[1]Sandhya B S

[2]Rohini Deshpande

Detection and analysis of cellular network traffic anomalies and SMS spammers

Journal of Electrical Systems

Abstract: **-** The unfolding usage of mobile users inclusive of both 4G and 5G creates huge accumulation of data in cellular network. The network service providers need to ensure proper management of resource in terms of uninterrupted service with cost-effectiveness. The detection of cellular traffic and Short Message Service (SMS) spammers is very challenging. In this paper a novel method is proposed to analyse and detect the traffic anomalies and SMS spammers. To achieve this, Call Detail Record (CDR) issued by service provider is used. The CDR is pre-processed to convert into machine understandable format using mean-normalization technique. K-means clustering elbow method proves to be the best tool in identifying the traffic clusters in the network that detects both high and low traffic in the network. The novelty of the proposed work is the detection of low traffic clusters which usually is misled as sleeping cell or cell outage. The paper also presents a model designed to predict whether the message is spam SMS or ham SMS. The proposed model is suitable to run different classifiers like Logistic Regression, Multi nominal Naive Bayes, Support Vector Machine (SVM), Random Forest Classifier. The model gives the highest accuracy rate of 98.277% with SVM in detecting SMS spam.

Keywords: Call Detail Record, machine learning, cellular network, K-means clustering, elbow method.

I. INTRODUCTION

With the growing population of cell phone users and advent of smart technologies there is a huge accumulation of data in the cellular network. The timely and effective analysis of this huge data is very much necessary for the network resource management and for better customer service. The details of this big data are available in the cellular network department which is known as Call Detail Record (CDR) and is used by only the internal operators for the required analysis. The CDR includes all the transaction details of the events occurring in the cellular network. The events recorded in the CDR are incoming calls, outgoing calls, incoming SMS, outgoing SMS and internet activity. The beginning and the end of every event is time stamped in the CDR.

The formal purpose for any cellular network is to provide wireless telephone and internet services to the customers with good coverage and better cost. It does not aim to detect cellular phones. However, the huge data accumulated in the cellular network can be used to detect the location of the cell phone users when required [1]. For the proper analysis of the CDR substantial volume of data must be gathered in the network including the location details [2]. The objective of CDR analysis is to account every activity of the customer which is generated by the cellular network and ensure non interrupted service [3]. The challenging part is transfer and storage of the CDR in the cellular network and communication service providers centers [4]. Prior to any analysis of CDR, the raw form of it has to be converted into processed or machine understandable form [5]. This is called pre-processing which is a vital step in the big data analytics. While pre-processing, the attributes of CDR must be considered. They are volume, velocity, variety and value [6].

Once when the data is ready in a processed form any required analysis can be performed. The analysis could be the detection of traffic activities, SMS spammers, fault detection etc. This is essential for the better network performance to improve the capacity and reliability nevertheless the availability of bandwidth. The analysis is also important to decrease the latency and the network failures [7].

Analysis of the processed CDR data with the suitable application of machine learning tool gives better outcomes with the less human intervention. However, it is important to know about the machine learning tools and methods

[1] *Sandhya B S: REVA University, Bengaluru, Karnataka

School of Electronics and Communication Engineering

[2]Rohini Deshpande: REVA University, Bengaluru, Karnataka

School of Electronics and Communication Engineering

before applying to the CDR data. As per the requirement of the analysis the suitable tool can be chosen to get the expected results. Hence it is preferred to understand about the machine learning and its application to communication networks.

Machine learning is an approach of replacing base knowledge acquisition by using many examples of required behavior to train the designed algorithm. The algorithm should perform well with the selected machine learning tool matching the data to be analyzed. The three main categories of machine learning methods are supervised, unsupervised and reinforcement learning. Supervised machine learning consists of labelled input data sets with output mapping to train the algorithm to get the desired output. Unsupervised machine learning has only unlabeled input data sets without mapping of desired output sets. Reinforcement learning is between supervised and unsupervised learning. In supervised learning there are many tools like Naive Bayes, Linear Regression, Logistic Regression, Random Forest, Support Vector Machine (SVM) etc. Unsupervised learning has tools like Principal Component Analysis (PCA), Independent Component Analysis (ICA) and K-means clustering [8].

However, considering the problem of traffic analysis in the cellular network an unsupervised K-means clustering elbow technique is proposed to identify the high and low traffic cluster. Also, a model suitable to run different classifiers is proposed to classify spam and non-spam messages. The paper is organized in the following fashion. Section II gives related works, section III explains the methodology to detect traffic anomalies and SMS spammers, Analysis with results is discussed in section IV and the paper is concluded with future work in section V.

II. RELATED WORKS

Md Salik Parwez et al. [9] recommended that the pre-processed data must be the combined events including calls, SMS etc. It also includes the time stamp at every grid.

Bilal Hussain et al. [10] proposed an approach of data pre-processing with the removal of irrelevant parameters. The empty fields in the CDR data are filled. The events that occurred in the network are summed up and transformed into machine understandable form.

Ramin Sharif et al. [11] proposed that pre-processing using Horton works sandbox. All the empty fields are set to zero and all the CDR events are totaled into one single event. Also, the time stamps are summed to 60 minutes for better memory and processing power.

Kashif Sultan et al. [12] stated that data has to be pre-processed to remove irregularities, misleading patterns and noise which makes the data ready for further analysis.

Pekka Kumpulainen et al. [13] have used unsupervised methods for the discretization of the two types of variables in the telecommunication network. The technique suggested is amplitude distribution of the data variables.

Anish Nediyanchath et al. [14] proposed SARIMA model which is a seasonal ARIMA (Auto Regressive Integrated Moving Average) model to detect the anomalies which supports seasonality.

Zhang C et al. [15] conducted experiments on cellular data set to design multi-scale convolutional recurrent encoder decoder network for obtaining correct window size for the detection of anomaly intensity.

Kashif Sultan et al. [16] have mentioned that information in the CDRs of cellular networks gives the details of operational efficiency and behavioral patterns of mobile subscribers. Spatiotemporal analysis is used to bring out the customers behavior markings in the network. The work proves that the real network traffic patterns can be detected and classified using a model based on machine learning technique.

Yaohua Sun et al. [17] have stated that there are many open issues yet to be resolved in the cellular network with the application of machine learning tools. The areas which need rigorous research are network resource management and mobility management.

Serkan Balli et al. [18] proposed a content-based classification solution to prevent spam SMS. The method filters out the unwanted message and uses Word2vec neural network to predict neighboring words.

Dea Delvia Arifin et al. [19] recommended SMS spam filtering performance by combining two of data mining task association and classification. FP-growth is used for mining frequent pattern on SMS.

Jialin Ma et al. [20] proposed a Message Topic Model (MTM) for SMS spam filtering. It addresses sparsity problem in SMS classification and symbol semantics is taken into account.

Nan Jiang et al. [21] have designed an algorithm to detect related spam numbers. It identifies additional spam numbers with similar SMS patterns during the same time period and same network location.

III. METHODOLOGY

The huge accumulation of data in the cellular network has to be analyzed for proper network resource management. With reference to this, a method is proposed to analyze the traffic anomalies in the network and identify the SMS spammers. Hence the proposed work is presented in two parts which are explained in upcoming sections *A* and *B* in detail. In the first section *A*, a technique is proposed to detect traffic anomalies in the network using K-means clustering elbow method and in the second section *B*, a model is proposed suitable to run different classifiers to identify SMS spammers.

A. Detection of traffic anomalies using K-means clustering elbow method

Initially the data set need to be preprocessed to make it ready for further required analysis. Call Detail Record (CDR) is the set of the mobile users' details produced by the cellular network operator. The CDR issued by Telecom, Italia in Milan, Trentino is used in the work here. The grid type CDR data is of the area of Milan grid and Trentino grid of 1000 squares and 6575 squares respectively, each with an area of about 235 square meters [6]. The CDR data set contains 8 numerical features. They are call in activity, call out activity, SMS in activity, SMS out activity, Internet traffic activity, square grid ID, country code, time stamp information.

There are many challenges to be considered in pre-processing the big data. The CDR data collected generally have few partially filled fields and unwanted noise. It includes irrelevant personal information. This has to be removed by data cleaning and pre-processing [22]. Once the unwanted data is removed, the accuracy and the reliability will be improved [23].

There are few fundamental steps which must be accommodated in pre-processing the CDR data. It is not compulsory to implement all the steps while pre-processing. However, as per the requirement the steps can be chosen [24].

The primary steps are:

1) Quality testing of the CDR data set

In this step the CDR data is checked for human errors like missing values and irrelevant values. For instance, irrelevant entries of values in the wrong address fields. The quality testing must assure the removal of such human errors.

2) Grouping of similar characteristic features in the CDR data set

This step helps in the better accounting of features in timely manner with less complexity. It requires very smaller memory and hence the processing time decreases.

3) Sampling of the CDR data set

Sampling of the CDR data set is very important step in pre-processing. The sample to be chosen should be illustrative of the entire data set. The sample must have all the characteristics of the whole data set while considering cost, memory and time. The popular sampling method used in pre-processing is random sampling [25]. Computational time and cost factors to be examined in the sampling process [26]. Sampling algorithm must be designed to extract the training set and calculate the important variables [27], [28].

4) Feature reduction of the CDR data set

It is the effective step in the pre-processing to check for the features to be retained and removed in the CDR data set. It helps in the removal of noise factors and provides finer visualization of the CDR data set for future analysis with less complexity. Feature reduction is emanated from Rough Set Theory. The features are reduced using Quick reduct [29].

5) Encoding of the CDR data set

This is the final step where in features of the CDR data set are to be converted into machine understandable form while not hampering the originality. Encoding enables us to understand the features in the data sets and the trade-offs required practically [30].

Depending on the relationship among the variables there are two categories of encoding. They are nominal encoding and ordinal encoding. When the variables are independent of each other then it is nominal encoding. When the variables are dependent on each other then it is ordinal encoding.

After preprocessing, the data set is ready for the further analysis. The analysis includes the detection of traffic patterns in the cellular network. K – means clustering is a popular unsupervised technique to recognize patterns. The algorithm finds optimal number of clusters in the data sets without any initialization and selection of parameters [31]. It has many types of methods to cluster the data points. However, the suitable method of clustering to be chosen according to the requirement of the problem and the data sets.

K-means clustering algorithm considers all the data points in the data sets with same priority. To reduce the feature components, entropy method is used. The entropy K-means clustering technique requires a smaller number of computations [32].

In the case when the target points are unknown, elbow method is considered. This method finds correct number of clusters [33]. It calculates distortion or inertia value for visualizing the data before clustering.

K-means clustering algorithm can also be accomplished in a semi – supervised way where finite number of dynamic changes are available. This method ensembles the semi – supervised cluster and reworks only on the consistencies of the clustering. It reduces the computational complexity and saves time [34].

The steps involved in the detection of traffic in the cellular network are as follows.

Step 1: The raw CDR data is preprocessed by normalizing empty fields by the mean values.

Step 2: The optimal number of clusters ‘k’ of mobile activities is determined by using k-means clustering algorithm.

Step 3: The number of clusters are determined such that the inertia value becomes constant without average distortion.

Step 4: The technique used to determine the optimal value of k is verified by Elbow method.

B. Prediction of SMS spammers using classifiers

Spam SMS are the unwanted messages from untrusted sources which lead to fraudulent sites which reveals personal information of individuals such as passwords or credit card No.’s etc. In most parts of the countries across the world up to 30-45% of messages were spam in 2011, in 2019 it raised to 50-65% and in 2022 it increased 30% more. Eventually Spam SMS accounts for half of all the mobile phone network traffic from 2019. The huge congestion caused due to spam SMS degrades the network performance. Hence a system is proposed which gives the end user whether the received SMS is spam or ham. A model is suitable to run different classifiers which can analyse the SMS features. The classifier is trained in a way that for every SMS message in the data set collected, every word is checked whether it is associated with spam message and categorized.

The data set used in our work is SMS Spam Corpus v.0.1 Big having 5574 messages in English. The suitable classifiers proposed for the detection and analysis of spam SMS are Logistic Regression, Multi nominal Naïve Bayes, Support Vector Machine and Random Forest.

Logistic regression is a predictive tool used in classifying the highest probability of entities. The tool is inherited from machine learning which establish the relationship between the binary depend variable and one or more independent variables.

Multi Nominal Naïve Bayes classifier is an easy and efficient computational tool used in problems of text classification. The tool is based on Bayes algorithm and a relationship is established between independent entities.

Support Vector Machine is a supervised tool used in distinguishing the planes of entities. The tool is suitable for complex data sets.

Random Forest is a machine learning tool which gives the consolidated output of multiple decision trees to solve classification problems. The tool is suitable for the datasets having continuous variables.

The steps involved in the detection of spam SMS are as follows.

Step 1: Data collection

The data set used is SMS Spam Corpus v.0.1 Big. It contains one set of SMS messages in English of 5,574 messages.

Step 2: Data cleaning

Data cleaning is the process of editing, correcting, and structuring data in a data set to make the content uniform and ready for analysis. Data cleaning removes irrelevant data and formats it into a machine understandable language for optimal analysis.

Step 3: Generation of training and testing sets

Data set is divided into 2 sets, 70% as training set and 30% as testing set. Training and testing data sets help to identify whether the proposed model is overfit or underfit.

Step 4: Generation of Term Frequency – Inverse Document Frequency (tfidf) vectorizer

Tfidf vectorizer converts a collection of raw SMS document into a term document matrix. Tfidf vectorization involves transforming every word of SMS into a score relative to that document and then creating the same information into a vector.

IV. Analysis, results and discussions

A. Analysis of traffic anomalies using K-means clustering elbow method

The CDR data which is in the unprocessed form is about 300MB which is very huge. Therefore, in the proposed work it is preferable to sample the data in the initial stage. With this regard the size is reduced to 8KB by sampling with sampling rate N = 100. Pre-processing of 5 such CDR data sets of different dates are conducted and the simulation results are analyzed.

The Table 1 shows the usability of the 8 numerical features of incoming call activity, outgoing call activity, internet activity, country code, square grid ID, time stamp information, incoming SMS activity and outgoing SMS activity in the CDR while preprocessing.

Table 1 Numerical features of CDR data set with the usability factor

Sl. No.	**CDR data set features**	**Usability factor (retained/removed)**
1	Incoming call activity	Retained
2	Outgoing call activity	Retained

3	Internet activity	Retained
4	Square grid ID	Removed
5	Country code	Retained
6	Time stamp information	Removed
7	Incoming SMS activity	Retained
8	Outgoing SMS activity	Retained

The pre-processing is carried out using mean-normalization technique. Therefore, mean values of all the retained parameters are evaluated and the missing values are normalized. The Table 2 shows mean values determined for the retained features.

Table 2 Mean values of the numerical features of CDR data sets

Date 2013	01/07	02/07	03/07	04/07	05/07
Incoming call activity	0.1013	0.1077	0.0708	0.2160	0.1816
Outgoing call activity	0.0914	0.0880	0.0823	0.1670	0.1343
Incoming SMS activity	0.1411	0.1305	0.1450	0.2328	0.1587
Outgoing SMS activity	0.1176	0.1205	0.1194	0.2069	0.1230
Internet activity	6.9103	5.8327	5.6055	5.8375	6.2851

The analysis of traffic in the cellular network necessitates the detection of the unusual behavior in the network like high traffic, low traffic. The high traffic may be due to large number of users. The low traffic may be due to non-usage of mobile by the users due to cell outage though there are ample number of users. Hence, it is very important to depict both high and low traffic in the cellular network for the network resource management.

Cellular network traffic detection is the realization of different traffic types by analyzing the CDR data set. It is the prime concern for the network management functions like ensuring the network quality-of-service (QOS) and cost management [35], [36]. There are three methods in the detection of traffic. They are port based, payload based and machine learning based methods. The port based is primeval method which uses port numbers from the protocol headers of packets to detect traffic [37]. The payload-based method uses payload packets with pre-defined patterns to analyze traffic in the network [38]. The recent and popular method of analyzing cellular traffic with less or no human intervention is machine learning based method.

In the proposed work to analyze the cellular traffic created by the mobile users, K-means clustering elbow technique is used to detect the high traffic activity and the low traffic activity. Depending on the distribution of the CDR data set the optimal number of clusters is determined [39].

The scattering details of the mobile users' activities are obtained by pre-processing the raw CDRs. The simulations of 5 sets of different dates are obtained. Fig. 1 – Fig. 5 are the graphs showing the distribution of the mobile users for the CDR activities dated 01/11/2013 considering all the country codes given. in the data set. In Fig. 1 the graph displays the distribution of the incoming call activity by the users across the country codes available in the data set. It records variations from the lowest value 0.0273 to the highest value 0.545103 for the incoming call activity. Fig. 2 gives the distribution of the outgoing call activity by the users. Fig. 3 and Fig. 4 records the distribution of the incoming and outgoing SMS activity of the users. Fig. 5 gives the distribution of internet activity of the mobile users.

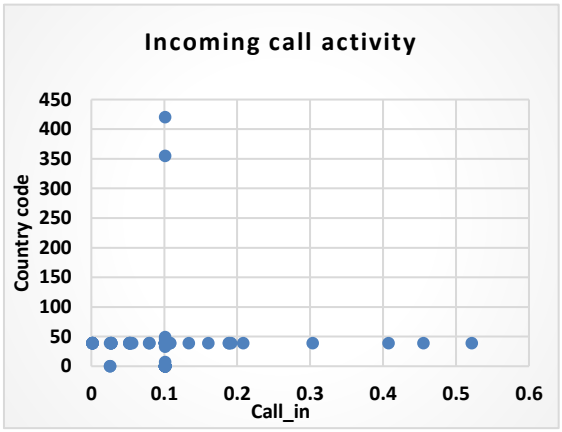

Fig. 1 Incoming calls activity on 01/11/2013

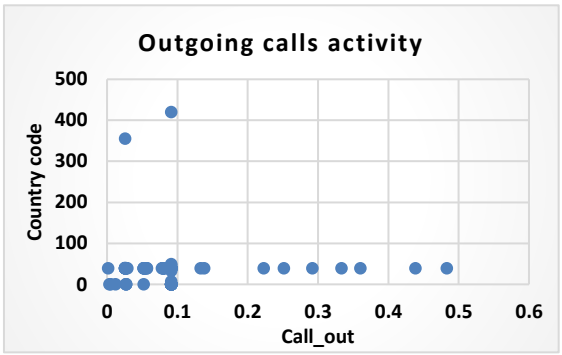

Fig. 2 Outgoing calls activity on 01/11/2013

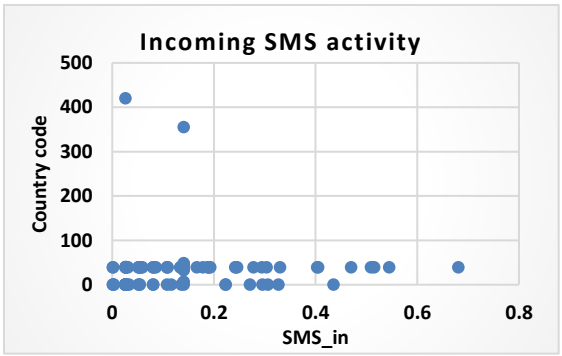

Fig. 3 Incoming SMS activity on 01/11/2013

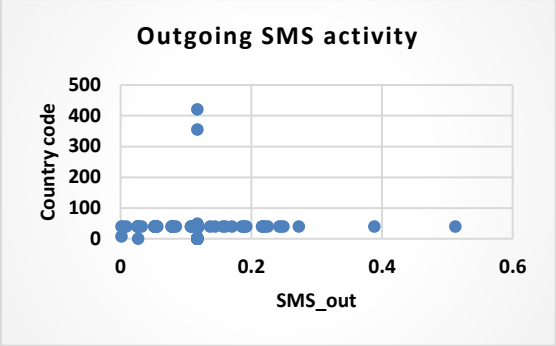

Fig. 4 Outgoing SMS activity on 01/11/2013

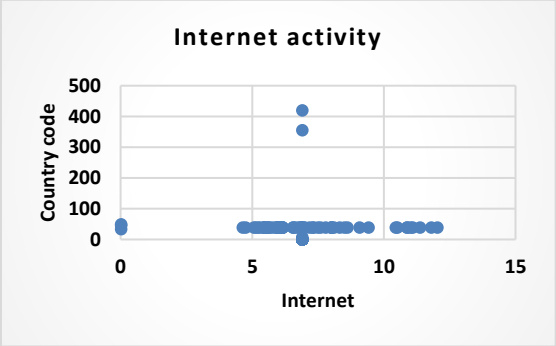

Fig. 5 Internet activity on 01/11/2013

For more clarity, the following simulations are carried out considering one country code 39 with 5 different dated CDR data sets. These simulations show the total number of records of each activity for 5 different dates. These results account the highest and lowest records of each activity. As an illustration, Fig. 6 depicts the highest record of incoming call 01-Nov-2013 and lowest on 03-Nov-2013. From the simulations, it can be observed that the highest call-in activity and call-out activity is on 01/11/2013. Also, highest SMS-in activity and SMS-out activity is on the same day. It reports highest internet activity too the same day. Thus, with the suitable pre-processing method, the CDR data set can be transformed into processed form and later appropriate machine learning tool can be applied to track the mobile users' activities in an advanced level.

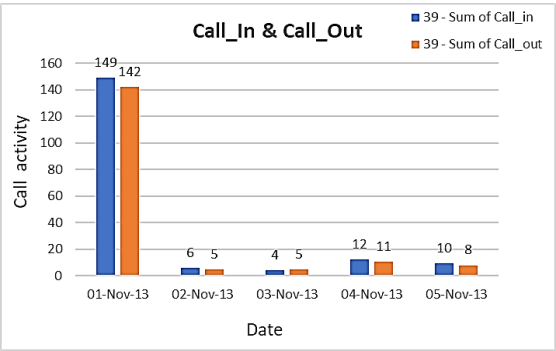

Fig. 6 Calls activity of country code 39 for 5 days

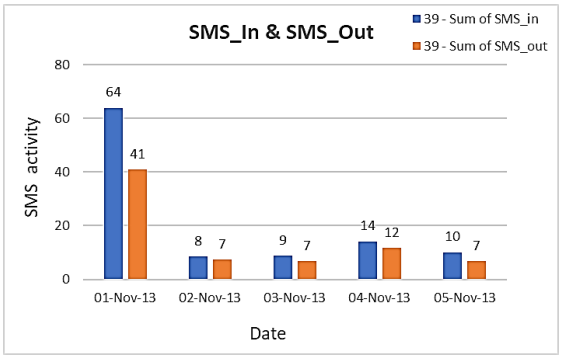

Fig. 7 SMS activity of country code 39 for 5 days

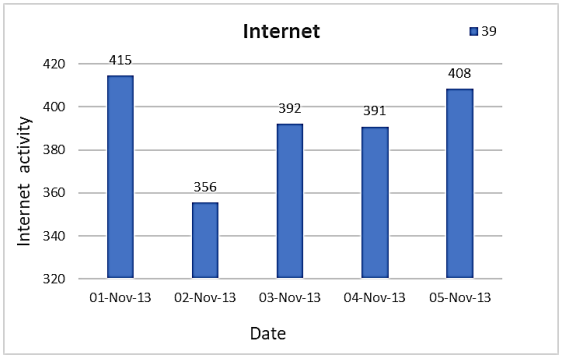

Fig. 8 Internet activity of country code 39 for 5 days

Further, it is essential to determine the optimal number of clusters to understand the traffic in the cellular network due to the mobile users. In this regard K-means clustering tool is used to find out the desirable number of clusters. The method used to achieve the right optimal value k is called elbow method. The number of clusters is chosen such that the inertia value becomes constant. Later, if the value of 'k' is increased it is unlikely that the average distortion is also increased. At this point, dividing the data into further clusters need to be stopped. If the plot of number of clusters versus inertia or sum of squared error gives an arm shaped graph then it verifies that the number of clusters chosen is optimum.

Fig. 9 to Fig. 13 are the clustering simulations obtained for the different CDR activities showing both high and low traffic intensities. For example, in Fig. 9 three clusters of traffic are obtained for the incoming call activity. Specifically, at country code 39 bigger cluster is formed with highest and lowest traffic records. Fig. 10, Fig. 11, Fig.12 show the three clusters of traffic obtained for outgoing call activity, incoming SMS activity and outgoing SMS activity respectively. Fig. 13 show four clusters of traffic for the internet activity.

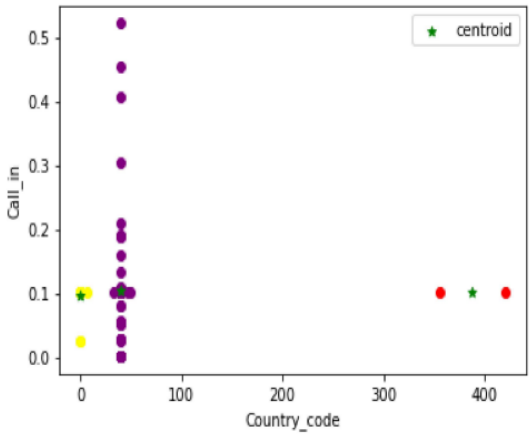

Fig. 9 Clustering for the incoming call activity

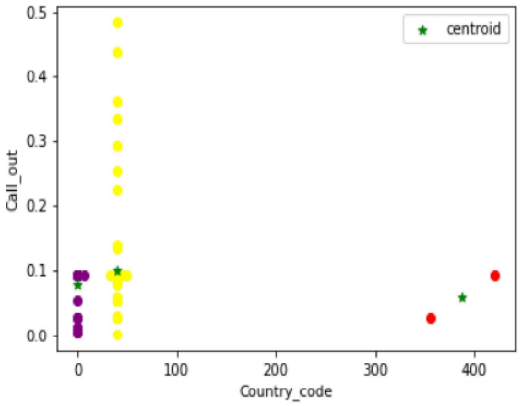

Fig. 10 Clustering for the outgoing call activity

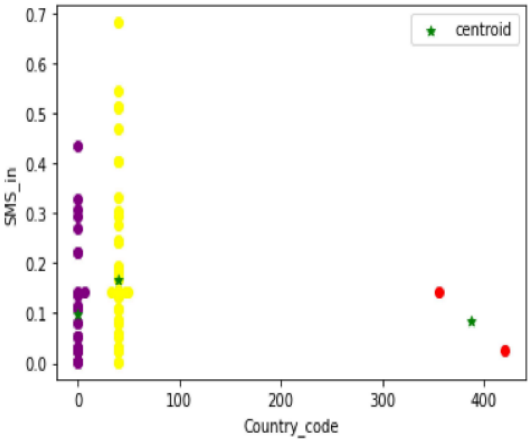

Fig. 11 Clustering for the incoming SMS activity

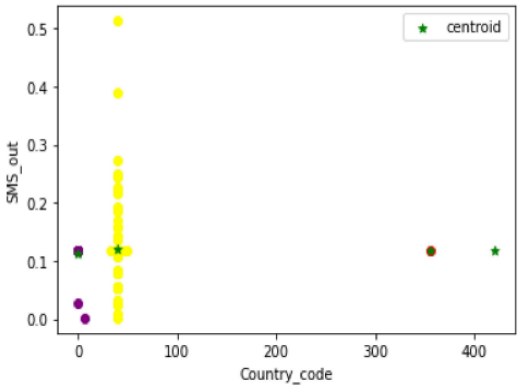

Fig. 12 Clustering for the outgoing SMS activity

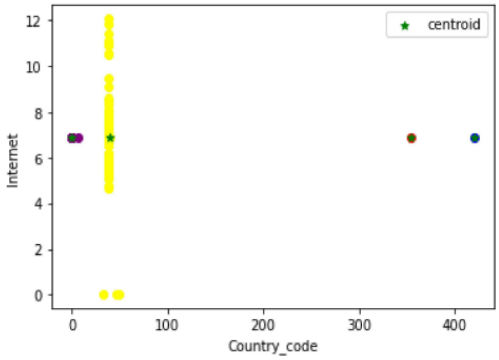

Fig. 13 Clustering for the internet activity

The Table 3 shows the number of clusters obtained for all the CDR activities.

Table 3 Number of clusters for all the CDR activities

CDR activity	**Number of clusters**
Incoming call	3
Outgoing call	3
Incoming SMS	3
Outgoing SMS	3
Internet	4

The following plot of arm shape in Fig. 14 verifies the optimal number of clusters obtained for all the activities.

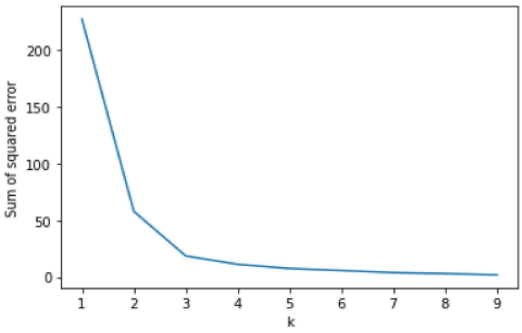

Fig. 14 Verification of the K-means clustering using elbow method

For comprehensive understanding, clustering for one single activity is carried out and for one particular country code. As an example, Fig. 15 indicates the clustering carried out for internet activity considering country code 39. The cluster in yellow color represents lower traffic, blue and purple color clusters represent moderate traffic and the red color cluster represents higher traffic activity.

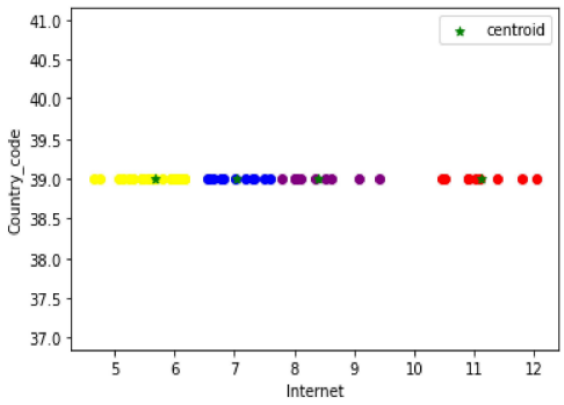

Fig. 15 K-means clustering for the internet activity of country code 39

The proposed K-means clustering elbow technique is experimentally compared with the following tools.

- Principal Component Analysis (PCA) with Random Forest (RF) regressor.
- Independent Component Analysis (ICA).

The comparison is shown in Table 4.

Table 4 Comparison of K-means clustering elbow technique with other machine learning tools

Tools/ Evaluation factors	**K-means clustering elbow technique**	**PCA with RF regressor**	**ICA**
Detection of traffic	Detects both high and low traffic	Detects only high traffic	Detects both high and low traffic
Detection of individual CDR activities	Detects the individual activities and groups them	Do not detect individual activities but only gives the correlation among them	Focuses on maximizing individual data points rather than grouping the data points of same activity
Complexity	Less complex	Very complex	Less complex
Cost	Cost effective as it is an unsupervised tool	Expensive as it is the combination of both unsupervised and supervised tool	Cost effective as it is an unsupervised tool

In the comparison analysis the key evaluation factors considered are detection of type of traffic, detection of individual CDR activities, complexity of the tool and cost.

The key comparison results deduced are as follows.

- PCA with RF regressor detects only high traffic. ICA and K-means clustering elbow technique detects both high and low traffic.
- K-means clustering elbow technique detects individual CDR activities. ICA detects individual CDR activities but will not form traffic clusters as required. PCA with RF regressor do not detect individual CDR activities.
- ICA and K-means clustering elbow technique are less complex and cost effective whereas PCA with RF regressor is very complex and costly.

Considering all these factors, it is evident that the proposed K-means clustering elbow technique is better than the other machine learning tools PCA with RF regressor and ICA in terms of required evaluation factors.

B. Analysis of SMS spammers using classifiers

The system model proposed to classify the spam and ham messages are run by suitable classifiers. The different classifiers used are like Logistic Regression, Multi nominal Naive Bayes, Support Vector Machine, Random Forest Classifier. The accuracy rates of prediction of spam SMS are determined for all the classifiers as shown in Table 5.

Table 5 Classifiers and its accuracy rate to predict spam SMS

Classifier	**Accuracy rate of spam detection (in %)**
Logistic Regression	96.841
Multi Nominal Naïve Bayes	96.697
Support Vector Machine	98.277
Random Forest classifier	97.343

Comparatively Support Vector machine and Random Forest Classifier are having better prediction rates i.e., 98.277% and 97.343% respectively. The Fig. 16 and Fig. 17 show snippets of SVM classifier Random Forest classifier identifying whether the SMS message is spam or ham.

```
In [135]: clf.predict(["Hi, this is Sandhya"])
Out[135]: array(['ham'], dtype=object)

In [147]: clf.predict(["Congratulations!, text 'WON' for free AIR tickets to the USA"])
Out[147]: array(['spam'], dtype=object)
```

Fig. 16 SVM classification of spam and ham SMS

```
In [145]: clf.predict(["Hi, How are you?"])
Out[145]: array(['ham'], dtype=object)

In [152]: clf.predict(["Congratulations!, text 'WON' for free AIR tickets to the USA"])
Out[152]: array(['spam'], dtype=object)
```

Fig. 17 Random Forest classification of spam and ham SMS

Further considering with the analysis of detection spam SMS, the length of each message and the punctuation marks can be accounted. The analysis helps in understanding the threshold values of the length and punctuations to predict if the message is spam or ham in the initial phase. The Fig. 18 shows the snippet where the length and punctuations of messages are calculated using the system model with the plots of the same shown in Fig. 19 and Fig. 20. The strokes in blue colour represent ham and in orange colour represents spam.

Out[11]:

	label	message	length	punct
0	ham	Go until jurong point, crazy.. Available only ...	111	9
1	ham	Ok lar... Joking wif u oni...	29	6
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155	6
3	ham	U dun say so early hor... U c already then say...	49	6
4	ham	Nah I don't think he goes to usf, he lives aro...	61	2

Fig. 18 Length and punctuation of spam and ham SMS

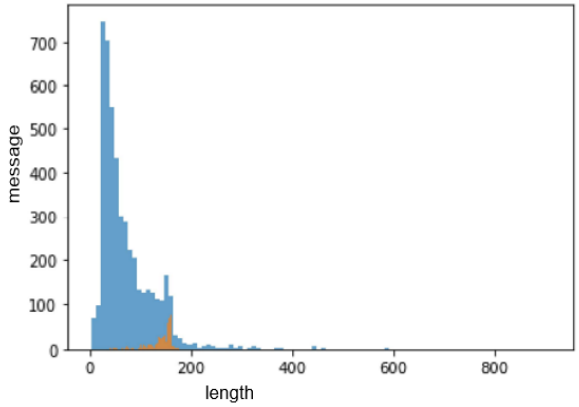

Fig. 19 Plot of length of ham and spam SMS

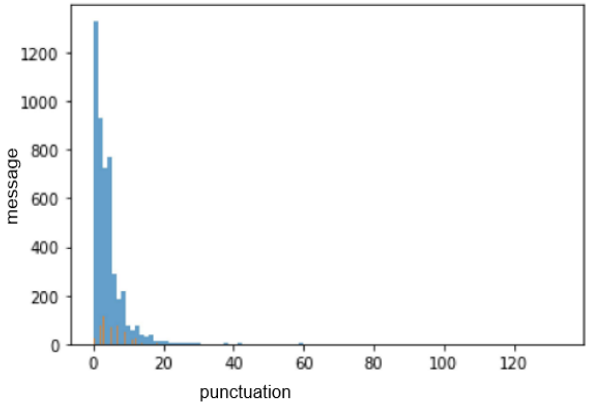

Fig. 20 Plot of punctuation of ham and spam SMS

V. CONCLUSION AND FUTURE WORK

This paper presents the pre-processing of CDR data set using mean-normalization method, the detection of traffic in the cellular network using K-means clustering elbow technique and prediction of spam SMS. As a result, the total activities of incoming calls, outgoing calls, incoming SMS, outgoing SMS and internet are obtained for different country codes and also observed for one individual country code. The pre-processed data can be used for the advanced analysis to understand the mobile user's behavior and to detect the traffic in cellular network caused by the different activities of the mobile users. The cellular network traffic is analyzed using K-means clustering elbow technique. The analysis also helps to understand the density of the network traffic for each activity. The novel work proposed depicts the detection of both high and low traffic in the cellular network created by the mobile users. The paper also presents a system model to predict spam SMS. The proposed model is run by different classifiers Logistic Regression, Multi nominal Naive Bayes, Support Vector Machine, Random Forest Classifier. The proposed spam detection classifier gives the accuracy rate of 98.277% in detecting Spam SMS. This benefits both the cellular network service providers and the users in terms of resource and cost management.

The future work can be extended to understand the actual cause for low traffic activity. Furthermore, research may lead us to classify the different causes of the same due to cell outages, sleeping cell or the non-usage of the mobile users appropriately. Also, SMS predictive model can be integrated with running time without hampering customer's privacy.

ACKNOWLEDGMENT

Authors acknowledge the support from REVA University for the facilities provided to carry out the research.

REFERENCES

[1] Larry E. Daniel, Lars E. Daniel, "Cellular system evidence and call detail records", ScienceDirect, pp. 225-237, year: 2012.

[2] Larry E. Daniel, Lars E. Daniel, "Discovery of call detail records", ScienceDirect, pp. 163-165, year: 2012.

[3] Mandy Chessell, Dan Wolfson, Tim Vincent, "Architecting to Deliver Value from a Big Data and Hybrid Cloud Architecture", ScienceDirect, pp. 33-48, year: 2017.

[4] Chenhan Xu, Kun Wang, Yanfei Sun, Song Guo, Albert Y. Zomaya, "Redundancy Avoidance for Big Data in Data Centers: A Conventional Neural Network Approach", IEEE paper, vol. 7, Issue 1, pp. 104-114, January-March 2020.

[5] Nirmal Ghotekar, "Analysis and Data Mining of Call Detail Records using Big Data Technology", IJARCCE, vol. 5, Issue 12, pp. 280-283, December 2016.

[6] Kashif Sultan, Hazrat Ali, Zhongshan Zhang, "Call Detail Records Driven Anomaly Detection and Traffic Prediction in Mobile Cellular Networks", IEEE paper, vol. 6, pp. 41728-41737, year: 2018.

[7] Muna Al-Saadi, Bogdan V Ghita, Stavros Shiaeles, Panagiotis Sarigiannidis, "A novel approach for performance-based clustering and management of network traffic flows", IEEE conference, pp. 2025-2030, year: 2019.

[8] Osvaldo Simeone, "A Very Brief Introduction to Machine Learning with Applications to Communication Systems", IEEE paper, vol. 4, Issue 4, pp. 648-664, December 2018.

[9] Md Salik Parwez, Danda B. Rawat and Moses Garuba, "Big data analytics for user activity analysis and user anomaly detection in mobile wireless network", IEEE paper, vol. 13, pp. 2058 - 2065, year: 2017.

[10] Bilal Hussain, Qinghe Du, Pinyi Ren, "Semi-Supervised Learning Based Big Data-Driven Anomaly Detection in Mobile Wireless Networks", IEEE paper, vol. 15, Issue 4, pp. 41- 57, year: 2018

[11] Ramin Sharif, Mahdiyar Molahasani Majdabadi, Vahid Tabataba Vakili "Mobile User-Activity Prediction Utilizing LSTM Recurrent Neural Network", IEEE Pacific Rim conference on communications, computers and Signal processing, year: 2019.

[12] Kashif Sultan, Hazrat Ali, Zhongshan Zhang, "Big Data Perspective and Challenges in Next Generation Networks", MDPI, year: 2018

[13] Pekka Kumpulainen, Kimmo Hätönen, Pekko Vehviläinen, "Automatic Discretization in Preprocessing For Data Analysis In Mobile Network", Research gate, pp. 813-816, year: 2003.

[14] Anish Nediyanchath, Chirag Singh, Harman Jit Singh, Himanshu Mangla, Karan Mangla, Manoj K. Sakhala, Saravanan Balasubramanian, Seema Pareek, Shwetha, "Anomaly Detection in Mobile Networks", IEEE conference, year: 2020.

[15] Zhang, C, Song, D, Chen, Y, Feng, X, Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H. and Chawla, N.V., "A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data", AAAI Conference on Artificial Intelligence, vol. 33, pp. 1409-1416, year: 2019.

[16] Kashif Sultan, Hazrat Ali, Adeel Ahmad, Zhongshan Zhang, "Call Details Record Analysis: A Spatiotemporal Exploration toward Mobile Traffic Classification and Optimization", MDPI, information, vol. 10, Issue 6, pp. 1-17, year: 2019.

[17] Yaohua Sun, Mugen Peng, Yangcheng Zhou, Yuzhe Huang, Shiwen Mao, "Application of Machine Learning in Wireless Networks: Key Techniques and Open Issues", IEEE paper, vol. 21, Issue 4, pp. 3072-3108, year: 2019.

[18] Serkan Ballı, Onur Karasoy, "Development of content-based SMS classification application by using Word2Vec-based feature extraction", IEEE paper, vol. 13, Issue: 4, publisher: IET, pp. 295 - 304, year: 2019.

[19] Dea Delvia Arifin, Shaufiah, Moch. Arif Bijaksana, "Enhancing spam detection on mobile phone Short Message Service (SMS) performance using FP-growth and Naive Bayes Classifier", IEEE paper, Date of Conference: 13-15 Sept. 2016, Date Added to IEEE Xplore: 16 January 2017, pp. 80-84.

[20] Jialin Ma, Yongjun Zhang, Jinling Liu, Kun Yu, XuAn Wang, "Intelligent SMS Spam Filtering Using Topic Model", IEEE paper, Date of Conference: 7-9 Sept. 2016, Date Added to IEEE Xplore: 27 October 2016, pp. 380-383.

[21] Nan Jiang, Yu Jin, Ann Skudlark, Zhi-Li Zhang, "Understanding SMS Spam in a Large Cellular Network: Characteristics, Strategies and Defenses", Springer paper, RAID 2013, vol. 8145, pp. 328-347.

[22] Liu Kesheng, Ni Yikun, Li Zihan, Duan Bin, "Data Mining and Feature Analysis of College Students' Campus Network Behavior", IEEE International conference on big data analytics, pp. 231-237, year: 2020.

[23] Siyang Qin, Youchen Zuo, Yaguan Wang, Xuan Sun Honghui Dong, “Travel Trajectories Analysis Based on Call Detail Record Data”, IEEE Chineses Control And decision Conference, pp. 7051-7056, year: 2017.

[24] Pranjal Pandey, online article: “Data pre-processing concepts”, available at https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c8, year: 2019.

[25] Mohammad Sultan Mahmud, Joshua Zhexue Huang, Salman Salloum, “A Survey of Data Partitioning and Sampling Methods to Support Big Data Analysis”, IEEE paper, vol. 3, Issue 2, pp. 85- 101, year: June 2020.

[26] Salman Salloum, Joshua Zhexue Huang, and Yulin He, “Random Sample Partition: A Distributed Data Model for Big Data Analysis”, IEEE paper, vol. 15, Issue 11, pp. 5846-5854, year: Nov 2019.

[27] Mingda Li, Hongzhi Wang, and Jianzhong Li, “Mining Conditional Functional Dependency Rules on Big Data”, IEEE paper, vol. 3, Issue 1, pp. 68-84, year: March 2020.

[28] Abdul Alim, Diwakar Shukla, “A Parameter Estimation Model of Big Data Setup Based on Sampling Technique”, IEEE International conference on data, engineering and applications, year: 2020.

[29] V R Saraswathy, M Prabhu Ram, A Vennila, S G Dravid, “Reduction of Features in the Data Set”, IEEE International Conference on Intelligent Computing and Communication for Smart World (I2C2SW), pp. 175-177, year: 2018.

[30] Duong Hoang, Brian Summa, Harsh Bhatia, Peter Lindstrom Pavol Klacansky, Will Usher, Peer-Timo Bremer, Valerio Pascucci, “Efficient and Flexible Hierarchical Data Layouts for a Unified Encoding of Scalar Field Precision and Resolution”, IEEE paper, vol. 27, Issue 2, pp. 603-613, year: February 2021.

[31] Kristina P. Sinaga, Miin-Shen Yang, “Unsupervised K-Means Clustering Algorithm”, IEEE paper, vol. 8, pp. 80716-8-727, year: 2020.

[32] Kristina P. Sinaga, Ishtiaq Hussain, Miin-Shen Yang, “Entropy K-Means Clustering with Feature Reduction Under Unknown Number of Clusters”, IEEE paper, vol. 9, pp. 67736-67751, year: 2021.

[33] Fan Liu, Yong Deng, “Determine the Number of Unknown Targets in Open World Based on Elbow Method”, IEEE paper, vol. 29, Issue 5, pp. 986-995, year: 2021.

[34] Yongxuan Lai, Songyao He, Zhijie Lin, Fan Yang, Qifeng Zhou, Xiaofang Zhou, “An Adaptive Robust Semi-Supervised Clustering Framework Using Weighted Consensus of Random k-Means Ensemble”, IEEE paper, vol. 33, Issue 5, pp. 1877-1890, year: 2021.

[35] Zhiyong Bu1, Bin Zhou, Pengyu Cheng, Kecheng Zhang, Zhen-Hua Ling, “Encrypted Network Traffic Classification Using Deep and Parallel Network-in-Network Models”, IEEE paper, vol. 8, pp. 132950-132959, year: 2020.

[36] A. Dainotti, A. Pescape, and K. Claffy, ‘‘Issues and future directions in traffic classification’’, IEEE paper, vol. 26, Issue 1, pp. 35–40, year: 2012.

[37] A. W. Moore, K. Papagiannaki, ‘‘Toward the accurate identification of network applications’’, Proc. Int. Workshop Passive Act. Netw. Meas. Cham, Switzerland: Springer, pp. 41–54, year: 2005.

[38] M. Finsterbusch, C. Richter, E. Rocha, J.-A. Muller, and K. Hanssgen, ‘‘A survey of payload-based traffic classification approaches’’, IEEE paper, vol. 16, Issue 2, pp. 1135–1156, year: 2014.

[39] Li Wenchao, Zhou Yong, Xia Shixiong, “A Novel Clustering Algorithm Based on Hierarchical and K-means Clustering”, IEEE Chinese Control conference, pp. 605-609, year: 2007.