Prediction of future throughput for cellular networks with radio KPIs

Arth Vyas
x17170516
MSc in Data Analytic

8th April 2019

Abstract

Across the world, there is rapid growth in the development of Cellular Network and with the advancement of the network more users are getting connected to the cellular network day by day. Users expectation has been raise due to the deployment of 4G LTE and upcoming 5G. This is because of the capability of the network to serve better due to its robust design by which mobile network operator promise the quality of service (QoS) to users. The common issue faced by users in a network is the poor quality of experience (QoE) though high-speed network provided by 4G LTE/5G. The issues like low throughput, network coverage, cell congestion, resource allocation, online gaming performance, and applications which are providing video streaming. If the behavior of throughput over time is analyzed than it can give insight to the mobile network operator to address the above issues. A reliable accurate throughput prediction for the cellular network can offer a better solution to the application uses adaptive behavior for streaming video, VoIP, and embedded logic by which provide better QoE. Estimated future throughput can help in adaptive bitrate (ABR) algorithm for video streaming, network decongestion plan, and resource allocation for better network management and QoE. This research will emphasize on building different probabilistic methods to predict future throughput to its best. The techniques that will be used here for achieving future throughput are ARIMA, ARIMAX, Random Forest, SVM, and SVR. Throughput is highly related and can fluctuate to a certain extent due to the impact of radio metric KPIs. The radio KPIs will be of the important factor to predict precise future throughput prediction, to consider these KPIs under the research will be the prime aim. Finally, the results interpretation will be addressing all video applications and network management by which important decision like ABR selection in the algorithm, and decongestion strategies and other important decision of network management.

Keywords:Throughput, Adaptive bitrate (ABR), Cell congestion, resource allocation, Quality of service (QoS), Quality of (QoE), Key performance Indicator (KPIs) 4G LTE/5G, ARIMA, ARIMAX, Random Forest, SVM, SVR.

Contents

1 Introduction

Telecommunication sector future seems to provide better quality of services since it is developing year by year. According to GSMA a huge amount has been invested primarily in the deployment of 4G Long term evolution (LTE) technology approximately $1.3 trillion has been invested and furthermore to be invested for the deployment of 5G in coming year. A more beneficial performance is assured because of new wireless spectrum, and deeper fiber backhaul. More cell sites need to be developed to deliver the benefits of 5G. The investment rise in this technology can be seen to ramp from 2020 and beyond. The issues like poor coverage, bad interconnectivity, flexibility, and quality of service was observed in 4G LTE for which all this issue will be address and improvement will be done on upcoming deployment of 5G. The base data bandwidth proposed by 4G LTE is approximately 2Mbps to 1Gbps and frequency band ranges from 2 to 8 GHz whereas on other side 5G promises data bandwidth higher than 1Gbps and frequency band with in 3 to 300 GHz. The basic issues faced by 4G LTE technology which is poor quality of service and quality of experience faced by user in using various application will be analyzed with respect to available throughput received by user. An appropriate use of future throughput will be addressed in this paper which can be used for various application and Mobile network operator can get insight for better network management as per there need. (Chris Mooney; 2018)

1.1 Background and Motivation

A general complain from a customer to an IT operation is network slowing down or poor application performance. Mobile data traffic is increasing tremendously over the decade this may be due to of 3G/4G LTE as well as the growth in smartphone adoption, new technologies, and applications. It is expected that the global data traffic will increase by tenfold between 2014 and 2019 approximately 24.3 exabytes per month by 2019 Liu and Lee (2015). According to Vodafone (2018) the annual mobile data usage will reach up to 701 thousand petabytes till 2021. also been known from Raca (2017) that share of mobile traffic at present is of mobile video which is around 60% and expected to increase to 78% by 2021. The increase of traffic posed problem for network for mobile network operator. The throughput for the cellular device achievable to in-network can fluctuate significantly over a period of seconds due to various reasons. The change can be observed in radio channel condition because of algorithm which can cause on condition. This is due to moving and can get to one base cell and immediately enter into the other base cell range at that time scheduler comes under picture. Can face some time a large amount of delay to receive the mobile data that is due to the network congestion Raca et al. (2017)

The applications use embedded logic for streaming video from Youtube, Netflix, and VoIP. The embedded logic will understand the behavior of network and automatically adjust the streaming in appropriate way in such manner that will not degrade the quality of experience for user. It adapts the details such as encoding bitrate this can possible only if the algorithm has the future estimated throughput. Accurate throughput prediction can be made available than applications can make better adaptive decision.

Apart from this user may get affect and experience the poor throughput due to the cell congestion. In this scenario more, user are connected to eNB and tower does not have enough resources for allocation to the user hence, such type of issue can be faced by users. Congestion typical effects are queueing delay, packet loss and in some case block

the new connections. This is possible when network scheduler uses Round Robin (RR) here, scheduler try to allocate cyclically resources to all user who send request to connect with base tower but provide bad cell throughput. The best method for scheduling is proportional fair (PF), it tries to provide maximum throughput to all user but do not sacrifices the minimum throughput range that to be provide for each user. In such type of scheduling if future throughput will be available than mobile network operator than scheduling can be modified, and new strategies can be made for better network management.

1.2 General findings

The related work has been done in the past for predicting the throughput for few seconds and milliseconds in future. The prediction of throughput can be done in two ways, one is based on mathematical approach considered and other by different machine learning algorithms on the history of throughput. Due to significant impact can be seen on throughput by radio metrics hence, an effort is taken to understand variance in throughput and use of radio access network (RAN). RacaDarijo et al. (2018),Raca et al. (2018) many adaptive algorithms for video is present in various applications which uses HTTP Adaptive streaming (HAS). Streaming can adjust according to the network condition and provide quality of video to users and thus improve the quality of experience.Bui et al. (2014) suggest that Piecewise Constant Threshold (PCT) provides maximum data to when is having good signal quality or else it on buffered data when quality is low. Hence, for scheduling algorithms to adopt the forecast and channel quality will help in appropriate scheduling the resources as required.

1.3 Area of Research and value

As it is clear from work and background that throughput gets affected highly due to a network condition and channel condition hence, radio metrics KPIs are an important feature in predicting future throughput. One of the present issue in predicting throughput for adaptation algorithms is lack of cellular dataset with the throughput variation, radio KPI, and context related information RacaDarijo et al. (2018) present the first ever publically available dataset with the context information. This dataset will be considered to carry out this research. Till date future throughput is predicted for 2s, 8s, and 12s, however it is said that longer horizon prediction can help the adaptation algorithm to optimize quality of experience, with the help of metrics such as average bitrate, rebuffering and bitrate switches. Raca et al. (2018) suggest video content is split into multiple chunks with the duration of 2s to 20s. Ideally most video play at 24 frame per second and adaptive streaming divide the video into small segment often set to 4 seconds long. After every 4 seconds different quality of video will be available it will get switch according to the network condition. This study focuses on addressing the following **research question:To what extent accurate future throughput can be predicted by using different probabilistic methods?**

Further rest of the paper is structured in such a manner. Section 2 presents brief literature review for importance of throughput in different application of video streaming, Cell congestion, and resource allocation. Section 3 represents the detailed methodology selected for implementing this research work in systematic manner.

2 Literature Review

2.1 Importance of throughput for Adaptive-Bit-rate (ABR)

This section will discuss about the findings on ABR used in applications who all support and provide the video content to the user Raca et al. (2017). There are several approaches to identify available bandwidth which is been used to analyze the available bandwidth and so ABR algorithm can be modified the approach and improved. The analysis of historical throughput, buffer occupancy, and traffic shaping are one of the appropriate data can help video provider to take necessary actions in ABR algorithm if required. The major challenge faced in this scenario is that data rate of the network over a few seconds of time it will get fluctuate these may be the various factors due to the change occur in channel, system load, and radio transmission. Following is the work done in the field of adaptive bit rate techniques by researchers.

First and foremost, the author has tried to explain importance of accurate throughput guidance which will help the applications like virtual and augmented reality to make better adaption decisions. The feasibility of predicting future throughput for seconds is tested. The author was able to achieve 8 future throughput prediction by analyzing throughput with different radio metrics. To perform this, a test was carried out through a framework as well as some field experiments. A 100 user dataset was gathered in which half of them were downloading with maximum speed and other half were uploading the content, all of the users were moving with the constant speed of 50 mph and they were latch to the seven-cell hexagonal layout configuration and provided of 10 MHz (50RBs). The method and algorithms used for accurate prediction of throughput for both static and mobile case were Exponential Weighted Moving Average (EWMA), Autoregressive Integrated Moving Average (ARIMA) and Random forest. The author was able to achieve accurate future throughput for 8 seconds where 50th percentile of all the errors are less than 15% for mobile and 2% for static devices. (Raca et al.; 2017)

User many time face the quality issue in terms of picture resolution, video resolution and data rate which indirectly give user the experience of low quality of service (QoS) by network provider. This is due the nature of downlink data rate because it tends to fluctuate from time to time. Hence this paper says that major part of poor quality of service is due to throughput which is very sensitive to context of use. So, if the content provider knows the future throughput on user end before the connection has been made then techniques like adaptive delivery strategies can be incorporated to mitigate this issue. This paper aims to achieve instantaneous throughput of connection over a period of few seconds for this various type of information were collected. First the Radio link information like Received Signal Strength Indicator (RSSI), Reference Signal Received Quality (RSRQ), Signal to Interference and Noise Ratio (SINR) was taken in consideration. Secondly the context information at user end were also collected which are (GPS) coordinates, speed, terminal category and frequency band used. Lastly the network performance details like average throughput, average number of users, connection success rate and Block Error Ratio (BLER) were collected. The analysis was carried out for 5,700 connections over 350 different cells in network. The algorithm used for this test was Random Forest and for the validation K-fold method. The results obtained were cross validation coefficient of determination at 0.85 and a median error ratio at 0.1. it was said that achievable throughput is majorly based on mobile device, radio link, cell capacity, the core network and some time at server of content provider. (Samba et al.;

2017)

The similar approach of predicting throughput was taken by the author of this paper by keeping in reference of radio channel quality, speed, distance from base station. The throughput was predicted for the next x seconds on the real throughput observed in past at y second. An author has tried to provide some statistical correlations between main throughput and contextual information. The dataset for this test was collected when users were downloading a file thousand times at different location, mobility and radio connection configurations. The method used for this experiment were Generalized Linear Model (GLM), Neural Networks (NNET) and Random Forests (RF). The major aim was to predict average throughput during download test the major factor keeping in mind of context information. A K-fold cross-validation with 10-fold was also applied. Finally, the results for half chunk of the throughput was under 7% to the actual throughput. (Samba et al.; 2016)

Recent studies indicates that accurate and semi-accurate prediction of throughput is possible for small and medium chunk of data on time scale. Prediction bitrate adaption and CrystalBall are the adaption algorithms which lead to semi-accurate throughput prediction show possible increase of user experience if the algorithms have idea of future throughput as well as prediction error in consideration for deciding the next video chunk quality Raca et al. (2017). This paper aim at predicting bandwidth quality for adaptive streaming for video and error mitigation. It is said that improving the bandwidth with good accuracy will improve the quality of experience of video streaming. Adaptive streaming downloads the future chunks according to the quality of network if the network is poor than user will face the buffering and user will get annoy. This analysis can be achieved through TCP-throughput or current buffer occupancy, accurate short time-horizon bandwidth prediction while long time-horizon is great aid to the solution. Author has designed a bandwidth prediction-aware adaptation algorithm known as crystalball and also aim to minimize the error rate of prediction. Crystalball algorithms maximize the minimum bitrate across all chunks but do not propose to maximize the average bitrate because it will degrade the quality of experience. Two sets of bandwidth trace were used for the testing which are a synthetic dataset to know the influence the effectiveness prediction based on quality of adaptation. The second data contain 40 traces of bandwidth reported every 3s was collected by riding campus bus connected with campus wifi each trace is 15 min long. (Mangla et al.; 2016)

This paper has proposed a question that if the prediction accurate bandwidth is possible than how much quality of video can be improved and so do the quality of experience of user? This question mainly depends on the condition network, If the network is providing good throughput than video streaming will take place in few seconds and user will be able enjoy uninterrupted video. It has been observed that adaptive bitrate algorithm also tries to predict the available bandwidth for the entire video and try to provide best quality with the available bandwidth that too 69%-86% of optimal quality. The data collected for this test was 20 LTE cellular traces each trace was having per-second available bandwidth for 360s. The algorithm used for prediction bitrate adaption was Nave but alone this was not enough to achieve the results hence, it was combined with the prediction of buffer occupancy and stability function was extracted by gaining maximum benefit. (Zou et al.; 2015)

To achieve good quality of experience for internet video bitrate adaption play an important role. Improvement in initial delay and high resolution as well as the initial bitrate and mid-stream bitrate is possible with the accurate prediction of throughput.

Several attempts was made to predict accurate throughput but did not able to justify it, This paper says that major three contribution were made to bridge this gap of delay with respect to bitrate which are as follows. Analysis of throughput characteristic for 20M+ sessions, secondly developed CS2P a throughput prediction by using data-driven approach and finally a prototype system was developed. The was collected in such a manner that allow to analysis the variability of throughput within and across the sessions. HTTP throughput measurement was taken form the operational platform of china leading online video provider contains total more than 219 million users monthly. Apart from this a dataset with 20 million sessions with 3 million unique IPs and 18 server IPs over 8 days. The model used for predicting mid-stream and initial low delay was Hidden-Markov-Model. The CS2P outperformed existing predicting approach done with SVR and Gradient Boosting Regression trees by 40% and 50% in terms of median prediction error. Finally, author provide an 3.2% improvement on overall quality of experience and 10.9% higher average bitrate over state of art which use Harmonic Mean for throughput prediction. (Sun et al.; 2016)

A different investigation for adaptive bitrate algorithm used to analyze the quality of experience for a user who are watching online video on their way which means they are not static user they are constantly moving. This type of users can experience a low quality of service because they are constantly moving by which their cell phone can get connect to different base tower (eNB) hence, can receive low throughput. The quality of service for moving user is still a concern.The author aim to predict TCP throughput which is essential for high quality of service for streaming videos. The method applied for this is TRUST which has two stages first it will identify the user moving pattern and then it will collect all the communication quality factors, sensor data and scenario information. In prediction stage long short-term memory (LSTM) model is used to predict the TCP throughput. Previously predicting future throughput was done with the conventional method now the question was how accurate the LSTM model is able to predict the throughput. It was observed that LSTM has predicted throughput with 44% maximum error less than conventional methods for the moving bus scenario. This paper incorporated history-based throughput prediction apart from calculation-based prediction. It has been said that history-based prediction is more accurate. (Wei et al.; 2018)

A normal average bitrate is necessary for watching high quality video at user end which in turn provide good quality of service. In this paper author says that this is not the ideal scenario every time for a user several time due to adverse network condition average bitrate is not possible to achieve on various occasion which will directly affect the user quality of experience as well as determine video quality will not be available for user to watch. This type of scenario user can experience while watching high quality video on YouTube and Netflix it has also been said that these are the major application which generate most of the IP traffic. To address this issue author, introduce ABR algorithm Storytelling Architecture Technology Experience (SATE), which provide stable and agile adaption even in bandwidth constraint and uneven network condition. The author has focused mainly on two most important network observation which are estimated throughput and playback buffer occupancy to perform this test the data was collected from University College Cork (UCC) and FCC broadband by which a specific buffer occupancy can be generated. The dataset consists of 4G LTE traces from US, Germany and Netherlands with mobility traces such as car, train and pedestrian. Other traces were gathered form YouTube which fall under category if web browsing, this was measured periodically by ISPs in the United States. Instead of good adoption and bandwidth

greater than Rmin if buffering occurs to overcome this issue a stable and agile adaption keeping in mind the fluctuation of network was build. Choi and Yoon (2019)

Adaptive video streaming for bandwidth sensitive applications normally affect sometime due to throughput. This paper says that accurate prediction of throughput will help in improve the quality as well as it will help in adaptive algorithms for the applications. Previously various step was taken by different authors to predict the throughput and tried to improvised the situation with the results but in this paper author has taken brave step to compare systematically the 7 different prediction algorithms and analyze the characteristic when applied to predict the throughput. The dataset for this test was a real-world production network trace data in 3 locations over a period of 9 months. This data is similar to the context of throughput scenario in real network. The algorithms used are Multiple Linear Regression, Neural Network Regression, Support Vector Regression, Arithmetic Mean. Author says that applying more complex algorithms does not give satisfactory results and hence, will suggest a lower error range for predicting throughput where the optimal throughput can be tested. This finding provides three insights over throughput prediction in mobile network, bandwidth is known as highly variable tends to be practicable then imaginations which can play good role in adaption video streaming. It has been found that Arithmetic Mean and Multiple Linear Regression outperformed the other algorithms and turned out to be most suitable candidate who can accurately predict the throughput than other algorithms. Another good find from this study is observed that throughput prediction horizon from 100 to 300 seconds will give accurate results can be useful in adaption strategies were the throughput is considered as input. Liu and Lee (2015)

The author aim to find the correlation between the throughput and signal quality indicators. For this several throughput indicators are been considered from the live LTE network like user throughput, cell throughput, and radio link throughput and tries to find out how these indicators are related to signal quality. It is also been found that if the analysis is carried out on the poorly selected indicators than outcome will be not as expected and can affect the user experience as well as overall network efficiency. This void is divided in five class, they are service availability, accessibility, retainability, mobility, and integrity to check the integrity average user throughput is analyzed and average cell throughput helps to understand the maximum cell capacity to serve. The dataset considered for the analysis was multiple different user latch to two different carrier at 734 MHz with 10 MHz BW and 2.132 GHz with 5 MHz BW respectively. This was measure across 234 LTE cells of an urban area and for more detailed analysis 60,000 calls trace performing DL and UL measurement were taken. It was observed that throughput was affected by the chatty applications due to last transmission time interval (TTI) and outer loop link adaption (OLLA). Hence, a weal correlation was found between throughput and radio link due to last minute TTI and with CQI because of the OLLA in short connection otherwise there is strong correlation between throughput with radio link and CQI. Buenestado et al. (2014)

An impact on users experience when they fall under MAC and physical layer of network is studied to predict TCP throughput for the wireless network with the high accuracy. Here two types of users are considered one who all are walking at normal speed and other the ones who are driving with 15-25 mph. it is evident that user are moving so they will fall under different access point due to the involvement of physical and MAC layer behavior which has significant impact on TCP throughput accurate prediction will be difficult. If the measurement of data is for long period of time than there is tendency

of predicting throughput more accurately. To test this author has used two different model, one is Exponentially Weighted Moving Average (EWMA) known as a standard time series forecast and Support Vector Regression (SVR) a machine learning based approach. The results came from these two methods lead to the expectation of author, the prediction percentage for two of the actual throughput was around 80% to 90% which is considered as accurate results on which important decisions for adaption techniques can be made. Mirza et al. (2009)

2.2 Importance of throughput for resource allocation and cell congestion

The issue which is faced by several cellphone user in terms of video stalls, choppy VoIP communication, poor web browsing experience and frustrating online gaming performance is happen due to network congestion. In other words, it can be said that user is trying to send more data to base station and in actual less data is properly sent is known as network congestion. User many times refer this as reduction in throughput and experience poor quality of experience. When users are suffering on internet they are aware of two things: throughput measured in bandwidth like megabits per second and latency time taken to surpass the traffic for connection measured in milliseconds. Hence this section focusses on impact of throughput prediction in congestion and factors affecting congestion. Predicting throughput is also essential for congestion control of LTE network management the author tends to predict throughput for two different scenario one is for single cell and other is for whole region. The dataset for this test includes internet downloads and uploads by mobile users in Hong Kong. It was collected from 1352 cell sites over the period for 21 days across the city. The data consist of timestamp, downlink, uplink cell ID as well as the GPS coordinates. To carry out the test, LTE throughput was modeled as time series and prediction carried out through two models ARIMA and exponential smoothing. It was observed in the end of experiment exponential smoothing was good for predicting throughput for whole region while ARIMA was good for single cell. (Dong et al.; 2015)

Future throughput information is vital for user equipment in Long Term Evolution (LTE). This can be useful to mitigate radio channel congestions in turn will provide better quality of experience to user, this will take place in rate adaption algorithms. Author aims to predicted application-level link quality by analyzing the performance of prediction at different velocities. The dataset for this experiment generated by Alcatel-Lucent sample level simulator in terms of eNodeB metrics where data was about Carrier Frequency, cell load, channel quality indicator, and various factors of user equipment were captured for 2400 second. The method used to for this experiment was ARIMA and authors was able to accurately predict application-level link quality by 90% to 97%. Mobile networks often suffer from fast fading or incorrect user location to distribute equal throughput to user hence the author propose stochastic model to predict user throughput in mobile network. The findings will be helpful in scheduling and resource allocation decisions apart from this author pitch in the various techniques to test errors sources and accuracy of prediction. The prediction of user throughput will cover and try to analyze whether the issue of user position or congestion of particular cell exist or not. The analysis was divided in to three group of users, first group will be placed under network area where the network is analyzed in respect with time to check the average throughput received by user who are placed across the base station cell. The second group will be looking for mobility of users to refine prediction granularity. Here the aim is to predict to which cell the user will

move into as well as the congestion level of cell and time to visit. The last group aims at the highest time granularity where the different filtering techniques and historical data is assessed. ARIMA and wavelet MultiResolution Analysis was used for the first group. For the second group a trajectory prediction was done by using Markovian and Lempel-Ziv models and then the comparison was done where Markovian outperform Lempel-Ziv with prediction possible in 98% of the cases. Lastly for the third group wavelet approximation and filtering was used to predict the bandwidth. In short the model allow to predict throughput which in turn helps to analyze the forecasting based optimization techniques to achieve objectives. (Sayeed et al.; 2015)

Multiple path exists for data sender and data receiver for which TCP throughput prediction is an important factor to analyze. In this paper authors describe new approach to predict TCP throughput. Stats reveals that prediction done for throughput within 10% of actual value 87% of time. A fair idea will be available for heavy file transfer in the traffic period. The test was carried on the testbed dataset where the end-hosts were connected over 1200 routers and traffic was allowed to flow from sending end to receiving end. The method used for this test is super vector regression and with confidence interval basic SVR predictor was extended by assuming that prediction error is normally distributed. Due to more advance innovation in mobile design and the rapidly increasing technology leads to the heterogenous network environment which will indirectly affects the performance of network and user will face the difference in quality of service. There are many challenges under this type of network which are switch over, resource management, location management, and QoS provisioning which are faced every single time, but users are unaware of it. The authour wants to address this type of issues specially focusing on enhanced throughput while switch over mechanism/algorithm regulate the methodology for the scenario occur of switch over form UMTS and LTE. All the critical parameters were considered of network and user end which allow to know the optimal vertical switch over or handover mechanism needs to adapt dynamic or non-dynamic parametric decisions. Here the authors have proposed to use the artificial intelligence and number of algorithms is run in the iteration to achieve optimal output. Several parameters were considered for initialization of handover are location, data rate, Bandwidth, RSSI, velocity, user network preference, and network coverage area. The comparison of throughput was made after the handover. The received throughput by user is classified in two categories the amount of data received by user in particular time and average cell throughput which sum average of all the users connected in a cell. Finally, the results were interpreted in such a fashion that if velocity increases average throughput seems to be decrease. Hence, more the handover occurs the throughput reduces. (Mirza et al.; 2010)

In this paper authors shade the light on the user who experience the poor quality of service when they are under the edge of eNodeB (eNB). The paper represents the improved extended modified largest weighted delay first (P-EMWDF) for the improvement of QoS for cell edge users which indirectly provide in improvement of throughput. The comparison is made between the extended and modified largest delay first. LTE network uses single carrier frequency division multiple access (SCFDMA) and the download as well as upload link supported is 100 Mbps and 50 Mbps respectively. The simulation proposed that a the different number of nodes improved throughput was observed for users with EMLWDF while the user outside the region suffer poor QoS. EMLWDF has outperformed MLWDF. (Lodwal et al.; 2019)

3 Research Method and Specification

3.1 Introduction

Research methodology is the important part to carry out any research because it gives the proper direction to the research, various researcher has incorporated different methodology to successfully achieve their targeted results and evaluate in systematic manner. For implementing this research proposal, the knowledge discovery and data-mining (KDD) methodology has been selected. KDD will guide to properly implement the project and this proposed structure will be followed throughout the research to successfully predict the future throughput using different probabilistic methods.(Fayyad et al.; 1996)

The selection of methodology differs from researcher to researcher, based upon the need of the project must be selected. The type of data and final approach of researcher for the project plays a vital role in selection. Knowledge discovery and data-mining (KDD) helps to extract knowledge from targeted data. KDD follows basic 5 step to achieve knowledge shown in below figure.

- Data selection
- Data preprocessing
- Data transformation
- Data mining
- Interpretation/Evaluation

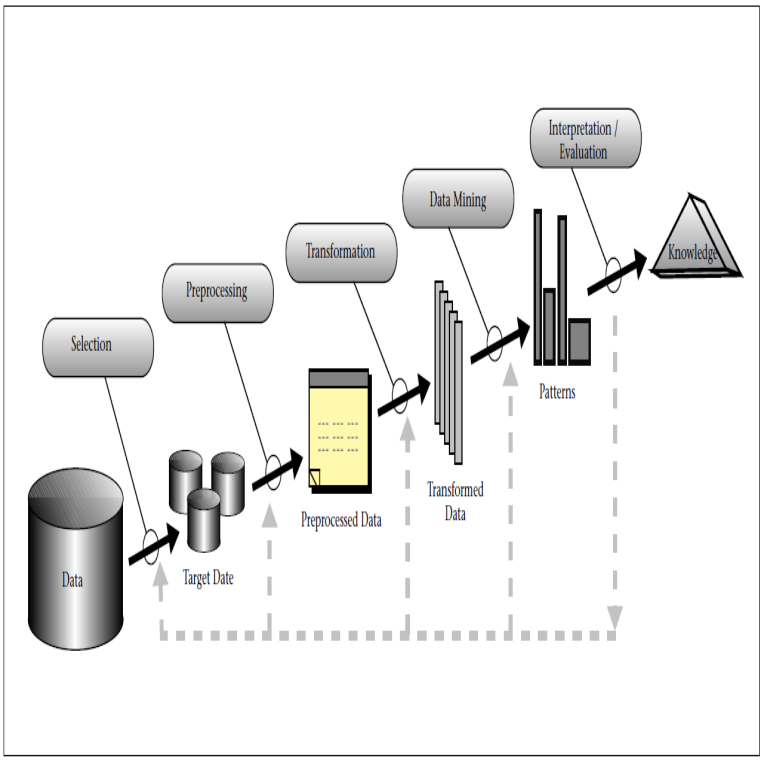

Figure 1: KDD daigram (Fayyad et al.; 1996)

4 Proposed methodology

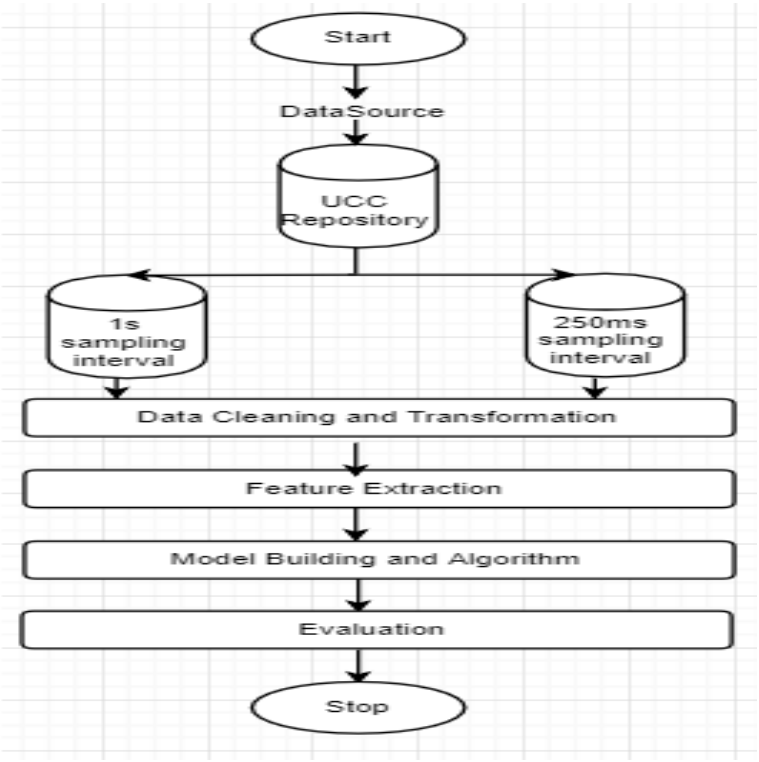

Figure 2: Process Flow Diagram

4.1 Data collection and extraction

The dataset is collected from University college cork (UCC), The dataset is about the two major Irish mobile operators with 4G trace represent the client-side cellular key performance indicators (KPIs). This dataset consists of 100 mobile users under seven-cell cluster which was generated by ns-3 simulator. According to 4050 this dataset is the first ever data about telecom which include radio matrix. Here the half of user are downloading with the maximum speed available and other half is uploading the content. All the users are moving with the constant speed of 50 mph. apart from this inter cell-site distance is fixed to 500 m and transmission power of cell to 44 dBm. Carrier bandwidth is 10 MHz (50RBs). The data consist of two sampling interval for all the users which are sampling interval of 1s and sampling interval of 250ms.[1]

4.2 Data Preprocessing

Data preprocessing in an important part before any decision taken on evaluation of data. General information regarding the data will be carried out by performing exploratory data analysis (EDA) in SPSS as well as in R studio. EDA will analyze the nature of data, distribution of data, statistical significance of independent variable on dependent

[1] https://www.ucc.ie/en/misl/research/datasets/ivid$_4g_lte_dataset$/

variable, and outlier if any. This analysis will help in identifying the important feature of dataset which can help in deciding the further step for analysis.

4.2.1 Data Cleaning

The Dataset will be extracted in its normal form hence, data prepressing will be required on the dataset. The basic cleaning, transformation, and feature selection before entering the dataset for analysis. The process of cleaning will be as follows:

- Loading of file into R studio.
- Renaming columns.
- Check for the Null value, if any understand the distribution of data and select the values which ever suit from mean, median, mode.
- Retaining the unique value.
- Removing the unwanted rows if any.

4.2.2 Data Transformation

- Combining the CSV for all users.
- Split column.
- Binary encoding.
- Changing datatype.

4.2.3 Feature Extraction

The dataset for this project consist of multiple feature for two types of data which are 1s sampling interval for 100 users and 250ms sampling interval for another 100 user hence, feature extraction techniques will help in identifying the important feature which affect the most to predicted variable. The feature will be decided on the regression analysis, and correlation of variable on prediction variable. The highly correlated variable will be selected and rest of them will be eliminated, apart from this a approach of weighted feature selection will be also be considered as the suitable option. If the correlation is negligible but according to technical specification if the variable is important than it will be considered. Below are the features present in the dataset:

- Time In seconds
- CCQI Competing channel quality indicators
- NDI Network Data intelligence
- RSRP Reference signal received power
- CSNIR Competing Signal interference noise ratio
- TBsize block size

- PDCP Throughput - Packet Data Convergence Protocol throughput
- CThroughput competing Throughput
- SINR Signal interference noise ratio
- Throughput rate of successful message delivery over communication channel.
- Delay Time taken to travel bit data from one node to endpoint.
- CellID_RSRP Cell ID reference signal received power
- TBLER Block error rate

4.3 Proposed data mining techniques

The most important part after initial analysis in the previous steps of KDD is deciding the algorithm. Now the foundation is ready, and data can be taken further for analysis of future prediction of throughput. The Time series model like Autoregressive integrated moving average (ARIMA), and ARIMAX. Other regression model like Random Forest, support vector machine, and support vector regression. After a brief literature review done for this domain, it is observed that due the nature of data and work done on historical data with the radio metrics. Various time series model and different machine learning model has been used to improve the prediction accuracy of throughput. None of the author has tried to compare two different time series model because only most used algorithm ARIMA was used till date to predict the throughput. Hence the dataset taken for this research contain multiple variable from network side as well as user side hence ARIMAX suitable option. In the end both time series model will be compared with the use of diebold mariano test and will check which model produce more accurate prediction.

4.3.1 Time Series

It is the statistical technique that deals the data which is in series of intervals or of specific time periods. There are three types of data considered for time series which are given below:[2] Time series data: A data taken at different times. Cross-sectional data: A data of more than one variable at same point of time. Pooled data: combination of series data and cross-sectional data. The model which will be used in this research will be ARIMA and ARIMAX the general brief about algorithm is given below and why it is considered for the dataset. ARIMA stands for autoregressive integrated moving average. It is also known as Box-Jenkins method. Autoregressive component: AR stands for autoregressive it is denoted by p. if p=0 no auto-correlation exist in series and p=1 series auto-correlation is still one lag. Integrated: It is denoted by d. if d=0 means series is stationary no differencing required, d=1 means series is not stationary differencing required to make it stationary, and d=2 means differenced twice. Further differencing is not reliable. Moving Average component: MA stands for moving the average and it is denoted by q. if q=1 means it is an error term and there is auto -correlation with the one lag. This method will be used to predict the future throughput for user who are

[2]https://www.statisticssolutions.com/time-series-analysis/

connected in the network based on the historical data. The equation for ARIMA(p,q) model:[3]

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q}$$

ARIMAX: An Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) this method is used when data is stationary or non-stationary, and multivariate with any type of data pattern. ARIMAX is related to ARIMA technique but ARIMA is used for univariate datasets while ARIMAX is capable for additional explanatory variables in categorical or numeric format. Assuming two time series Yt and Xt as both stationary. Then, the transfer function model (TFM) can be written as follows:[4]

$$Yt = C + v(B)Xt + Nt$$

Where: Yt = output series
Xt = input series
C = constant term
Nt = stochastic disturbance
v(B)Xt = transfer function
B = backshift operator

$$v(B)Xt = (v0 + v1B + v2B2 +)Xt$$

The dataset for this research contains different key performance indicator of network as well as from user end which means that the throughput received by user is dependent on multiple variable of network. Hence, ARIMAX is considered as a suitable technique for carrying out the prediction of future throughput. Apart from this the future forecast achieved from ARIMA and ARIMAX will compared in the terms of accuracy. Another method will be used for this research is Random Forest (RF) Random forest is the powerful algorithm it can carry out regression as well as classification. It comes under supervised machine learning techniques can also be considered as the extension of decision tree. It is normally trained with the bagging method and the forest build is an ensemble of decision trees. In general, it builds multiple decision trees and merge it together to achieve accurate results. For improving the accuracy of the model hyperparameter tuning will be done, below are some of the parameter which will be taken into consideration in this research:

- n_estimators = number of trees in the forest, more the trees better the accuracy will be.
- max_features = max number of features used for split nodes.
- Max_depth = max number of levels in each decision tree.

[3]https://www.statisticssolutions.com/time-series-analysis/

[4]https://www.smarten.com/blog/what-is-arimax-forecasting-and-how-is-it-used-for-enterprise-analysis/

Support Vector Machine/Regression (SVM/R): The SVM will be used for this dataset and aim is to obtain 3 categories for the users which will indicate minimum, average, and maximum throughput hence user, are connected to the cell tower (eNB). This category will be useful to mobile networking operators (MNO) to take necessary action to improve the performance of network by which quality of experience will be improved. Apart form this the dataset will be also predicted with SVR model too. SVR is used for regression type problem and basically works on the principle of SVM Liu and Lee (2015) K-fold cross validation will be used to improve the accuracy of model. Key terms of SVR:

- Kernel: This function is used to map the lower dimension variable into higher dimension data.
- Hyper Plane: In SVM it is the line which separates the data class in between, this line will be define and will help in predict the continuous value.
- Boundary line: Apart from hyper plane there are other two line which create the margin between two classes the data can either be on the line or outside the boundary line.
- Support vectors: These the data points which are closest to the boundary and the distance is minimum or least.

5 Evaluation

After analyzing the data through algorithm, now it is time to interpret precisely the results of model. The values which will be checked are Absolute Percentage Error (APE), Root Mean Square Error (RMSE), residual error (RE), and Autocorrelation factor. This all error value will provide strong base to discovery the results, after analyzing the value it can be compare with the actual value and the forecast value. This thorough analysis done by using radio KPIs of the network with the different model. Finally the results have to be compared and best model to be selected and appropriate benefits from future throughput will be stated for the respective application/network management decision.

6 Future Plan

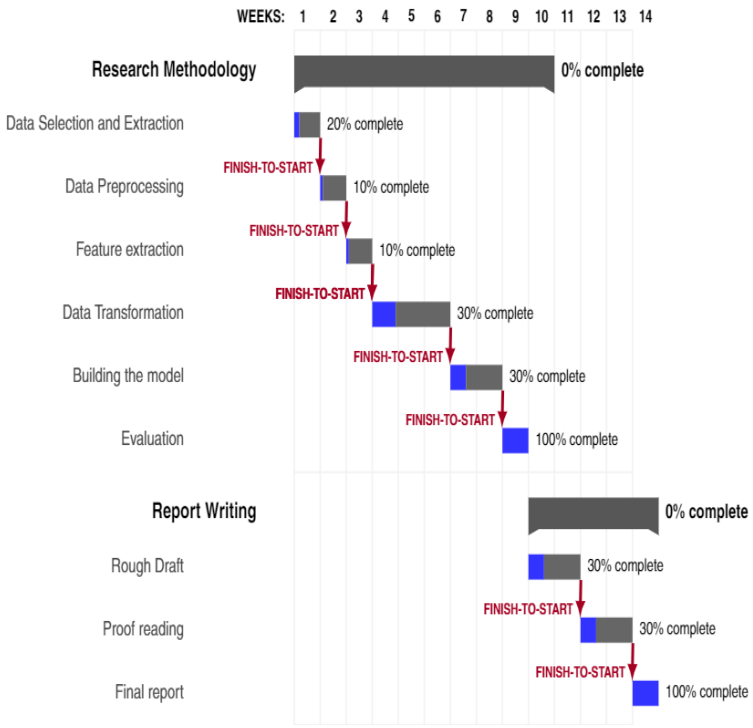

Figure 3: Gantt Chart

References

Buenestado, V., Ruiz-Aviles, J. M., Toril, M., Luna-Ramírez, S. and Mendo, A. (2014). Analysis of throughput performance statistics for benchmarking lte networks, *IEEE Communications letters* **18**(9): 1607–1610.

Bui, N., Michelinakis, F. and Widmer, J. (2014). A model for throughput prediction for mobile users, *European Wireless 2014; 20th European Wireless Conference*, VDE, pp. 1–6.

Choi, W. and Yoon, J. (2019). Sate: Providing stable and agile adaptation in http-based video streaming, *IEEE Access* .

Chris Mooney, Madhav Hari, A. A. M. H. J. P. L. M. M. (2018). Industry top trends 2019 - telecommunications, *SP Global Ratings*, VDE, pp. 1–2.

Dong, X., Fan, W. and Gu, J. (2015). Predicting lte throughput using traffic time series, *ZTE Communications* **13**(4): 61–64.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *American Association for Artificial Intelligence* **17**: 37–54.

Liu, Y. and Lee, J. Y. (2015). An empirical study of throughput prediction in mobile data networks, *2015 IEEE Global Communications Conference (GLOBECOM)*, IEEE, pp. 1–6.

Lodwal, H., Yadav, A. and Panchal, M. (2019). A novel algorithm to improve quality of service of cell edge users in lte, *International Conference on Advanced Computing Networking and Informatics*, Springer, pp. 237–245.

Mangla, T., Theera-Ampornpunt, N., Ammar, M., Zegura, E. and Bagchi, S. (2016). Video through a crystal ball: effect of bandwidth prediction quality on adaptive streaming in mobile environments, *Proceedings of the 8th International Workshop on Mobile Video*, ACM, p. 1.

Mirza, M., Sommers, J., Barford, P. and Zhu, X. (2010). A machine learning approach to tcp throughput prediction, *IEEE/ACM Transactions on Networking (TON)* **18**(4): 1026–1039.

Mirza, M., Springborn, K., Banerjee, S., Barford, P., Blodgett, M. and Zhu, X. (2009). On the accuracy of tcp throughput prediction for opportunistic wireless networks, *2009 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, IEEE, pp. 1–9.

Raca, D. (2017). Throughput prediction in cellular networks, *MISL* .

Raca, D., Zahran, A. H., Sreenan, C. J., Sinha, R. K., Halepovic, E., Jana, R. and Gopalakrishnan, V. (2017). Back to the future: Throughput prediction for cellular networks using radio kpis, *Proceedings of the 4th ACM Workshop on Hot Topics in Wireless*, ACM, pp. 37–41.

Raca, D., Zahran, A. H., Sreenan, C. J., Sinha, R. K., Halepovic, E., Jana, R., Gopalakrishnan, V., Bathula, B. and Varvello, M. (2018). Incorporating prediction into adaptive streaming algorithms: A qoe perspective, *Proceedings of the 28th ACM SIGMM Workshop on Network and Operating Systems Support for Digital Audio and Video*, ACM, pp. 49–54.

RacaDarijo, Quinlan, J. J., Zahran, A. H. and J, S. C. (2018). Beyond throughput: a 4g lte dataset with channel and context metrics, *Proceedings of the 9th ACM Multimedia Systems Conference*, ACM, pp. 460–465.

Samba, A., Busnel, Y., Blanc, A., Dooze, P. and Simon, G. (2016). Throughput prediction in cellular networks: Experiments and preliminary results, *1ères Rencontres Francophones sur la Conception de Protocoles, lÉvaluation de Performance et lExpérimentation des Réseaux de Communication (CoRes 2016)*.

Samba, A., Busnel, Y., Blanc, A., Dooze, P. and Simon, G. (2017). Instantaneous throughput prediction in cellular networks: Which information is needed?, *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, IEEE, pp. 624–627.

Sayeed, Z., Grinshpun, E., Faucher, D. and Sharma, S. (2015). Long-term application-level wireless link quality prediction, *2015 36th IEEE Sarnoff Symposium*, IEEE, pp. 40–45.

Sun, Y., Yin, X., Jiang, J., Sekar, V., Lin, F., Wang, N., Liu, T. and Sinopoli, B. (2016). Cs2p: Improving video bitrate selection and adaptation with data-driven throughput prediction, *Proceedings of the 2016 ACM SIGCOMM Conference*, ACM, pp. 272–285.

Vodafone (2018). Annual mobile data usage worldwide from 2015 to 2021 (in thousand petabytes), *Statista.com* p. 6.

Wei, B., Kawakami, W., Kanai, K., Katto, J. and Wang, S. (2018). Trust: A tcp throughput prediction method in mobile networks, *2018 IEEE Global Communications Conference (GLOBECOM)*, IEEE, pp. 1–6.

Zou, X. K., Erman, J., Gopalakrishnan, V., Halepovic, E., Jana, R., Jin, X., Rexford, J. and Sinha, R. K. (2015). Can accurate predictions improve video streaming in cellular networks?, *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, ACM, pp. 57–62.