2021 6th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 02-04, 2021

Machine Learning Based Predictive Mechanism for Internet Bandwidth

Swapnil R. Pokharkar
Department of Computer Science and IT, Babasaheb Ambedkar Marathwada University,
Aurangabad, India
swapnil223469@gmail.com

Sanjeev J. Wagh
Department of Information Technology, Government College of Engineering,
Karad, India
sjwagh1@yahoo.co.in

Sachin N. Deshmukh
Department of Computer Science and IT, Babasaheb Ambedkar Marathwada University,
Aurangabad, India
sachin.csit@bamu.ac.in

***Abstract*- Internet of Things refers to the way that more and more physical devices are collecting and exchanging data over the internet. Internet of Things will have an increasing impact on bandwidth. Many Internet of Things devices operate wirelessly, while others are connected. Most IoT devices use less bandwidth, but many devices going online mean high bandwidth will be needed. As IoT grows, it will be necessary to have a platform which can accommodate this huge change. Due to the development of technology amount of data that is transmitted by devices is increased, which will need for increased bandwidth. For example, when smartphones start transmitting images and streaming video, need for bandwidth increases tremendously. There is no particular solution available for spectrum predictions. In this paper, we propose a machine learning prediction algorithm for internet bandwidth.**

Keywords: Internet Bandwidth, Location Area Code, Machine Learning, Network Traffic

I. INTRODUCTION

Internet bandwidth has tremendous impact on network performance. Network performance can be measured with different prediction models. In present time network traffic prediction is a research area. Network bandwidth also impact on economy of any nation. In future there is a demand for device to device communication, IOT indicates requirement of bandwidth. In future internet bandwidth will provide big market up to 2025.

To predict internet traffic, we used India's well-known network as dataset. Dataset mainly focuses on location area code, total data in G.B, total uploaded data, total downloaded data. Experimental result provides us actual and predicted data on location area code using different prediction algorithm. This paper focuses on prediction of future bandwidth with maximum accuracy [1] [2].

In order to achieve this objective, we use different machine learning algorithms to predict internet bandwidth [3].

II. BACKGROUND WORK

In paper "Bandwidth comparison model" we tried to analyze network traffic [4]. We used India's well known network data as a dataset for prediction model [5]. As per dataset we analyze uploaded and downloaded data on each location area code for Pune region. We performed comparison of data uploaded and downloaded in 2018 and 2020 [4].

III. MACHINE LEARNING ALGORITHM USED FOR PREDICTION

A. Random Forest

Random forest builds number of decision trees and combines them to get a more precise and stable prediction [6]. Random forest is a bagging technique which intends to decrease complexity of models [7]. The trees in random forests are run in parallel. There is no communication between these trees at the time of building trees.

Random forest select random sample from dataset. To get prediction result from every decision tree it will construct decision tree for every sample. Then, voting performed for every predicted outcome. Finally, the most voted prediction result is selected as final prediction outcome.

B. Decision Tree

Decision tree is used to predict data of future to produce meaningful continuous result [8]. A decision tree asks series of questions to the data for arriving at final estimate. Each question narrowing our predictable values until the model gets sufficiently sure to make a solitary prediction. Model decides the order of question as well as their content. All questions are in the form of True or false. During decision tree algorithm training, the model is fitted with any historical data which is relevant to problem statement. The model learns any connection between actual data and target variable.

To decide split of node in two or more sub-nodes decision trees regression use mean squared error (MSE) [9].

Mean Square error is calculated by using formula

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

The smaller mean squared error, the closer you are to finding the line of best fit.

C. K Nearest Neighbor

To predict values of any new data points KNN algorithm [10] uses feature similarity. It means new point is assigned a value based on how closely it resembles point in the training set.

2021 6th International Conference for Convergence in Technology (I2CT) | 978-1-7281-8876-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/I2CT51068.2021.9418164

In KNN, Distance between new point and each training point is calculated. After calculating distance, closest k data points are selected. Euclidean Distance [11] and Manhattan Distance [12] are two different methods to calculate distance.

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

$$\text{Manhattan} = \sum_{i=1}^{k} |x_i - y_i|$$

Final prediction for new point is the average of these data points. In order to achieve better accuracy, important question is that variable supports prediction of future bandwidth [13]. We have selected input parameter as date, LAC and output parameter as Total Data. We analyse availability of these variables which helps to forecast the future bandwidth.

IV. EXPERIMENTAL RESULTS

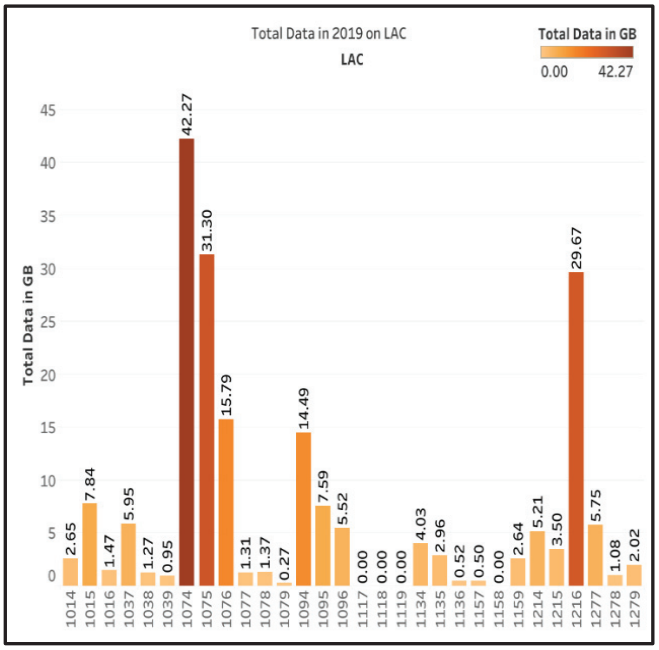

Fig. 1. Total Data in 2019 on LAC

Fig.1 shows the bandwidth used in 2019. X-axis represents LAC and y-axis represents Total data in GB. We can understand the usage of data on different LAC, with the help of variation in bins and color. Total data used on each LAC indicated by number above the bins.

Fig.2 shows the bandwidth used in 2020. X-axis represents LAC and Y-axis represents Total data in GB. We can understand the usage of data on different LAC, with the help of variation in bins and color. Total data used on each LAC indicated by number above the bins.

We have compared both graphs, and then concluded usage of data in 2019 and 2020 is varied significantly.

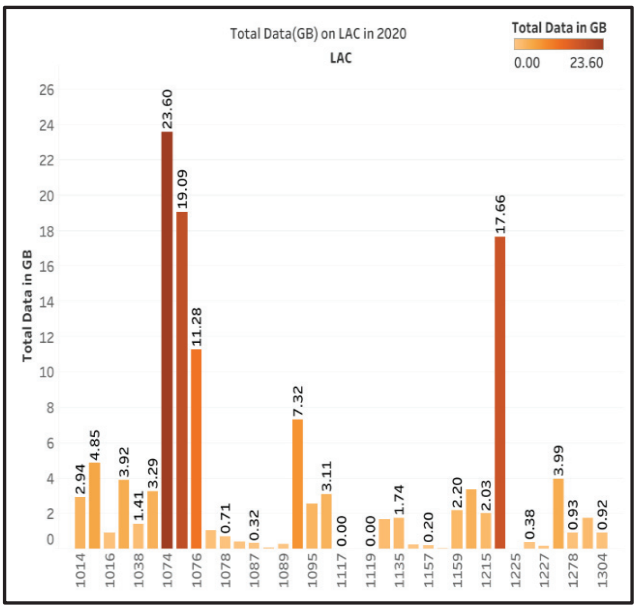

Fig. 2. Total Data on LAC in 2020

A. Prediction Using Decision Tree Algorithm

Step 1: Import Numpy, Pandas and matplotlib libraries.

Step 2: Import the dataset using pd.read_csv() method

Step 3: Pre-processing of dataset, data cleaning if necessary

Step 4: Select all input parameters from dataset as x and all output parameter as y

Step 5: Apply train_test_split in order to create training and testing parameters. We are using test_size = 1/3

Step 6: Train the decision tree regression model on training set

Step 7: Predict values using predict () method

Step 8: Compare real values with predicted values

Step 9: Visualize decision tree prediction

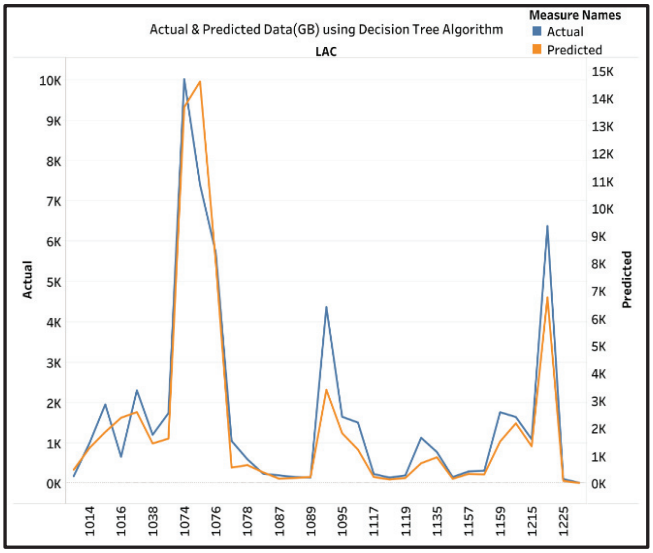

Fig. 3. Actual and Predicted Data on LAC Using Decision Tree

Fig.3 shows comparison of actual and predicted data on each LAC. In fig.3 Blue line indicates Actual data while orange line indicates predicted data. Prediction is carried out by using Decision Tree algorithm. With reference to above graph, we analyze accuracy of Decision Tree algorithm in our problem. If actual and predicted lines match then accuracy is higher. But in our graph actual and predicted lines somewhat differ from each other.

B. Prediction Using K Nearest Neighbor Algorithm

Step 1: Import required libraries and dataset.

Step 2: Pre-processing of dataset and data cleaning if necessary

Step 3: Select all input parameters from dataset as x and all output parameter as y

Step 4: From training set select random K data points.

Step 5: With the help of selected data points build decision trees.

Step 6: Select the number N for decision trees that you want to build.

Step 7: Repeat Step no. 4 & 5.

Step 8**:** Find the predictions of each decision tree, to achieve new data points. Assign new data points to category that wins majority votes.

Step 9: Compare real values with predicted values

Step 10: Visualize decision tree prediction

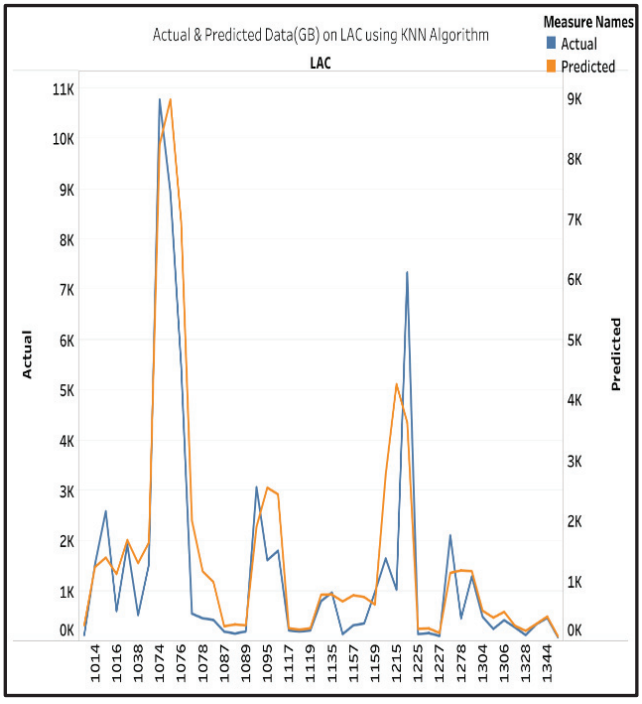

Fig. 4. Actual and Predicted Data on LAC using KNN

Fig .4 shows comparison of actual and predicted data on each LAC. In fig.4 Blue line indicates Actual data while orange line indicates predicted data. Prediction is carried out by using K Nearest Neighbor algorithm. With reference to above graph, we can conclude relations between actual and predicted lines are not much similar. So, accuracy provided by this algorithm is moderate.

C. Prediction Using Random Forest Algorithm:

Step 1: Import required libraries.

Step 2: Import and print the dataset

Step 3: Pre-processing of dataset and data cleaning if necessary

Step 4: Select all input parameters from dataset as x and all output parameter as y

Step 5: Apply train_test_split in order to create training and testing parameters.

Step 6: Create random Forest regressor and fit to dataset

Step 7: Predict values using predict () method

Step 8: Check Accuracy of prediction

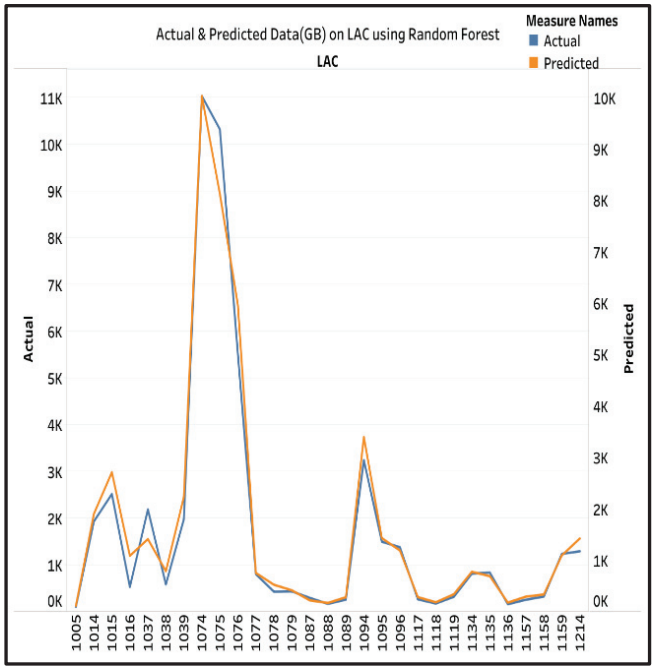

Fig. 5. Actual and Predicted Data on LAC using Random Forest

Fig.5 shows direct comparison of actual and predicted data on each LAC. Blue line indicates Actual data while orange line indicates predicted data. Prediction is carried out by using Random Forest algorithm. We analyze accuracy of random Forest algorithm in our problem. As compared to other prediction algorithm random forest provides more accuracy. Random Forest model is able to provide relationships between actual and predicted data. Random Forest model gives more accurate forecasts of future bandwidth.

We used three machine learning algorithms to predict bandwidth requirement. From these three algorithms random forest provides highest accuracy in prediction.

V. LIMITATION AND FUTURE WORK

Nevertheless, too reliably prediction of the network bandwidth requirement is a complex task and still considered an open challenge. In our future research, we plan to optimize the bandwidth prediction with machine learning evaluation to simulate different use cases. Hereby, we intend to estimate the real advantage of our model. In future work,

we will use different machine learning algorithm for better prediction**.**

VI. CONCLUSION

In this paper, we proposed a machine learning based prediction approach for a future bandwidth prediction. We visualized the result of prediction using different graphs. We can conclude that random forest algorithm gives better performance in terms of accuracy. Random forest algorithm provides approximately 82% accuracy, which is better than other algorithms. This contribution can uphold situation where limited bandwidth needs to be predicted.

REFERENCES

[1] Muhammad Usman "A Bandwidth Friendly Architecture for Cloud Gaming" 978-1-5090-5124-3/17/$31.00 ©2017 IEEE

[2] Muhammad Faisal Iqbal "Efficient Prediction of Network Traffic for Real-Time Applications" Volume 2019, Article ID 4067135

[3] Jaiswal Rupesh Chandrakant "Machine Learning Based Internet Traffic Recognition with Statistical Approach" 978-1-4799-2275-8/13/$31.00 ©2013 IEEE

[4] Swapnil R. Pokharkar, Sanjeev J. Wagh, Sachin N. Deshmukh "Bandwidth Comparison Model for Future Internet Using Machine Learning" International Journal of Future Generation Communication and Networking Vol. 13, No. 3, (2020), pp. 1249–1257

[5] Mahanagar Doorsanchar Bhawan "The Indian Telecom Services Performance Indicators" Telecom Regulatory Authority of India

[6] Yu Liu, Lu Liu, Yin Gao, Liu Yang "An Improved Random Forest Algorithm Based on Attribute Compatibility" 978-1-5386-6243-4/19/$31.00 ©2019 IEEE

[7] Amir Saffari, Christian Leistner, Jakob Santner, Martin Godec, Horst Bischof "On-line Random Forests" 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 978-1-4244-4441-0/09/$25.00 ©2009 IEEE

[8] Ahmed Mohamed Ahmed, Ahmet Rizaner "A Decision Tree Algorithm Combined with Linear Regression for Data Classification" 978-1-5386-4123-1/18/$31.00 ©2018 IEEE

[9] William C. Ogle, Hanna E. Witzgall, Michael A. Tinston, J. Scott Goldstein "Independent Sample Mean Squared Error for Adaptive Detection Statistics" O-7803-8870-41051$20.OO©g2005 IEEE

[10] Aiman Moldagulova, Rosnafisah Bte. Sulaiman "Using KNN Algorithm for Classification of Textual Documents" 978-1-5090-6332-1/17/$31.00 ©2017 IEEE

[11] Chin-Chen Chang, Jer-Sheng Chou, and Tung-Shou Chen "An Efficient Computation of Euclidean Distances Using Approximated Look-Up Table" 1051–8215/00$10.00 © 2000 IEEE

[12] Mrs.M. D. Malkauthekar "Analysis of Euclidean Distance and Manhattan Distance Measure in Face Recognition" 978-1-84919-859-2

[13] Abdelnaser, Mohammad Adas "Using Adaptive Linear Prediction to Support Real-Time VBR Video Under RCBR Network Service Model" 1063–6692/98$10.00 □ 1998 IEEE