Data Mining for Mobile Internet Traffic Flow Forecasting

Saber Kh. Elmabrouk
School of Applied Sciences and Engineering, Libyan Academy
Janzur, Tripoli, Libya
saber.elmabrouk@academy.edu.ly

Alhasan Nureddin Salem
Libyana Mobile Phone, Tripoli, Libya

Abstract
Internet traffic can be described as a general term that includes the transmission of internet data between different devices and systems. The analysis and prediction of internet traffic is a proactive approach to ensure secure, reliable and qualitative network communication. Several linear and non-linear models are proposed and tested to analyze the prediction of network traffic, including techniques based on regression analysis, artificial intelligence and data mining. These interesting combinations of internet traffic analysis and forecasting techniques are implemented to achieve efficient and effective results. Timely and accurate prediction of the use of internet data is a topic of great importance in the telecommunications industry. In addition, internet traffic data is important for many applications in telecommunications management, such as understand customer behavior, optimal planning of the capacity of networks, successful decision making and maintaining the quality of services at guarantee level in the future. This situation inspires us to rethink data mining with internet traffic forecasting problems. This study provides an overview of data mining in telecommunications and proposes a novel model for forecasting mobile internet traffic based on artificial neural networks. The analysis of mobile Internet traffic data during the last five years shows that the month of August has a higher traffic flow, while the lowest flow was the month of June. The selected model has three layers with a determination coefficient of 97.5%.

Keywords
Data mining, Internet traffic, Forecasting, Telecommunication, ANN.

1. Introduction
Each manager wants to know the exact nature of future events, act accordingly, or plan activities when there is sufficient time to execute the plan. The effectiveness of the plan depends on the awareness of future events. But all managers are planning for the future, whether or not future events are known. This means that they are trying to predict the future based on the historical data and on their judgment and experience. In telecommunications companies, the managers using historical data to predict the future to understand customer behavior, maintain quality of service, optimize network capacity, monitor network traffic flow, network failures, maximize profit, fraud detection, anticipate new products and services to improve existing products and services, and much more. Accurate estimation of network performance is critical to the success of any type of network, whether voice or data. The two most important factors that it needs to consider in managing the telecommunications are service and cost. Quality of service is essential to maintain customer satisfaction whilst cost always influences profitability. This study aims to provide an overview of data mining in telecommunications and proposes a novel model for forecasting mobile internet traffic based on artificial neural networks (ANNs).

2. An Overview
Recently, the concept of data mining has been increasingly used in various industries. This discipline is primarily concerned with designing algorithms to extract new information from large, existing structured data. Data mining, also known as Knowledge Discovery in Databases (KDD). KDD was coined in 1989 by Gregory Piatetsky-Shapiro during the first workshop on KDD (KDD-89). The term data mining appeared in the database community in 1990 (Fayyad et al., 1996). However, the challenge of data mining is to draw valuable analytical conclusions. Thus, most telecommunications companies have realized that the vast volume of data they collect and possess could be effectively utilized for solving their business problems by converting them into information and knowledge. Data mining can be considered as a technique that automatically generates this knowledge from available data (Joseph, 2013). As a result,
many authors have analyzed the use of telecommunication services in different contexts, but little research has focused on the data traffic generated, particularly on smartphones and their ability to use them. The methods used in data mining, such as machine learning (ML), belong to the field of artificial intelligence (AI). A system becomes artificially intelligent by feeding it with relevant patterns. ML is a set of algorithms that analyze data and learn from them to make informed decisions. Whereas ANNs is one such group of algorithms for ML. Thus, ANN is a special set of algorithms that has revolutionized ML. However, as ANN best identify patterns or trends in data, they are well suited for prediction and forecasting applications. Anyway, different AI techniques have been successfully applied to a wide range of extraordinary activities in the telecommunications industry. Previous research has shown that data mining has used various types of telecommunication data that can be classified into three different sets, namely: call details, network and customer data.

**Call details data:** This is information related to calls, which is stored as a record of the details of the call that contains information such as originating and terminating phone numbers, date, time and duration of call. Usually this dataset is not directly used for data mining (Joseph & Madhuri, 2013). The following features can be generated from this data: Average call duration, Average number of call created per day, Average number of call received per day, Percentage of no-answer calls, Percentage of day time calls, and Percentage of weekday call.

**Network data:** Telecommunication networks contain thousands of interconnected components. These components are capable of generating big data. This data is used for network management such as network diagnosis and fault detection. Data mining technologies are used to identify network faults by automatically extracting knowledge from this dataset. Chaudhary & Singh (2015) defined the network fault as an abnormal operation or defect at the component, equipment, or sub-system level that is significantly degrades performance of an active entity in the network or disrupts communication.

**Customer data:** Since telecommunications companies have millions of customers, it is very important to have a database to store information about their customers. This information may include; customer name, address, billing data and payment history, call detail data, subscription information, service plan, and more.

To survive in a highly competitive telecommunications market, companies need to adapt to external changes. Customer Relationship Management (CRM) has proven to be an important control tool for identifying environmental changes in business processes and making difficult decisions. The literature review shows that most data mining applications in telecommunications fall into three categories: CRM, fraud detection, and network management. However, most of these studies were devoted to customer relationship management. These three categories are described below with a general description of their application in data mining.

**Customer Relationship Management** (CRM): A CRM is defined as a set of activities a business performs to identify, qualify, acquire, develop and retain loyal and profitable customers by delivering the right product or service to the right customer, through the right channel, at the right time, and the right cost (Galbreath & Rogers, 1999). Data mining applications in CRM can take many forms, among them; 1) Basket Analysis: developing inventory and store layout strategies. 2) Sales Forecasting: examines time-based patterns for make restocking decisions, helps in supply chain management, and financial management. 3) Database Marketing: increasing productivity, optimization company resources based on customer profile, demographics, tastes, preferences, and purchasing behavior, and more. 4) Predictive Life-Cycle Management: helps the organization predict customer lifetime and properly serve each segment. 5) Market Segmentation: increase efficiency by finding out which customers are interested in buying your products and designing marketing campaigns and promotions based on their tastes and preferences. 6) Product Customization: predict which features need be bundled to meet customer demand by customize products according to the exact needs of customers, and 7) Fraud Detection: By analyzing what has turned out to be fraudulent in the past, the company can take corrective action and stop the future occurrence of such events. Thus, the lessons learned and collected data from customer information should be used to improve business and customer relationship behavior.

Moreover, nowadays, the telecommunication market matured and the churn rates became high. This maturing of the market and increasing competition leaded the companies to focus on their existing customers and to find a way not to let them go. Note that the term "churn" refers to the monthly or annual turnover of the clientele as defined by Strouse (1999). Customer churn rate designates the percentage of customers the company has lost over a given time period. Churn rate is a reverse side of retention rate, which shows the percentage of customers a company has retained over a specific period. The most practical ways to calculate customer churn rate is; Customer Churn Rate = (Customers lost...
during the period/Total customers at the start of the period) *100, where: Customers lost = Total Customers at the start of the period + New customers during that period - Total Customers at the end of a period (Bitrix24, 2019).


Fraud Detection: More than a third of organizations have experienced economic crime as reported by over 6,000 respondents to PwC's Global Economic Crime Survey in 2016. In many existing literature, subscriber intent plays a central role in defining fraud. Johnson (1996) defines fraud as any transmission of voice or data across a telecommunications network where the intent of the sender is to avoid or reduce legitimate call charges. Also, Kaski et al (1998) defines fraud as obtaining unbillable services and undeserved fees. Likewise, Blavette (2001) noted that telecommunication fraud occurs whenever a perpetrator uses deception to receive telephony services free of charge or at a reduced rate. It is a worldwide problem with substantial annual revenue losses for many companies.

There are many fraud techniques, some of which are quite sophisticated and combine more than one known method. Telecommunications fraud is not static; new techniques evolve as the companies put up defenses against existing ones. The fraudsters are smart opponents, continually looking for exploitable weaknesses in the telecom infrastructure. Part of their motivation is accounted for by the fact that once an exploit is defined, there are thousands (or millions) of potential targets. New types of fraud appear regularly, and these schemes evolve and adapt to attempts to stop them (Becker et al, 2010). However, most of the data mining work done in detecting telecommunications fraud aims to detecting or preventing fraud (Fawcett & Provost, 1997 and Hilas, 2009) and subscription (Estevez et al, 2006 and Farvaresh & Sepehri, 2011) methods, because these are the fraud methods that have caused the most damage to the telecommunications industry. Some papers, such as Collins (1999 A), Collins (1999 B), Collins (1999 C) and Hoath (1998) describe state of telecommunication fraud. However, these papers do not discuss the details of the detection process or the organized fraud model. In 2002, Bolton & Hand produced a report on a review of statistics and ML tools as an effective technology for telecommunications fraud. Likewise, Becker et al (2010) discussed some of the major fraud systems and their techniques. The study led to general conclusions about fraud detection. They specifically support the use of simple, understandable models, effective use of visualization and a flexible environment, and emphasize the importance of data management. In 2018, Zhou & Lin summarized the anti-fraud measures available in the telecommunications industry. Besides, many data mining techniques exist in the literature for (a) supervised learning based on training data of known fraud and legitimate cases and (b) unsupervised learning with data that are not labeled to be fraud or legitimate. The literature shows that there are three data mining techniques used for fraud analysis: Bayesian network, decision tree, and back-propagation. The Bayesian network is used for the classification task. Classification determines which class a given data belongs to by defining predefined categorical classes. Decision trees are used to create descriptive models. Descriptive models are created to describe the characteristics of the error. Akhter & Ahamad (2012) discussed in detail the role of ANN in the prevention of telecommunications fraud. They came to the conclusion that ANN is a better way to detect phone fraud because it incorporates adaptability, speed and efficiency. Farvaresh & Sepehri (2011) used a hybrid approach consisting of a pre-processing, grouping and classification phase. In the clustering phase, SOM and K-means were combined, and in the classification phase, decision tree, ANN, and supporting vector machines were examined as a single classifier and bagging, boosting, stacking, majority and consensus voting as ensembles were examined. The performance of all single and ensemble classifiers was evaluated using various indicators and compared with statistical tests. It was shown that support vector machines among single classifiers and boosted trees among all classifiers have the best performance in terms of various metrics. Recently, Arafat et al (2019) suggested the use of various ensemble classifiers to overcome the highly biased dataset and to facilitate more accurate classification. They found that the Extreme Gradient Boosting algorithm gives the best results in terms of correctness and performance.

Network Management: Network monitoring is so essential because it has live insights. Having up to the minute knowledge about what’s happening in the network is invaluable for any networking team. we are able to act when problems occur without relying on end-user reports. The real added value of the AI network management tool is that it intelligently automates tasks that are only becoming more general and complex. As network tools become more sophisticated and accurate, it becomes increasingly difficult to effectively manage and manage a limited workforce.
Network monitoring tools have begun to introduce AI and ML into network systems. In addition to recognizing and alerting users to problems, however, AI can apply solutions to common problems without human intervention. AI and ML allow network performance monitors to learn about common network issues. AI and ML don’t just work on its own. It needs to be trained to respond to events that happen on your network. This means that you can customize AI to analyze and respond to event types. If you identify the same problem over and over again, AI can figure out the best solutions to deal with them. Finally, it will provide enough data to make a decision. In addition to being able to find network problems, AI is an incredible resource for preventing malicious problems before they damage the network. AI can improve problems when they are noticed, allowing users to focus more on issues that require deeper analysis and insight (Hein, 2019 and Enterprise Integration, 2019).

In 1993, Sasisekharan introduced a method to store data and identify trends in faulty networks. Using this approach, a number of patterns have been identified on AT&T’s worldwide network, which can be used by AT&T to improve network reliability. In another work, Sasisekharan et al. (1996) provided an approach to maintain a massively interconnected communications networks over time. This approach has been used to detect and predict chronic transmission faults in AT&T’s digital communications network. Liu (2014) presented a monitoring and analysis system for large-scale networks based on Hadoop. Their results demonstrated that the system can efficiently processes 4.2 Tb of traffic data from 123 Gb/s links with high performance and low cost every day. In 2016, Qu et al. introduced a multilevel model mining architecture that supports automatic network management by discovering interesting patterns from telecommunication network monitoring data. The proposed architecture leverages and combines existing frequent item set discovery over data streams, association rule deduction, frequent sequential pattern mining, and frequent temporal pattern mining techniques while also making use of distributed processing platforms to achieve high-volume throughput.

3. Data Mining for Forecasting
Telecommunication service provides some kind of time series data, that is, data collected over time, including a time stamp. In the context of forecasting, the decision maker has to find ways to get the value of big telecommunication data. However, predictive data mining offers an opportunity to leverage the time series data sources that are already available to decision makers that can directly impact profitability. For reference purposes, Rey et al. (2012) defined short-range forecasts as one to three years, medium-range forecasts at three to five years and long-term forecasts at more than five years. Generally, the authors agree that a period of more than ten years should be considered as a scenario rather than a forecast. Most business modeling involves quarterly forecasting. Quarterly data is the frequency with which the vast majority of external data providers store and forecast historical data. Gor (2003) defined forecasting as the process of estimating a future event by transmitting historical data. Historical data is systematically combined in a predetermined manner to obtain an estimate of the future. Tuovila (2019) reported that forecasting is a technique that uses historical data as inputs to make informed estimates that are predictive in determining the direction of future trends. Businesses use forecasting to determine how their budget or plan their costs over the next period. This is usually based on forecast demand for the products and services offered. Sanders (1998) has shown that poor forecasting can lead to catastrophic decisions and prediction accuracy can make successful decisions, so forecasting has improved in many areas, such as long-term forecasts by airlines. Whereas, prediction is concerned with estimating the outcomes for invisible data. For this purpose, it fits a model into a training dataset, which results in an estimator \( f^*(x) \) that makes predictions for new \( x \) samples. Forecasting is a sub-discipline of prediction, in which we are making predictions about the future based on time series. Thus, the only difference between prediction and forecasting is that we consider the temporal dimension. The form of the forecasting is \( f^*(x_1, \ldots, x_t) \) where \( x_1, \ldots, x_t \) indicate historic measurements at time points \( 1, \ldots, t \), while the estimate relates to time point \( t+1 \) or some other time \( t \) in the future (Data Science, 2019). Nonetheless, there are many ways to forecast. Each method is different and depends on the purpose of the forecast, the data required, the availability of the data and the time frame for forecasting the event. Gor (2003) presented three different forecasting methods: qualitative, time series, and causal methods.

3.1 The Process of Data Mining for Forecasting
Figure 1 summarized the seven steps of the data mining forecasting process; 1) identify the problem, 2) select the forecasting period (time horizon), 3) select the forecasting method. This requires knowledge of the different forecasting methods, which situations to use, how reliable they are; what type of data is needed. Based on these considerations; one or more methods can be selected. 4) data selection: use the various metrics identified in step 2 as an appropriate source that is compatible with the method(s) selected in Step 3. 5) make a forecast: apply the model using the collected data and calculate the value of the forecast. 6) evaluation: before using the forecast model obtained.
by one of the methods (step 5), it should be evaluated based on the confidence interval and error analysis. 7) the model is ready to run.

Fig. 1 – The Seven Steps of the Data Mining Forecasting Process

4. ANN Applications for Telecommunications Forecasting

Recently, the use of ANN has contributed to improving the performance of the telecommunication service and defining future technical and capacity requirements. This section provides a brief introduction to ANNs, followed by some previous ANN studies on telecommunications applications. Basically, ANNs are used to model non-linear statistical data and mimic the functioning of human brain neural networks. Figure 2 shows a simple neuron model.

![Simple Neuron Model](image)

It can be seen from Figure 2 that the neuron is a processing element that takes a number of inputs, $x_1, x_2 \ldots x_n$, weighs them, $w_1, w_2, \ldots, w_n$, and sum them up together with a bias parameter to get the value $a$. Then, $a$ value is processed by the activation function $f(u)$ to get $d$, and then, the neural output error, $e$ is calculated by subtracting the result $d$ from the target, $t$. The essence of the neuron model is to minimize the output error, according to some optimization criteria, to improve the fit. Thus, all the neurons in the ANNs are trained using a block of input and output datasets so that the NNs know the problem. The training can be classified into two groups: supervised and unsupervised training. The former needs predefined training data that reflects network behavior. The network target is already known and compared to the output when the input is applied to the network. Learning rule modifies $w_1, w_2, \ldots, w_n$ and bias to bring outputs closer to target. However, the weight ($w_1, w_2, \ldots, w_n$) and bias parameters are only updated with the input data when using an unsupervised training. The multilayer perceptron neural network (MLPNN) is an extension of the neuron model, and connections can only be one-way in the network. However, in the field of ANN modeling, a radial basis function network (RBFNN) is an ANN activated by radial basis functions. It is often used to create models for regression type problems (Pijush et al, 2017). Also, Group Method of Data Handling Polynomial Neural Network (GMDH-PNN) is a model for obtaining high order input output relationship in time-series problems. GMDH-PNN is an inductive unidirectional polynomial network that is made of large number of layers and each layer contains many neurons (Koo et al, 2014).

Application of ANNs to traffic prediction by Ardhan et al (2007) Markus et al (2012), Raheem & Okereke (2014), Ozovehe (2015) and Aliyu et al (2017) shows that ANN, RBFNN and GMDH-PNN can predict traffic with very high accuracy. In 2004, Zhao et al. proposed two ANN models for network traffic forecasting to achieve two goals: forecasting a 24-hour load shape at a time and forecasting peak load per day. They collected network traffic data for a year and a half. Data traffic from February 1, 1998 to July 31, 1999 was used as a training data set, and data from August 1, 1999 to August 31, 1999 was used as a test data set. The 24-hour load shape prediction model consists of four layers [24-20-30-1]. The input layer includes 24 inputs, two hidden layers with 20 nodes in the first hidden layer, 30 nodes in the second hidden layer, and one node in the output layer. The activation function used in the hidden layer was a sigmoid, and the activation function for the output layer was a linear activation function. The peak load...
forecasting model consists of five layers [16-15-20-30-1]. The input layer has 16 nodes, the output layer has only one node connected with three hidden layers. The 1\textsuperscript{st} hidden layer has 15 nodes, the 2\textsuperscript{nd} hidden layer has 20 nodes, and the 3\textsuperscript{rd} hidden layer has 30 nodes and trained by Levenberg-Marquardt algorithm. However, the names of the input variables of both models were not mentioned in their study.

Cortez et al (2006) analyzed the effectiveness of three different time series forecasting approaches when applied to TCP/IP internet traffic; Holt-Winters, ARIMA methodology and a Neural Network Ensemble (NNE) approach. Comparison of these time series forecasting methods shows that NNE produced lowest errors. Data was collected from two different ISPs in every five minutes and every hour. However, the input variables used in their study was not specified. They used a logistic activation function in the hidden layer and a linear activation function in the output layer. The Levenberg-Marquardt algorithm was used to construct the model. The average squared error was used to measure the prediction performance of the models.

Oravec et al. (2008) proposed three different moving average neural network models for predicting video traffic stream; multilayer perceptron, radial-based function networks and back propagation through time neural networks. The multilayer perceptron network configuration was [3-10-1], the radial-based function network configuration was [3-500-1], and the back propagation through time configuration was [3-20-1]. The activation function was logically in the hidden layer and the linear activation function in the output layer. The Quasi-Newton algorithm was used to train the model. The data used for training and the testing was taken from the video stream files of the telecommunications network group. Training and testing dataset consists of 2000 points. The authors did not specify the three input variables used in the first layer for the three proposed models. The best results of the forecast were achieved from the back propagation through time network model.

In 2009, Gowrishankar & Satyanarayana modeled wireless network traffic in a short period of time using neural networks and statistical methods. The results of both methods were compared at different time scales; 1 second, 10 seconds and 1 minute. The ANN model has a hidden layer, with a sigmoid activation function in the input layer and a linear activation function in the output layer. They selected Quasi-Newton algorithm to train the model. The authors did not specify the variables used in the first layer of the model.

Junsong et al (2009) proposed Elman neural network model to predict Internet traffic as a time series. The model architecture was [2-2-1]. This means two variables in the first layer, one hidden layer with two nodes, and one target (web traffic) in the output layer. Both input and output layers activated by sigmoid activation function. The training algorithm was Levenberg Marquardt. The authors did not specify the two variables used in the model.

In 2010, Zaleski & Kacprzak introduced a Radial based neural network with a self-organizing algorithm that predicts short-term values of traffic generated in packet switching networks. The selected model architecture [7-3-1] consists of one hidden layer with three nodes activated with Gaussian activation function in the input layer. Whereas on the output layer, one neuron acts as a linear activation function of hidden layer weighted outputs. The authors did not define the seven variables used in the model as input layer.

Barabas et al. (2011) used two types of ANN (single-task learning (STL), multi-tasking (MTL) network approaches) to predict the traffic load on Ethernet. The ANN model configuration consists of [4-5-1] for the STL approach and [8-3-1] for the MTL approach. The activation function of neurons in the hidden layer was a sigmoid function, while the output nodes have a linear transfer function. The Levenberg-Marquardt algorithm was the training algorithm. They used 200 Mb consecutive traffic loads to build the prediction models, and 20 Mb values to test the models. They concluded that the STL-based ANN model had the worst forecast performance, while the ANN model with the MTL approach had the best performance. The authors did not define the selected variables used in the models as input layer.

In 2012, Botoca & Patrascu modeled two different approaches to ANN (Multilayer Perceptron (MLP) and Jordan-Elman Neural Network) to predict channel quality in long-term evolutionary networks. Data was collected from radio networks, 200 recorders were used to train the model, 77 records to validate it and 102 records to test it. The MLP network structure consisted of [5-8-1]. Hyperbolic touch activation function used for all layers. The generated propagation error algorithm used for training and validation. In the case of Jordan, the Elman NN structure was [4-8-1], with a hyperbolic touch activation function in each layer. The generated propagation error algorithm used for training and validation. They found that the Jordan-Elman Neural Network approach achieves better forecasting results than the MLP network approach. The authors did not define the selected variables used in both models.
Pandey et al. (2013) used ANN method to predict cellular network traffic during busy hours (between 5 pm and 7 pm). The ANN model consisted of [9-10-4-1]. Nine parameters were selected as input variables in ANN; SDCCH Reservation Attempts, SDCCH Successful Calls, TCH Assignment, TCH Successful Calls, TCH Availability Ratio, Inbound Transfer Success Rate, Outbound Transfer Success Rate, Mid-Term Traffic, and Average Hold Time. The Tan sigmoid transfer function provided the best generalization of data across all layers using the Levenberg-Marquardt as a training algorithm.

Ozovehe et al. (2018) formed three types of ANN (MLPNN, RBFNN, GMDH-PNN) and Neuro-fuzzy Inference Systems models based on traffic measurements during busy hours taken from certain GSM / GPRS sites in Abuja to predict traffic congestion in certain macrocells. They found that GMDH-PNN more predictive of TCH congestion than other models as it resulted in lower errors for mean absolute error, standard deviation, root mean square error and higher determination coefficient (R²). The authors did not define the selected variables used in the models.

Recently, Erunkulu et al. (2019) introduced an ANN model for predicting call drop events in a global network of mobile networks. The accuracy of the model was 87.5%. The architecture of the model consisted of five layers [5-4-5-4-1]. Inputs for the first layer are: received signal level, received quality, Image error rate, Bit error rate and Timed transmission. However, the authors did not provide the training algorithm and activation functions used to build the model.

5. Forecasting Mobile Data Traffic Over Mobile Phone Networks in Libya

5.1 Data Acquisition and Analysis

Mobile voice and internet data was recorded as a daily traffic monitor from January 1, 2013 to December 31, 2017 on the Libyan real mobile network, which covers the Tripoli region. Between 2013 and 2017, it was noticed that the demand for voice services increased insignificantly, while the demand for Internet services increased significantly. Whereas the proliferation of smartphones supporting new mobile technologies has enabled mobile users to initiate voice calls over Internet services, resulting in high data traffic of up to 1.5 Mb per user in 2017, thus reducing the need for lower voice calls over the last five years from 2013 to 2017. Figure 3 shows a consumer comparison of voice and data usage between 2013 and 2017.

![Fig. 3 - Consumer Comparison of Voice and Data Traffic (2013-2017)](image)

Nowadays, many applications such as voice, video calling, and social media have been replaced by voice over internet services. The spread of modern mobile devices as the smartphone can support new mobile technologies. Today, the behavior of mobile users has changed, and most of them are using smart devices to manage new mobile networks, each of which increases data traffic and affects the network's ability to support traffic in the future. while demand for voice services is expected to decline in the coming years. Table 1 shows that data traffic gradually increased by 93.2% between 2013 and 2017. This indicates that the demand for Internet services will increase in the near future. This is a challenge for telecom service providers to make decisions that can be used to avoid the risks of network congestion and maintain quality of service. Figure 4 shows the distribution of mobile traffic data over the last five years (2013-2017).

<table>
<thead>
<tr>
<th>Years</th>
<th>Average Data Traffic Annually (Mb)</th>
<th>Annual growth rate %</th>
<th>Average annual data traffic/user (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>2,338,745</td>
<td>-</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Table 1 - Data Traffic and Growth Rate on Mobile Network
The analysis of the highest and lowest congestion of the Libyan telecommunications company in Tripoli region is summarized below:

**Winter Season:** Due to national holidays and celebrations, February saw the highest traffic, contributing to the increased use of telecommunications and the Internet to transport events online, which will increase traffic this season. Although traffic is at its lowest in January, it may be because most students are busy for final exams, and not interested in the Internet.

**Spring Season:** The highest internet traffic flow was in May. The reason for this is that the big event of the Telecommunications and Information Technology Exhibition is held in Tripoli every May. While the lowest traffic occurred in April, an investigation is needed to determine the probable cause.

**Summer Season:** Students make up a large proportion of internet consumption. In August, most students take summer vacations, so they have the highest traffic of this month. In addition, the data analysis shows that the lowest internet traffic occurred in June. However, further investigations are needed to determine the correct probable cause. 

**Autumn Season:** Although the highest Internet traffic was in November and the lowest in September, it still needs to be examined for the right reasons.

Overall, the internet traffic data analysis shows that August has the highest internet traffic flow in the past five years (2013-2017) and the lowest traffic in June. Additionally, it should be noted that the average increase in per capita internet traffic increased gradually between 2013 and 2017, as shown in Figure 4 and Table 1. Internet usage in 2014 and 2015 was significantly slow. In 2017, the network was optimized and upgraded with the introduction of new HSPA technology. As a result, the average consumption per user jumped from 0.53 in 2013 to 1.5 in 2017. This means that there will be a strong demand for Internet services in the coming years. As a result, we need to develop a predictive mobile internet traffic model that can provide predictive capabilities for a variety of purposes, such as network capacity planning, customized traffic development, quality of service improvement, resource management, and making decisions that can be used to avoid risks of network congestion.

### 5.3 Features Selection

In machine learning and statistics, features selection, also known as variables selection to build the model. One important thing to keep in mind is trade between predictive accuracy and model interpretation. Because if we use a large number of variables, the predictive accuracy is likely to increase but the interpretation of the model will decrease. However, there are many ways to choose the model features, some treat it strictly as a work of art, others as a science, while some form of domain knowledge and a disciplined approach are probably the best solutions. In fact, there is an excellent method of selecting features, so-called stepwise, which is a combination of two wrapper methods for selecting by forward and backward selection techniques.

Based on our knowledge of the telecom domain and the mobile internet traffic low historical data, we have selected three main features of mobile network services data to model internet traffic forecasts as follows: mobile data traffic, mobile data traffic per user, and mobile users (subscribers). Based on the data from these main features, other variables were generated as: 5 days moving average data traffic, 10 days moving average (MA) data traffic, 15MA data traffic, users, 15MA Mobile users, traffic per user at \( t -1 \), 5MA consumption per user, 10MA consumption per user, 15MA consumption per user.
consumption per user, Data traffic at $t-1$, Data traffic at $t-2$, and Data traffic at $t-5$. Then, stepwise technique was applied. It shows that the best combination of features is (traffic per use)$_r$, users, and (Data traffic)$_r$. Therefor, three inputs are selected to be incorporated into the input layer. The daily observation of mobile traffic data provided from mobile service during January 01, 2013 to December 31, 2016 was used to build the model and denoted as group A, while the mobile traffic dataset during 2017 used for evaluating forecast accuracy and denoted as group B. However, traffic data of group A is randomly divided into training, validation, and testing sets to avoid over-fitting and improve network model generalization. The training set (68%) is used to update network weights and biases. The validation set (16%) is used to monitor network performance during the training process. Training can be stopped if the execution of the model results in a minimum error in the validation dataset. The testing dataset (16%) is used to fine tune the models. This is not used in training or validation, but to determine the optimal network architecture, select the appropriate model, and evaluate the performance (generalization) of the fully defined model.

In order to find an efficient network model and analyze the performance, the data must be pre-processed in order to have a small variation of the output data. It is well known that, in theory, the output data can have a large domain of variation. However, if the domain of variation of the output data is large, the ANN model tends to be less stable. The dimensionless output data ranges from 0 to 1, while the input data ranges from $-1$ to 1. Table 2 shows the results of the pre-process stage.

### 5.4 ANN Model Design

Typically, training a basic multilayer ANN boils down to minimizing some kind of error function. Commonly, the sum of squared errors is chosen as a function of this error which is often called objective function or cost function or loss function. The learning problem is presented as a search or optimization problem and an algorithm is used to navigate the space of possible sets of weights that the model can use to make good or sufficiently good predictions. To specify the number of hidden layers and the number of units in each layer, a trial and error approach was carried out, beginning with one hidden layer and one hidden unit. Hidden units were then gradually added. Based on the error function (Sum of squared error, SSE), several architectures were compared with different network parameters, three nodes in the input layer and one node in the output layer. Hidden layers are between 1 and 3, nodes of hidden layer 1 are between 1 and 10, nodes of hidden layer 2 are between 1 and 6, nodes of hidden layer 3 are between 1 and 4, and three different types of activation function (AF), called logical, hyperbolic tangent and linear. The nine best architectures are listed in Table 3. For the best selection, the nine networks were trained with seven different training algorithms; Quick Propagation, Conjugate Gradient Descent, Quasi-Newton, Limited Memory Quasi-Newton, Levenberg-Marquardt, Online backpropagation, and Batch back propagation. In each case, the architecture was retrained seven times with different initial weight randomizations. However, one method to avoid local minima is to retrain the network with a different weight initialization. This will locate the error function (SSE) at different minima in the error surface, and then choose the acceptable solution. However, the selected data traffic model is multilayered with three input nodes and one hidden layer with eight nodes, [3−8−1]. The network was trained using a Quasi-Newton training algorithm. The neurons in the backpropagation used a logistic function as an input activation function and linear function as an output activation function. The proposed traffic model provides low minimum errors in training, validation, testing, and overall dataset, as shown in Table 4. In addition, the statistical error analysis of the prediction capability obtained from the proposed ANN model is summarized in Table 5. Obviously, the proposed ANN model provides traffic forecast values with $R^2$ of 97.5% and an average error of 706 Mb, which indicates that the model describes the data very well.

### Table 2 - Results of The Pre-Process Stage

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Column type</th>
<th>(internet traffic)t-1</th>
<th>Users</th>
<th>(traffic /user)t-1</th>
<th>internet traffic (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>input</td>
<td>Input</td>
<td>input</td>
<td></td>
<td>Output</td>
</tr>
<tr>
<td>Scaling range</td>
<td>[-1..1]</td>
<td>[-1..1]</td>
<td>[-1..1]</td>
<td>[0..1]</td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>1042994.76</td>
<td>2316830.3</td>
<td>0.413134</td>
<td>1042994.76</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>3499321.4</td>
<td>2695701.4</td>
<td>1.353166</td>
<td>3499321.4</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2560315.1</td>
<td>0.835849</td>
<td>2143037.16</td>
<td>2141794.90</td>
<td></td>
</tr>
<tr>
<td>Std. deviation</td>
<td>68391.39</td>
<td>0.209051</td>
<td>543270.14</td>
<td>543602.59</td>
<td></td>
</tr>
<tr>
<td>Scaling factor</td>
<td>0.000005</td>
<td>2.127586</td>
<td>4.07E-07</td>
<td>8.14E-07</td>
<td></td>
</tr>
</tbody>
</table>
Moreover, Figures 5, 6 and 7 show the scatter plots of actual traffic versus predicted model output for training, evaluation and overall data points respectively. The Figures show all the points are scattered around the \( y = x \) line, indicating excellent agreement between the actual and estimated data traffic points.

### 6. Mobile Traffic Forecast for 2017

As mentioned above, daily monitoring of mobile traffic data over the period 2013-2016 (Group A) served to build the model, while the 2017 Mobile Traffic Data Set (Group B) used for evaluating forecast accuracy.

#### 6.1 Quarterly and Semi-Annual Mobile Traffic forecasts

In the quarterly forecast, the input variables represent the historical data traffic recorded in Q4 2016 and applied to the model to forecast Q1 2017 mobile traffic. Similarly, for Q2 2017 mobile data traffic forecasts, the inputs are Q1 2017 predicted data traffic. Likewise, in the semi-annual forecast, the input variables represent data traffic recorded in the second half of 2016 and are applied to the model to forecast mobile data points for the first half of 2017. Table 6 presents the quarterly and semi-annual mobile traffic forecasts. The results of the forecasting error analysis show that the quarterly forecasting scenarios performed better than the semi-annual forecasting scenarios with the lowest average error of 54402 Mb and higher \( R^2 \) of 91%. Other related statistical error analysis is presented in Table 7. Because the purpose of this study is to develop an accurate model to estimate the demand for data traffic on the Libyan mobile network. This model enables resource management to allocate optimal capacity resources, better service quality and cost savings.
Proceedings of the International Conference on Industrial Engineering and Operations Management
Dubai, UAE, March 10-12, 2020

Fig. 5 – Scattered plots of Actual Vs. Model Output traffic; Training Dataset

Fig. 6 – Scattered Plots of Actual Vs. Model Output Traffic; Validation Dataset

Fig. 7 - Scattered Plots of Actual vs. Model Output Traffic; Overall Points of Dataset Group A

Table 6 - Quarterly and Semi-Annual Mobile Traffic Forecasts

<table>
<thead>
<tr>
<th>Quarter</th>
<th>(internet traffic)_(t-1)</th>
<th>Users</th>
<th>(traffic/user)_(t-1)</th>
<th>Traffic Forecast</th>
<th>Actual Traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 2017</td>
<td>2952852</td>
<td>2492787</td>
<td>1.17</td>
<td>2921850</td>
<td>3033873</td>
</tr>
<tr>
<td>Q2 2017</td>
<td>3031088</td>
<td>2514026</td>
<td>1.22</td>
<td>3015280</td>
<td>3333114</td>
</tr>
</tbody>
</table>
7. Conclusions

This study provides an overview of data mining in telecommunications and proposes a novel model for forecasting mobile internet traffic based on artificial neural networks and draws the following conclusions:

- Previous research has shown that telecommunication datasets used in data mining can be classified into; call details, network and customer data.
- The applications of telecommunications data mining fall into three categories; customer relationship management, fraud detection, and network management. Most of these studies were devoted to customer relationship management, such as; customer value and retain profitable customer, generate customer profitable and mine these for marketing purposes, acquire new customer, maximize profit, and churn analysis which includes churn prediction and churn management.
- Application of ANNs to traffic prediction show that ANN, RBFNN and GMDH-PNN can predict traffic with very high accuracy.
- Many previous telecommunication data mining research has employed time series forecasting techniques. Most of these studies did not specify the name of the input features. Only the number of input nodes was defined.
- Mobile voice and internet data was recorded as a daily traffic monitor from January 1, 2013 to December 31, 2017 on the Libyan real mobile network, which covers the Tripoli region. Between 2013 and 2017, it was noticed that the demand for voice services increased insignificantly, while the demand for Internet services increased significantly.
- The trend mining of Internet traffic during the past five years (2013-2017) shows that traffic is highest in August and lowest in June.
- The daily observation of mobile traffic data provided from mobile service during January 01, 2013 to December 31, 2016 (group A) was used to build the model, while the dataset during 2017 (group B) used for evaluating forecast accuracy.
- Data of group A is randomly divided into training, validation, and testing sets to avoid over-fitting and improve network model generalization. The training set (68%) is used to update network weights and biases. The validation set (16%) is used to monitor network performance during the training process.
- The proposed ANN forecasting model has three layers with eight nodes in the hidden layer. The network was trained using a Quasi-Newton training algorithm.
- In features selection, stepwise technique shows that the best combination of features is (traffic per use)\(_t-1\), users, and (Data traffic)\(_t-1\).
- The neurons in the backpropagation used a logistic function as an input activation function and linear function as an output activation function.
- The proposed traffic model provides low minimum errors in training, validation, testing, and overall dataset. It provides traffic forecast values with 97.5% R\(^2\) and an average error of 706 Mb, which indicates that the model describes the data very well.
- The results of the quarterly and semi-annual forecasts of mobile traffic show that the quarterly forecast performed better than the semi-annual forecast, with its lowest average error of 54402 Mb and higher correlation coefficient of 91% R\(^2\).

References


KDD- 89 <https://www.kdnuggets.com/meetings-past/kdd89/index.html> Copyright © 2000 KDNuggets


© IEOM Society International 3084
Proceedings of the International Conference on Industrial Engineering and Operations Management
Dubai, UAE, March 10-12, 2020


Rey, Tim, et al. (2012) Applied Data Mining for Forecasting Using SAS. SAS Institute Copy bibliography citation


**Biographies**

**Saber Kh. Elmabrouk** is a talented veteran with over 20 years of experience teaching and researching in Modeling, Artificial Intelligence and Data Mining, Big Data, Information System Management, Uncertainty analysis, and Risk Management. He received his Ph.D. in Petroleum Engineering from Regina University in Canada. Prior to his Ph.D., he had earned his Master and Bachelor degrees in Petroleum Engineering from the University of Tripoli, Libya. He is currently department head of Chemical and Petroleum Engineering at Libyan Academy, Janzour, Libya. In addition, he is an adjunct faculty with department of Engineering Project Management at Faculty of Engineering, University of Tripoli, Libya.

**Alhasan Nureddin Salem** holds a M.Sc. in Engineering Management from Libyan Academy, Janzour in 2019 and a B.Sc. degree in electrical and electronic engineering from University of Zawia, Libya in 2001. Mr. Alhasan began his career at ABB in 2002 as an energy management engineer. Subsequently, in 2005, he worked for Libyana Mobile Phone Company, where he began working as a wireless radio engineer, then supervised the mobile installation phase of GSM Mobile projects from 2005 to 2007. He is then a mobile switching engineer from 2008 to 2009, and in 2010-2012 worked as an expert in wireless network quality in the field of performance management. He is a radio network planning manager in telecommunications projects for the period 2013-2015 and a director of management for 3G and 4G mobile technologies for the period 2016-2017. Currently Mr. Alhasn is an expert in core network strategy and planning in Libyana Mobile Phone Company.