International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST207536

Big Data Analytics For 6G

Aarushi Sharma*, Rahul Aggarwal, H. Srikanth Kamath
Department of Electronics and Communications, Manipal Institute of Technology, Manipal, Karnataka, India

Article Info
Volume 7, Issue 5
Page Number: 158-164
Publication Issue :
September-October-2020
Article History
Accepted : 20 Sep 2020
Published : 30 Sep 2020

ABSTRACT

The scope of this paper is to analyse traffic demands for 6G communication systems. Since it is important to maintain the sustainability and competitiveness of the communication system, we need to invest in researching about what 6G would be like. We have taken an important possible application of 6G which is Data Analytics. We will be performing Predictive Analysis to predict the traffic patterns based on the traffic of each cell at any given time of the day.

Keywords : Data analytics, 6G traffic, Intelligent Network, Machine learning

I. INTRODUCTION

Future cellular communications will require higher data rates and a much more reliable transmission link to stay in pace with the growth of multimedia services, while also maintaining decent quality of service. In the past two years, some countries have released relevant research plans concerning the development of 6G.

Around 2030, our society will likely become data-driven, enabled by nearly instantaneous, unlimited wireless connectivity [1]. As a vision for the future, in terms of speed, 6G will probably utilise higher frequency spectrum than previous generations in order to improve the data rate expected to be 100 to 1000 times faster than that of 5G [2].

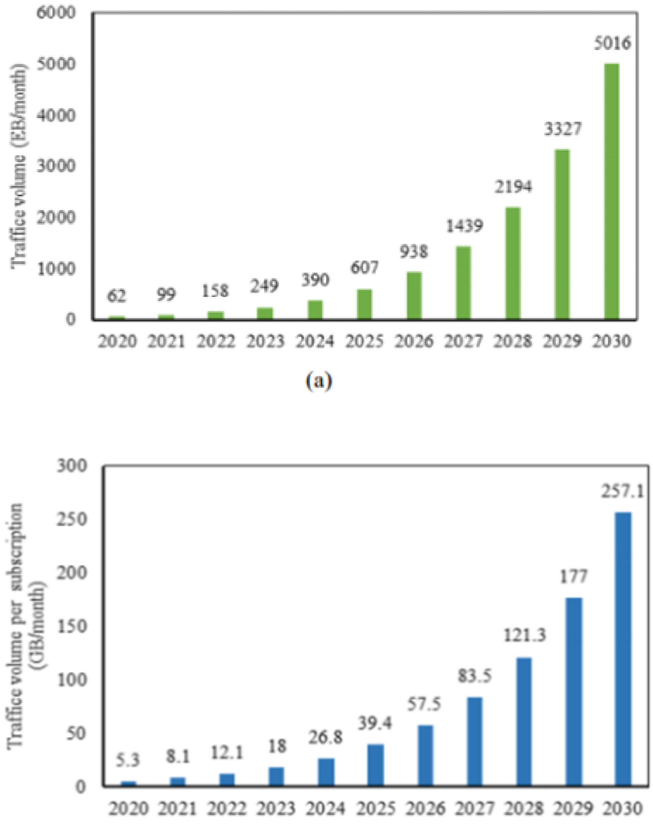
Understanding the dynamic of the traffic demands in a wireless network represents a complex task, due to the massive densification of the mobile devices attached to the network. Since the first natural application of AI is big data analytics, our objective is to use Predictive analytics on 6G and use the data to predict future events such as traffic patterns. We should be able to predict whether the traffic of a particular cell will be high or low at any given time.

In the following sections we will analyse each of the steps taken and discuss the results obtained from the same.

II. TRENDS IN TRAFFIC

We jump from one generation of communication system to the other after every 10 years. As we move on to the next generation, this jump helps in the improvement of Quality of service and also includes new features. Due to the introduction of new services such as mobile to mobile communications and use of smart phones, the mobile traffic has increased a lot compared to the previous generations and this same trend is said to be followed for the future generations as well. Fig [1] depicts the exponential growth of data traffic.

It is expected that the global mobile traffic volume will increase 670 times in 2030 as compared with the mobile traffic in 2010 [3]. The traffic volume for each of the mobile devices will also increase. The traffic volume of a mobile device in 2010 was 5.3 GB per month. However, this volume will increase 50 times in 2030. The number of M2M subscription will increase 33 times in 2020 and 455 times in 2030, as compared with 2010.



(a)

(b)

Figure 1: (a) Total global traffic volume. (b) Traffic volume per subscription [3]

III. EXPLORATORY ANALYSIS

The process of Data Analysis includes examining of raw data and deriving information and conclusions. The process involves several steps starting from collection of right data. There are many sites online which provide data sets to work on. The next step in the analysis is processing of raw data. It is just a sequence of operation performed to convert raw data into usable form of data.

Data set used contains the information about the traffic of 4G cells. 4G cell traffic is understood as: When a user uses a mobile data service, the mobile device will be served by a nearby 4G cell [3]. The total data capacity of all users served by a cell within an hour is called the traffic of that cell within 1 hour. Example: Cell 039872 is serving 50 subscribers, each subscriber in 1 hour x uses an average of 10Mb => Traffic of cell 039872. So the traffic of this cell in hours x = 50 * 10 = 500Mb. Data set has: 57 cells Data is collected in approximately 1 year x 24 hours x 57 cells [3].

The data we received was already pre-processed so we were able to skip this step. The next important step was performing Exploratory Analysis. It is an approach for analyzing data sets to summarize their main characteristics. The main objective of performing this analysis is for seeing what the data can tell us beyond the formal modeling or hypothesis testing.

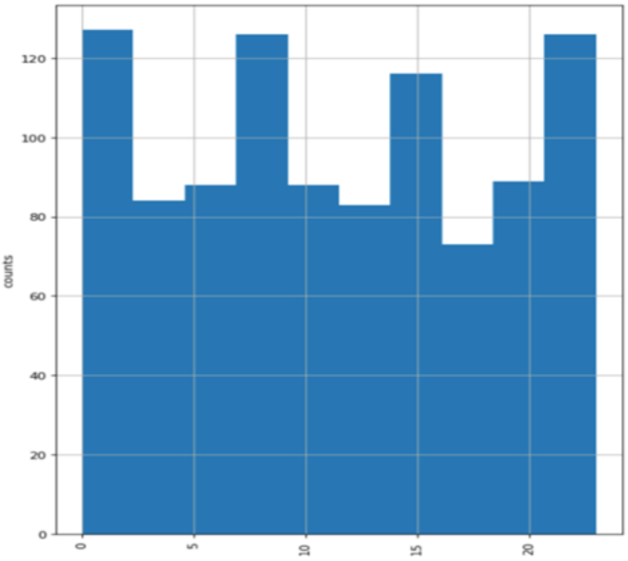


Figure 2: Histogram showing count of number of samples values in a given interval

Figure 2 and Figure 3 (next page) show the results obtained after performing exploratory analysis on my data. This analysis is best performed using visual methods. My data set includes 4 columns – Date/Time, Cell name, Hour and Traffic. Figure 2 is a histogram for traffic values. The X-axis has traffic values and Y-axis is the number of samples. The plot shows the count of number of sample values in a traffic range. For example, from the plot we can see that there are around 127 sample values whose traffic values are between 0 and 2.

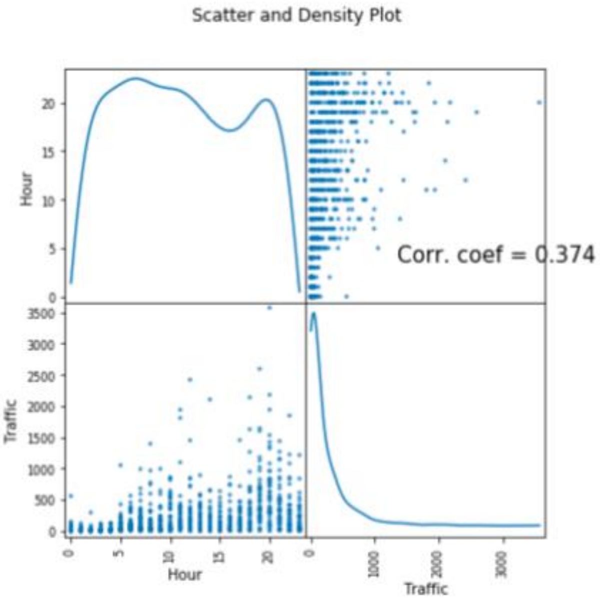


Figure 3: Scatter and Density plots for two features – Traffic and Hour

In the second plot (Fig. 3) we have plotted a scatter plot and a density plot for both Hour and Traffic feature. The first scatter plot in the first row has traffic values on the x-axis and hour values on y-axis. Every sample has a (traffic, hour) pair which is plotted as points. The plot directly below this scatter plot is a density plot which is only for the traffic values. Same as histogram, this plot shows the count of sample values over a range of traffic values. In other words the density plot is a smooth curve that fits the histogram. Similarly a density plot for our feature Hour is plotted.

We have also calculated correlation coefficient between hour and traffic. This provides with a numerical measure of the correlation or a statistical relationship between the two quantities. As shown on the plot itself, the correlation coefficient obtained was 0.374. From this we infer that there is some positive correlation between hour and traffic. Since the coefficient is much less than 1 we can say that it is not even close to a perfect correlation. The two quantities are related and dependent to some extent.

The purpose of this analysis is to look for any missing data and other mistakes. It allows us to gain maximum insight into the data set and its underlying structure.

IV. MACHINE LEARNNING MODEL

After performing exploratory analysis and uncovering the parsimonious model, one which explains the data with a minimum number of predictor values, we have to choose an appropriate machine learning model for training our data set. There are two broad categories of machine learning – supervised and unsupervised learning. Supervised learning requires that the algorithm's possible outcomes are already known and the data that is used to train the algorithm is already correctly labelled. Unsupervised learning is used when the train set is unlabelled.

Since supervised learning algorithm is much simpler and accurate, we decided to work with it. Further, there are different types of supervised learning algorithms. The most important task before choosing the algorithm is to define the problem statement clearly. The main aim is to be able to predict whether the cellular traffic would be high or low during any given time at any given day. We need to train our data in such a way that we are able to classify the traffic values into three categories – High, medium or low traffic.

Out of various types of supervised learning methods, we need to follow a type of classification/prediction algorithm. As we know that supervised learning only works with labelled data, our first step will be to label our data. In order to do so we have to add a fifth column named 'Label' in our working data set. We fill it values for each sample based on the traffic values.

We used if – else condition for doing the same. If traffic value of a sample is between 0 and 3500 label 0 will be added for that row. If the traffic value is between 3501 and 7500, label 1 will be assigned and if the traffic value is greater than 7500 label 2 is assigned. Here label 0 will denote low traffic, label 1 will indicate medium traffic and label 2 will denote high traffic.

Decision Tree Algorithm is a widely used classification algorithm. We have used our features to build the decision tree. Initially we consider the entire data set as our root node. In order to choose the feature for making the split, we make use of something called gini index. Gini index or gini impurity measures the probability of a random variable being wrongly classified when chosen. Since we have 4 variables we find gini index for each of them and choose the variable with the least gini index. After choosing the variable we perform our split. We continue to split after calculating gini index for variables at each level until we have performed the classification. The last level represents the outcome and the leaf nodes represent the class labels or decisions taken after computing all attributes.

All the steps mentioned above are all performed using a single statement. The python library sklearn.tree included all the necessary functions that are required to build the tree correctly.

V. TRAINING

After completing the design for the decision tree we need to run the entire algorithm for our training set. We need to check whether our algorithm is working for different input variables as well or not. For this we need to determine certain control parameters. These parameters may be adjusted by optimising performance on a subset or a validation set. We have obtained this validation set by performing K-Fold Cross validation.

In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data [5]. The k results can then be averaged to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10- fold cross-validation is commonly used, but in general k remains an unfixed parameter [5].

We specifically made use of Stratified K-fold cross validation. In stratified k-fold, the partitions are selected so that the mean response value is approximately equal in all the partitions. This means that each partition contains roughly the same proportions of each type of class labels.

The goal of cross-validation is to estimate the expected level of fit of a model to a data set that is independent of the data that were used to train the model. It can be used to estimate any quantitative measure of fit that is appropriate for the data and model. The idea of building machine learning models works on a constructive feedback principle. You build a model, get feedback from metrics, make improvements and continue until you achieve a desirable accuracy. Evaluation metrics explain the performance of a model.

An important aspect of evaluation metrics is their capability to discriminate among model results. So we choose Precision and Recall Metrics. Precision and recall are two extremely important model evaluation metrics. Precision refers to the percentage of our results obtained which are relevant and Recall refers to the percentage of total relevant results correctly classified by our algorithm. We have also made use of another performance metric, Accuracy. It is the most intuitive performance measure and it is simply a ratio of correctly predicted observations to total observations.

VI. RESULT ANALYSIS

Timely and accurate prediction of traffic flow is essential for proactive traffic management. If a cell is experiencing more traffic at a certain point in a day, the connection will be switched to the nearest neighbor.

This planning and resource allocation must be done in advance to ensure seamless connectivity. We have prepared a model which can predict in which range the traffic falls into, based on the time of the day and cell, and enables us to take steps accordingly. We have considered different cases based on the K-split validation method and decision tree classifier. We have considered different values of K and also changed the criteria of the spilt and compared the prediction and accuracy scores for each.

Stratified K Fold is a cross validation method which divides the data set into K samples. It will consider one sample as test set and remaining k-1 samples as training set. This is performed k times such that each sample is taken as a test set at least once. In this case, we specified the number of splits as 5 and we also chose the Gini Index as the criteria for selecting features for the decision tree. With these specification we got the following result:

Precision% = 81.902%

Recall% = 81.639%

Accuracy% = 81.639%

In the next case, we specified the number of splits as 20 and we also chose the Gini Index as the criteria for selecting features for the decision tree. With these specification we got the following result:

Precision% = 82.223%

Recall% = 81.941%

Accuracy% = 81.941%

Next, we specified the number of splits as 20 and we also chose the Information Gain as the criteria for selecting features for the decision tree. With these specification we got the following result:

Precision% = 82.062%

Recall% = 81.771%

Accuracy% = 81.771%

In this case, we specified the number of splits as 20 and we also chose the Information Gain as the criteria for selecting features for the decision tree. In this case we also changed the random states to 4 in the function just to see how he result varies. With these specification we got the following result:

Precision% = 82.147%

Recall% = 81.872%

Accuracy% = 81.872%

Then we specified the number of splits as 20 and we also chose the Gini Index as the criteria for selecting features for the decision tree. In this case we also changed the random state to 10 in the function just to see how he result varies. With these specification we got the following result:

Precision% = 82.223%

Recall% = 81.909%

Accuracy% = 81.940%

We got the maximum precision using the following specifications. We specified the number of splits as 20 and we also chose the Gini Index as the criteria for selecting features for the decision tree. In this case we also changed the random state to 70 in the function just to see how he result varies. With these specification we got the following result:

Precision% = 82.257%

Recall% = 81.970%

Accuracy% = 81.970%

After changing various different parameters and recording the results we were able to make a prediction model whose Precision score is **82.257% and give 81.970% accuracy**. This model can now be used to predict the range of traffic values in each cell at a given time of the day.

VII.CONCLUSION

The objective of this paper was to analyze the traffic patterns and be aware of the traffic demands in advance. In 6G wireless communication network the densification of mobile devices in a network will tend to increase. To meet the strict requirements it is fundamental that the network becomes aware of the traffic demands. The analysis of the traffic and the precise forecast of the user demands are essential for developing an intelligent network. Knowing in advance the user demands makes the network able to promptly manage the resource allocation among the contending users.

Through this project we were to build an algorithm which classifies the traffic as high, medium or low. With the help of trained data we were able to give cell name and just the time of the day as input and our algorithm would tell us the traffic pattern so that we can address the demands beforehand. High data rate, spectral efficiency and reliable wireless communication can be achieved in 6G communication strategies. Therefore, it is better to be fully prepared for all the challenges that we may face for smooth deployment of 6G.

VIII. REFERENCES

1. K. David and H. Berndt, "6G Vision and Requirement," IEEE Vehic. Teh. Mag., vol. 13, no. 3, Sept. 2018, pp. 72– 80.
2. J. G. Andrews et al., "What Will 5G Be?," IEEE JSAC, vol. 32, no. 6, June 2014, pp. 1065–82.
3. "Beyond 5G: The Roadmap to 6G and beyond", Online]. Available: https://www.cablefree.net/wireless-technology4g-lte-beyond-5g-roadmap-6g-beyond/, Posted Jul 4,2017 Accessed on - March 18, 2020
4. "Predicting Traffic for 4G LTE network", Online]. Available: https://www.kaggle.com/naebolo/predict-traffic-of-lte-network, Posted Nov 20, 2018 Accessed on - March 18, 2020
5. "A gentle introduction to cross validation", Online]. Available:https://machinelearningmastery.com/k-fold- cross-validation/Posted May 23, 2018 Accessed on - May 14, 2020

Cite this article as :

Aarushi Sharma, Rahul Aggarwal, H. Srikanth Kamath, "Big Data Analytics For 6G", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 7 Issue 5, pp. 158-164, September-October 2020. Available at doi : https://doi.org/10.32628/IJSRST207536
Journal URL : http://ijsrst.com/IJSRST207536