Mobile network traffic prediction

Introduction

Mobile phone devices such as smartphones, tablets, wearable devices as well as mobile phone subscribers are increasing rapidly.According to a report presented by Ericsson, the mobile devices have surpassed the world population. Due to such a huge growth in mobile devices and mobile phone subscribers, the congestion of mobile network is not unusual. This increases traffic congestion on the base station and energy consumption as well.

Problem statement

To overcome this network traffic congestion the traditional approach or brute force method is deploying more base stations, and adding more data processing units (increased capital expenditure) thus increasing energy consumption and maintenance spending (increased operating expenditure). This project proposes to use call detail record (CDR) to analyse the various patterns and predict the network utilization so that an informed network expansion can be taken up by the mobile service provider resulting in improved Quality of Service (QoS) to the end customers.

Dataset Description

In this project, a multi-source dataset released by Telecom Italia in 2015 is used. The dataset is one of the most comprehensive collections from an operator and also publicly available. Originally, the collection was created for a big data challenge with projects ranging from mobile networking to social applications. Provided data points include records in telecommunication, weather, news, social network, and electricity from the city of Milan and Trentino during November and December 2013. For mobile Internet traffic forecasting, we focus on telecommunication records. Geographical grids are first defined for data recording. The city is divided into 100×100 areas with aggregated call detail record (CDR) data. Each grid has a unique square ID covering an area with the size of 235 × 235 meters. The telecommunications dataset contains the following information used in this work:

- **Square ID**: the identification of the square of Grid.
- **Time interval**: Data is aggregated for 10 minute time interval, the beginning of the time interval of the record is given. The end interval time can be obtained by adding 10 minutes to this value
- **Internet traffic activity**: the number of CDRs generated during the time interval in a square id
- **Received SMS** a CDR is generated each time a user receives an SMS
- **Sent SMS** a CDR is generated each time a user sends an SMS

- **Incoming Call** a CDR is generated each time a user receives a call
- **Outgoing Call** a CDR is generated each time a user issues a call

Data source: telecom italia open big data challenge dataset - A multi-source dataset of urban life in the city of Milan and the Province of Trentino Dataverse

Methodology

1. Cleaning and preprocessing
2. Carrying out visual EDA for getting a general sense of data and understanding the relationships between various attributes. Since the dataset has very few attributes available, we have to analyse and understand each of the attributes in the dataset and derive additional attributes to predict the behaviour of subscribers.
3. Ascertaining critical attributes and their effect on target
4. Applying ML models to predict network traffic

Deliverables

1. Plots showing the relationship between various attributes, and their relevance in final target prediction.
2. Model to predict the network traffic trends per grid for the next hour