTTP) uses RF algorithm[9] and Machine learning algorithms is used for predicting waiting time[1].

The massive unstructured data is called Big Data. Now in present days very less amount of data is generated in structured form as compare to unstructured data e.g. Text files, sensor data, web data, social networking data or different varieties of data. For Big Data management Hadoop is used. Hadoop is a framework that provides distributed processing of large data sets across cluster using a simple programming model. It is an open source data management which uses distributed processing.

II. DIFFERENT METHODS OF PREDICTING WAITING TIME

2.1. Predicting waiting time using Patient Treatment Time Prediction (PTTP) algorithm

2.1.1. Pre-Processing of data

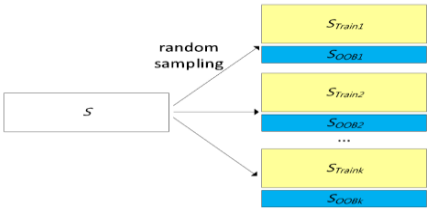
In the pre-processing phase, hospital treatment data from different treatment tasks are gathered such as the numbers of patients visit the hospital every day, patients who has been registered and his information and patient treatment tasks according to his health condition. Where each treatment task record can consist of multiple information for example, task name, task location, department, start time, end time, doctor, and attending staff.

2.1.2. Choose the same dimensions of the data

The data generated from different treatment tasks have varying dimensions. To train the patient time consumption model for each treatment task, choose the same features of these data, such as the patient information (patient card number, gender, age, etc.), the treatment task information (task name, department name, doctor name, etc.), and the time information (start time and end time). Other features of the treatment data are not chosen because they are not useful for the PTTP algorithm, such as patient name, telephone number, and address.

2.1.3. Training of dataset

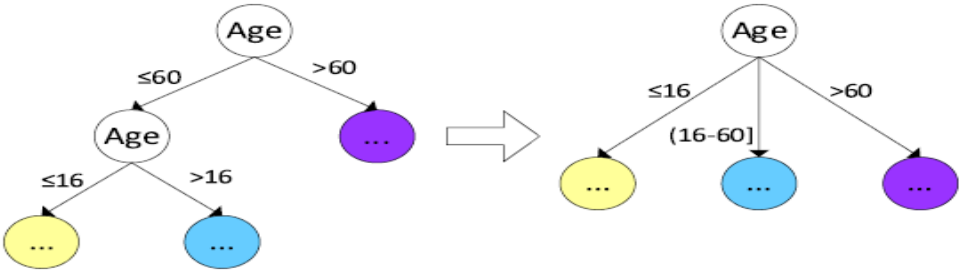
K training subsets are sampled from the original training dataset *S* in a bootstrap sampling process. *N* samples are selected from *S* by a random sampling and replacement method in each sampling period. After the current step, *k* training subsets are constructed as a collection of *Strain*.The unselected data in each sampling period are composed as an out-of-bag (OOB) data set[1].



"***Figure 1.Process of training dataset random sampling for the PTTP model***
"

2.1.4. PTTP model based on the RF algorithm

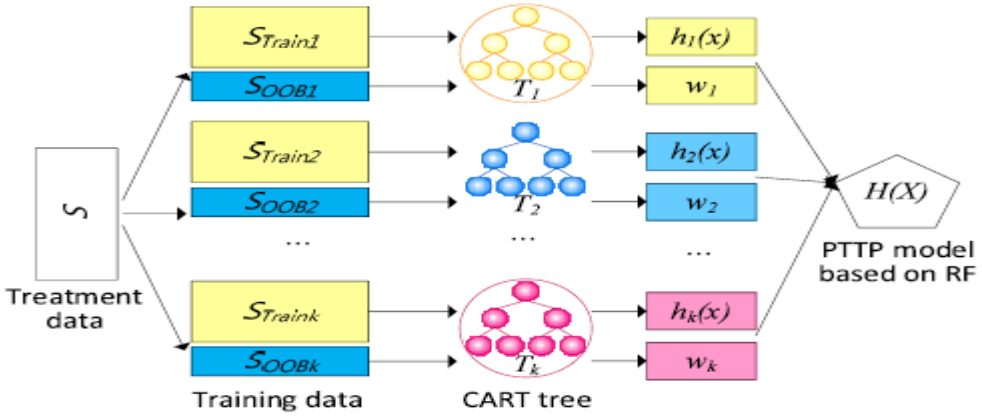
Although we have removed part of error data, other types of noise data might exist. Therefore, the third optimization aspect of the RF algorithm is to reduce the impudence that the noisy data that affect on the algorithm accuracy.Noise removal method is performed in the value calculation of each CART leaf node[4]. CART regression tree model is created for each training subset. CART model is a binary tree form, each root node represents a single input variable and a split point on that variable. The leaf nodes of the tree contain an output variable y which is used to make a prediction. CART tree is actually a partitioning of the input space, and each input variable as a dimension on an p-dimensional space. The decision tree split this up into forks, new data is filterd through the tree and the output value is predicted by the model[5].



"***Figure 2. Example of splitting of CART tree***"

2.1.5. Collecting *k* cart trees for a RF model

After the construction of the *k* CART regression trees, these trees are collected for a random forest model as shown in below fig.



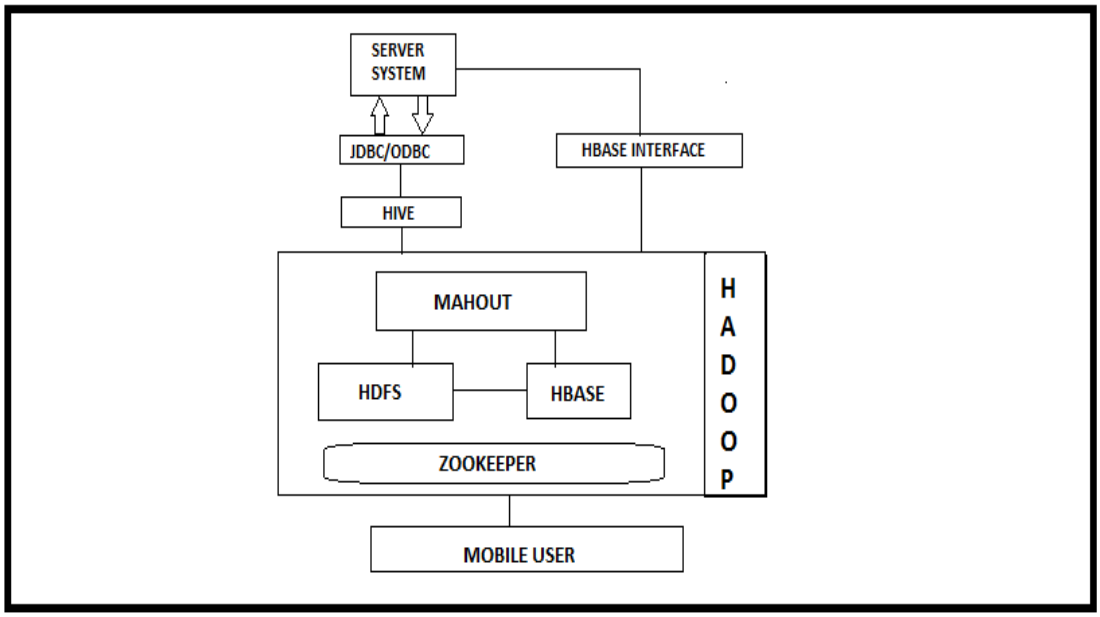
"***Figure 3. PTTP model based on RF algorithm***"

2.2. Predicting waiting time using RF algorithm

Hospital data generally stores the Structured And Unstructured Data. Most data used is Structured Data which includes information of a patient and information of treatments. This above data is stored in the hadoop cluster with the help of an JDBC/ODBC interface and then the data is stored in an HDFS with the help of an MAPREDUCE and the HIVE .The Structured data in the HDFS is written using an HIVE and its SQL like Query language HQL[2].

Hadoop is a framework which provides distributed processing of large data sets across cluster using a simple programming model. Mainly Apache Hadoop Framework consistsMapReduce and Hadoop distributed file system. Hadoop distributed file system, as Map reduce provide a simple programming model well as other related projects e.g. Apache Hive, Apache HBase etc.

There are various technologies belongs to HadoopHbase for storing large dataset, Apache pig is scripting language for processing of large data set, Hive is designed for OLAP is fast and Scalable, Scoop is used for import and export data from RDBMS to Hadoop, Zookeeper is used for distributed application and flume is for moving large amount of data to centralized data[2].

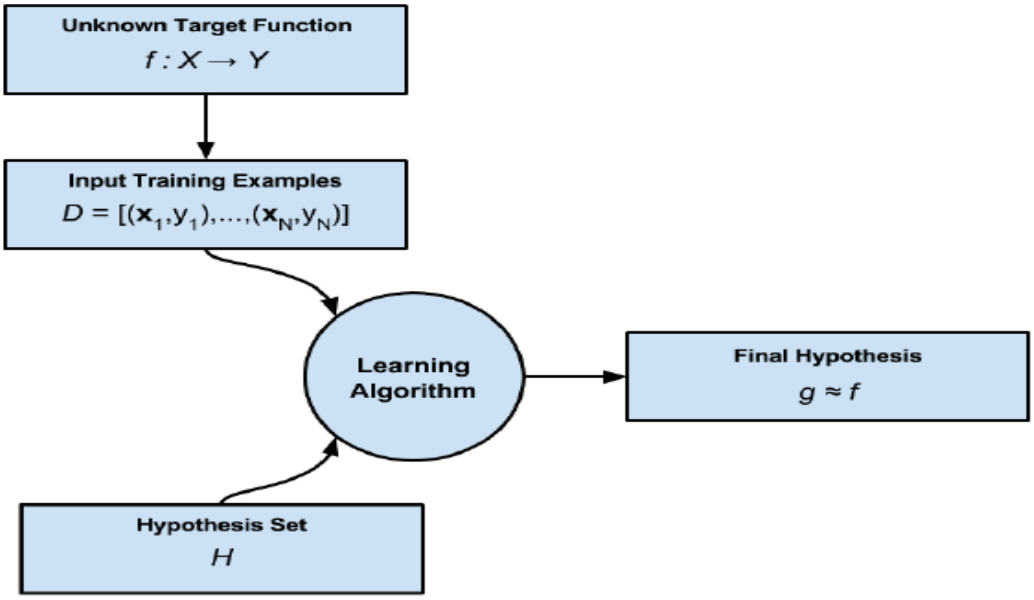


"***Figure 4. Diagrammatic representation of proposed system***"

Random forest is most popular data classification and regression algorithm. This system introduced a Scalable Random Forest Algorithm which is based on Map-Reduce Technique.The algorithm is divided into three stages: initializing, generating and voting. RF algorithm has main objective of improving the traditional random forest algorithm based on MapReduce model. RF algorithm provide scalable performance, and it can negotiate with the distributed computing environments to decide its trees scale.

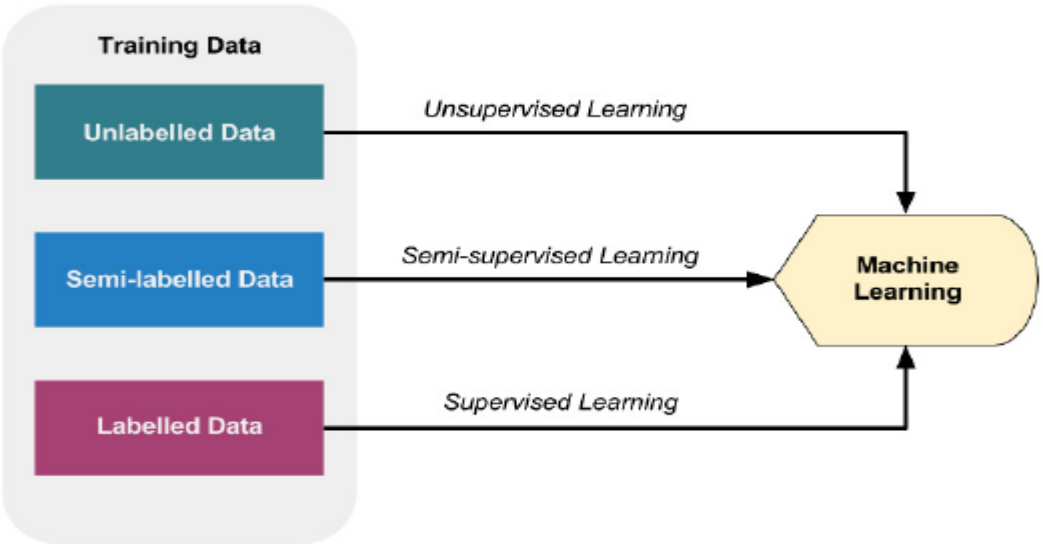
2.3. Predicting waiting time using Machine learning(ML) technique.

Using treatment appointment durations as an example, the main components of a machine learning problem are as follows. There exists an input vector, **x**, where each*xi* represents a unique "feature" that describes the output, *y* (for example, patient information that influences a patient's appointment duration), an unknown target function$f : X \rightarrow Y$, which corresponds to the ideal formulate predict the duration of an appointment, where *X* is the input space (theset of all possible inputs **x**), and *Y* is the output space (the set of all possible outputs *y*, in this case, a real-valued duration). There exists a dataset *D* ofinput-output training examples (**x**1, *y*1)……..(**x***N*, *yN*), where *yn*= *f*(**x***n*) for*n* = 1, . . . , *N* (inputs corresponding to previous patient data and their corresponding appointment durations known in hindsight). Finally, there existsa learning algorithm that uses the dataset *D* to pick a formula $g : X \rightarrow Y$that approximates *f*.



"***Figure 5.An overview of machine learning setup***"

The algorithm chooses *g* from a set of possible formulas, called the hypothesis set *H*. The algorithm chooses the *g* that best matches *f* on the training examples of previous patient data, with the expectation that it will closely match *f* on new patient data. When a new patient checks in for their appointment, their waiting time is inferred by applying machine learning to estimate the appointment durations of those patients ahead in the queue who have yet to be seen. The predicted individual durations of those appointments is based on *g* (the hypothesis that the learning algorithm produced), not on *f* (the ideal target function which remains unknown).Figure 6 illustrates the components ofthe learning setup. And classification of data can be done by using LR, SVM and Artificial Neural Networks[3].



"***Figure 6.Components of machine learning***"

III. ANALYSIS

Each model was fit to a training set and evaluation of its performance was done ona testing set. The training set was constructed by randomly sampling ofthe whole retrospective dataset, while the remaining was used for testing. Based on decision trees and combined with aggregation and bootstrap ideas, random forests, were powerful nonparametric statistical method allowing to consider regression problems as well as two-class and multi-class classification problems, in a single and versatile framework. On a practical point of view, RF are widely used and exhibit extremely high performance with only a few parameters to tune[9]. Since RF are based on the definition of several independent trees, it is thus straightforward to obtain a parallel and faster implementation of the RF method, in which many trees are built in parallel on different cores.

3.1. Accuracy

Accuracy of algorithm is influenced by a large volume of noisy data. As the proportion of noisy data increases, the accuracy of the algorithm decrease. PTTP algorithm and SVM can reduce the influence of noisy data effectively and achieve good robustness.

3.2. Performance

To evaluate the performance of the PTTP algorithm, groups of historical hospital treatment data are trained at different scales of the Spark cluster. The scale of slave nodes of the Spark cluster in each case increases, by observing the average execution time of the PTTP algorithm in each case, different performances across various cases are compared and analyzed. The advantage of the parallel algorithm in cases of large-scale data is greater than in cases of small-scale data[1]. The benefit is more obvious when the number of slave nodes increases. In case of machine learning which uses SVM for classification, speed and size is limited with respect to training and testing.

	PTTP model	**ML model**
Methods used	Random Forest	SVM, LR, ANN
Classification is by	CART	SVM
Accuracy	High	Low
Performance	High	Low

"***Table 1.Comparision between PTTP model and ML***"

IV. CONCLUSION

In this paper we have reviewed on the some of the techniques which are being used for hospital queuing management and the implementation of those different techniques. Present methods include PTTP algorithm, RF algorithm,Hadoop, Sql, Hbase, and other Machine learning algorithms. Later on we have compared the different techniques used by researchers in their systems, such as extension in RF algorithm, storing of structure data into the database etc. This comparison will help us in building our system more convenient and useful. In this paper, we focus on helping patients complete their treatment tasks in a predictable time and helping hospitals schedule each treatment task queue and avoid overcrowded and ineffective queues.

REFERENCES

[1]. JIANGUO CHEN[1,2], KENLI LI[1,2], ZHUO TANG[1,2], KASHIF BILAL[3,4], "A Parallel Patient Treatment Time Prediction Algorithm and Its Applications in Hospital Queuing-Recommendation in a Big Data Environment", Received March 8, 2016, accepted April 12, 2016, date of publication April 25, 2016, date of current version May 9, 2016

[2].ChetanShelake, SohamPatwardhan, SourabhTeli, KrantikumarMhetar, "Predicting the waiting time of patients in hospital by using RF algorithm and Design of HQR system",International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 03 Issue: 09 | Sep -2016.

[3].John Kang, MD, PhD,* Russell Schwartz, PhD,JohnFlickinger, MD, and SushilBeriwal, MD "Machine Learning Approaches for Predicting Radiation Therapy Outcomes-A Clinician's Perspective" International Journal of Radiation Oncology Received Mar 5, 2015, and in revised form Jul 21, 2015. Accepted for publication Jul 27, 2015.

[4]. G. Chrysos, P. Dagritzikos, I. Papaefstathiou, and A. Dollas, ``HC-CART: A parallel system implementation of data mining classi_cation and regression tree (CART) algorithm on a multi-FPGA system," *ACM Trans.Archit. Code Optim.*, vol. 9, no. 4, pp. 47:1_47:25, Jan. 2013.

[5]. Y. Ben-Haim and E. Tom-Tov, ``A streaming parallel decision tree algorithm," *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 849_872, Oct. 2010.

[6]. L. Breiman, ``Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5_32, Oct. 2001.

[7]. G. Yu, N. A. Goussies, J. Yuan, and Z. Liu, ``Fast action detection via discriminative random forest voting and top-K subvolume search," *IEEETrans. Multimedia*, vol. 13, no. 3, pp. 507_517, Jun. 2011.

[8]. C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, ``Robust and accurate shape model matching using random forest regression-voting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1862_1874, Sep. 2015.

[9] K. Singh, S. C. Guntuku, A. Thakur, and C. Hota, ``Big data analytics framework for peer-to-peer botnet detection using random forests," *Inf.Sci.*, vol. 278, pp. 488_497, Sep. 2014.

[10]. Jiawei Han, Yanheng Liu, Xin Sun," A Scalable Random Forest Algorithm Based on MapReduce",2013 IEEE