Big Data Analytics for User-Activity Analysis and User-Anomaly Detection in Mobile Wireless Network

Md Salik Parwez, *Student Member, IEEE*, Danda B. Rawat, *Senior Member, IEEE*, and Moses Garuba, *Member, IEEE*

***Abstract*—The next generation wireless networks are expected to operate in fully automated fashion to meet the burgeoning capacity demand and to serve users with superior quality of experience. Mobile wireless networks can leverage spatio-temporal information about user and network condition to embed the system with end-to-end visibility and intelligence. Big data analytics has emerged as a promising approach to unearth meaningful insights and to build artificially intelligent models with assistance of machine learning tools. Utilizing aforementioned tools and techniques, this paper contributes in two ways. First, we utilize mobile network data (Big Data)—call detail record—to analyze anomalous behavior of mobile wireless network. For anomaly detection purposes, we use unsupervised clustering techniques namely k-means clustering and hierarchical clustering. We compare the detected anomalies with ground truth information to verify their correctness. From the comparative analysis, we observe that when the network experiences abruptly high (unusual) traffic demand at any location and time, it identifies that as anomaly. This helps in identifying regions of interest in the network for special action such as resource allocation, fault avoidance solution, etc. Second, we train a neural-network-based prediction model with anomalous and anomaly-free data to highlight the effect of anomalies in data while training/building intelligent models. In this phase, we transform our anomalous data to anomaly-free and we observe that the error in prediction, while training the model with anomaly-free data has largely decreased as compared to the case when the model was trained with anomalous data.**

***Index Terms*—5G, anomaly detection, call detail record (CDR), machine learning, network analytics, network behavior analysis, next generation wireless networks, wireless cellular network.**

Manuscript received October 31, 2016; revised December 16, 2016; accepted December 25, 2016. Date of publication January 9, 2017; date of current version August 1, 2017. This work was supported in part by the U.S. National Science Foundation under Grants CNS-1650831 and CNS- 1658972. Paper no.-TII-16-1268. *(Corresponding author: D. B. Rawat.)*

The authors are with the Department of Electrical Engineering and Computer Science, Howard University, Washington, DC 20059 USA (e-mail: mdsalik.parwez@bison.howard.edu; db.rawat@ieee.org; mgaruba@howard.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2017.2650206

I. Introduction

MASSIVE amount of data and information are produced by and about people, things, and their interactions. In wireless communication industries, the major drivers of Big Data are the increasing number of smart devices, machine-to-machine communications, and penetration of social media. With communication network evolution towards 5G, a multitude of technologies like base station densification and massive multiple input multiple output (MIMO) are expected to elevate the size of data exponentially. The data are generated at very large scale (volume) with expected size of 24.3 Exabyte per month [1], with fast input/output to/from the network (velocity) and from various sources within and outside the network (variety) and the quality, and trust of the data available at an incomparable level of volume, velocity, and variety (veracity). These unconventional 4V features (volume, velocity, variety, veracity) of current data generation give birth to *Big Data* and thus its management and analysis require schemes for *Big Data analytics* [2].

Big Data analytics is an umbrella term, that incorporates methods and technologies, hardware and software for collecting, and managing and analyzing large scale structured and unstructured data in real-time. Big Data analytics works on entire data as opposed to only sample data in conventional data analytics schemes. In the case of small data, analysis were performed by randomly selecting samples (partial data) that were considered as representative of the whole data. Due to analysis of only partial data, the information extracted are inaccurate and incomplete and thus the decisions made are suboptimal and the performance achieved are poor and suboptimal. Especially in the case of real network analysis and troubleshooting, precise and quick information are desired for providing exact solution, which can only be possible if whole/Big Data is analyzed. For current and the envisioned 5G mobile networks, Big Data offers a number of solutions in a variety of ways; some of them are outlined below [2].

1) Big Data analytics offer end-to-end visibility of the wireless network.
2) Big Data analytics enables self-coordination among network functions and entities.
3) Big Data analytics enables assessment of long-term dynamics of the network.

1551-3203 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: University of North Texas. Downloaded on October 17,2024 at 20:18:40 UTC from IEEE Xplore. Restrictions apply.

4) Big Data analytics builds faster and proactive network.
5) Big Data analytics enables smart and proactive caching in wireless network.
6) Big Data analytics enables energy efficient network operation.
7) Big Data analytics would enable unified performance evaluation.

In mobile wireless network, there are a number of network measurements and parameters which are continuously exchanged among, reported, and gathered at/from the user end (UE), and nodes in the radio access network, and core network (CN) of the long term evolution-advanced (LTE-A) network. Example includes call detail record (CDR), reference signal received power, radio link failure report, location information, UE movement behavior, etc. Additionally, with the emergence of 5G, there will be huge increase in the number of devices and nodes in the network; the dynamicity and heterogeneity of user and network environment, etc., causing exponential increase in the number and types of data. By analyzing the network measurements and information, network performance can be improved in three major ways. First, it enables effective control and optimization of network. Second, it facilitates service providers to optimize and enhance customer's experience by gauging all the relevant historical data. Last but not the least, *network analytics*-enabled insights can facilitate efficient network planning and deployment. Thus, network performance can be continuously monitored, instantly optimized, and proactively protected from possible faults, enabling an intelligent and self-organized network (SON).

In this paper, we exploit CDR information collected from CN of a real mobile cellular network. The spatio-temporal information contained within the CDR helps us to analyze user specific activity in a certain region at a particular time and date. By anomaly in this paper, we mean abnormal or unusual behavior or activity pattern of user and thus an accordingly effect on the network. Anomaly in network performance can be noticed due to multiple reasons, such as sleeping cell, hardware failures, surge in traffic, etc. Successful anomaly detection and its proper treatment results in multiple benefits, e.g., at a place like stadium, when the provided bandwidth resource does not suffice the highly-dense user demands, then it will cause choke in bandwidth, and it will appear as anomaly in the network. On one hand, such behaviors are considered normal in telecom and are not usually counted amongst anomalies. On the other hand, categorizing them into anomalies would help the network operators not only in detecting them but also in determining the region of interests (ROIs), for which proper network resources can be proactively allocated for enhanced quality of experience.

Our contributions toward anomaly detection in this paper are as follows.

1) We utilize machine-learning tools namely K-means clustering and Hierarchical clustering to detect anomalies.
2) We verify and compare the anomalies with those found through ground truth observation to determine their accuracy.
3) After anomaly detection and verification, the performance of anomaly-free data is evaluated by passing it through a neural network-based prediction model that forecasts future traffic pattern. We observe that the mean square error (MSE) between the training, validation, and test data is largely reduced as compared to that in the case when the model was trained with anomalous data.

Although the above machine learning tools have widely been used in literature, we utilize them here with the purpose of comparing the anomaly detection performances among themselves before comparing with anomalies found through ground truth data. This paper demonstrates a key use case and underlying processes of analyzing user and network behavior utilizing data collected from mobile network. Additionally, it discusses and presents forecasting network activity with anomaly-free data, which is an important step behind making the network intelligent and accurate. The network analytics presented is one of the viable approaches for enabling smart societies where almost everything will be connected and each task will be performed intelligently, such as health services, diet instruction, etc. Based on similar concept, an idea of ambient assisted living has been presented in [3] to understand needs of elderly people, by analyzing data collected from various types of nodes in the surrounding network.

The rest of the paper is organized as follows. Section II describes the relevant work in the literature. The dataset and the preprocessing use has been explained in Section III. In Section IV, we discuss the machine learning algorithms utilized for detecting anomalies. Section V investigates the anomalies found through the learning algorithms. The prediction model and its performance comparison in case of data with anomaly and without anomaly have been discussed in Section VI. Section VII concludes the paper.

II. Relevant Work

In the literature, anomaly detection has been performed through various supervised, semisupervised, and unsupervised learning methods. Naboulsi *et al.* [4] proposed a framework that categorizes large size CDR into different call profile and accordingly classify network usages. As a by-product, the proposed framework identifies unexpected traffic that is anomalous as compared to the normal. They consider large-scale dataset of voice calls, however, our dataset is comprised of voice calls as well as text messages. Article [5] categorized anomaly detection techniques for Big Data based on nearest neighbors, clustering, and statistical approaches. Their finding suggests that for real datasets, clustering-based anomaly detection technique performs the best and second best performance attained by the nearest neighbor-based technique k-NN. In [6]–[8], k-means clustering approach has been performed on CDR for different purposes, such as identifying industrial parks and office areas, commercial areas, nightlife areas, leisure areas, and residential areas. It was validated that there was a strong correspondence between land uses and the infrastructures included in the geographical representation of each cluster. Using the k-means clustering, the authors were able to detect volume anomalies in real network traffic, achieving satisfactory results. K-means clustering has been applied in [9] and [10] for anomaly detection

Authorized licensed use limited to: University of North Texas. Downloaded on October 17,2024 at 20:18:40 UTC from IEEE Xplore. Restrictions apply.

in traffic data. The data contains unlabeled flow records, which k-means algorithm separates into clusters of normal and anomalous traffic. Cici *et al.* [11] utilize agglomerative hierarchical clustering to perform segmentation and anomaly/outlier detection from a single dendrogram, utilizing Pearson correlation as distance function. Rule-based approach has been utilized in [12] have been for detecting anomaly CDR information generated from wireless network. However, study in [12] overlooks the effect of nontravelling users in the analysis, which is a common case in business areas. On the contrary in this paper, we first categorize the call activity information into different clusters. This in turn helps to identify the unnatural behavior in the activity on a particular date and time, which further can be traced back for root cause analysis. Few similar works along with security concern are also presented in [13]–[16].

Run-time data analysis and solution are the other perspective of Big Data analytics approach. To cope with the rapidly changing network and user behavior, networks needs to perform not only adaptive but also agile, and to achieve that, networks are required to analyze the real data and provide solution in real time and in time-efficient manner. Toward this, some interesting examples can be found in [17]–[19]. This becomes more important when the user is mobile such as in vehicular communication as discussed by these papers [17]–[19]. A similar example on data analytics scheme for water sustainability has also been discussed in [20].

Motivated from above works, we chose clustering-based approaches to detect anomalies. However, our work distinguishes from all above works with the fact that they are limited to detecting anomalies, whereas we move a step further to extract the meaning and cause of those anomalies so that appropriate action can be taken, such as proactive resource allocation or cell outage avoidances, as discussed earlier in Section I. Additionally, we evaluate the effect of anomalous data by using it for prediction over a neural-network-based prediction model and we observe that the prediction error decreases by significant amount after making the data anomaly-free. Thus the ultimate objective is circled around detection of anomalies and to see its effect on normal operation of a mobile wireless network. Using above concept, when the network observes any abnormal behavior in its operation, it identifies them as anomalies. Moreover, these kinds of approaches are crucial for efficient 5G network applications where accuracy and precision are extremely important, such as health care, defense, etc.

III. System Model and Dataset Preprocessing

The system model basically consists of LTE-A cellular network. The overall LTE-A network architecture with network elements and the standardized interfaces is shown in Fig. 1. At a high level, the network has three layers; the user end (UE), the access network (E-UTRAN) and the CN. The UE consists of only the user mobile devices; the access network is also made up of essentially just one node, the evolved-NodeB (eNodeB) to which UE connects. The CN on the other hand, consists of many logical nodes such as mobility management entity (MME), service gateway, packet data network gateway, etc. The CDR information is collected at CN layer of the network. We utilize this data to understand user and network activity and determine anomaly therein.

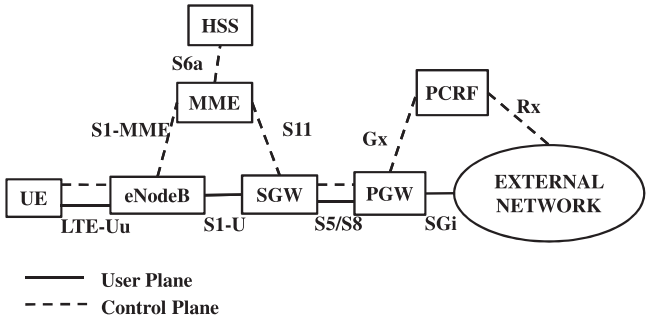

Fig. 1. LTE-advanced network architecture.

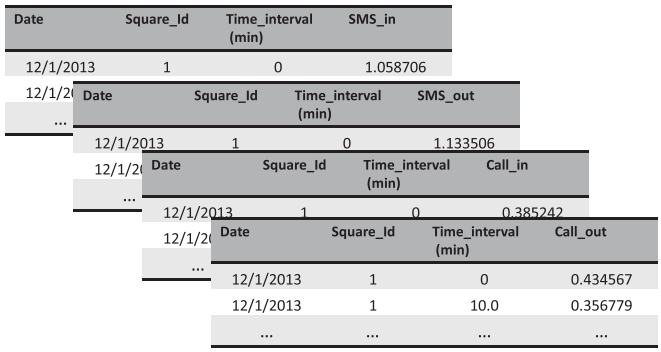

Fig. 2. Dataset before preprocessing.

A. Description of the Dataset

The CDR dataset used in this paper was obtained from real network Telecom Italia that had been made publicly available as part of Big Data Challenge 2014 competition [21]. This dataset provides information about the telecommunication activity over the city of Milan. The Milan region is divided into 100×100 square grids, which are named as Milan Grid. Each grid has side length of "0.235 km" and an area of 0.055 km^2. The dataset contained separate files for different activities such as inbound call, outbound call, inbound SMS, and outbound SMS from every grid with time interval of 10 min over a day for the duration of a week as shown in Fig. 2.

The CDR dataset contained the following fields before preprocessing.

1) Square ID: This indicates the identification number of the square grids.
2) Time Stamp: In the raw dataset, data were recorded over an interval of 10 min.
3) Inbound Call Activity: It indicates the duration of the inbound call at a particular grid within time stamp of 10 min.
4) Outbound Call Activity: It indicates the duration of the outbound call at a particular grid within time stamp of 10 min.

Authorized licensed use limited to: University of North Texas. Downloaded on October 17,2024 at 20:18:40 UTC from IEEE Xplore. Restrictions apply.

Date	Square_id	Time_interval (hours)	Activity
12/1/2013	1	0	3.012022
12/1/2013	1	1	1.652472
...	...	...	...

Fig. 3. Dataset after preprocessing.

5) Inbound SMS Activity: It indicates the duration of the inbound SMS at a particular grid within time stamp of 10 min.
6) Outbound SMS Activity: It indicates the duration of the outbound SMS at a particular grid within time stamp of 10 min.

B. Dataset Preprocessing

The dataset provided by the telecom service provider were in raw form that we preprocessed before analyzing them. The preprocessing phase consist of cleaning and filtering of data in order to avoid the generation of misleading and inappropriate rules or patterns as discussed in earlier section. The irregularities in the data can be in the form of noise, missing data fields in some of the data records, etc. We utilized Apache Pig tool on Hadoop-based data processing architecture. Since the sizes of the data were large, Hadoop served as ultimate choice for time efficient processing of large dataset. For details on Apache Pig and Hadoop architecture, please refer to detailed tutorial explained in [22].

The dataset were aggregated to one-hour intervals, which were for only 10 min before processing. Next, we combined the separate datasets corresponding to the grids, date, and time into one single dataset. We transformed the datasets in such a way that various information about a particular grid, such as multiple calls and SMS activities, on a particular date and time could be done in simple few steps. In other words, combining the dataset made the data processing and its investigation easy and smooth. Additionally, it decreased the required memory and processing resources of the data mining algorithms.

After uniting the separate dataset, the final dataset looks as shows in Fig. 3. In the final dataset, we summed up the inbound and outbound call and SMS activities at every grid and named them as activity at that grid. It is because; this shows the combined call and text activity by the user at every grid at every time step. Once we acquired aggregated activity of each grid, we use this to identify anomalies in the data. In the next section, we discuss about the machine learning algorithms that we utilized to detect the anomalies that we later remove to make the dataset anomaly-free. In the following section, we explain the machine learning algorithms that were applied to anomalies.

IV. Anomaly Detection Algorithms

Anomalies represent the behavior of the network, which is different than what is usually expected and observed. In the following, we discuss the four machine learning algorithms detailing and how do they detect anomalies.

A. K-Means Clustering

K-means is one of the simplest unsupervised clustering methods utilized to solve well-known clustering problems, especially for large datasets. It belongs to the category of partitioning methods in which a database containing n objects is partitioned into a set of k clusters. It assumes fixed number of cluster k known *a priori*. Determining the optimal value of k in itself differs from dataset to dataset and method to method. However, for dataset under consideration, the optimal number of cluster was determined using Elbow method [23]. After determining k, the main task involves finding partition of k clusters that optimizes the chosen partitioning criterion. Given the input to the algorithm is k, the k-means algorithm partitions a given dataset into k clusters so that the resulting intracluster homogeneity is high and intercluster homogeneity is low. Cluster similarity is measured in terms of mean value of the objects in a cluster, which is usually cluster's centroid or center of gravity.

The k-means algorithms clustering can be summarized in following few steps.

1) Choose randomly the initial value of k from the space represented by the objects being clustered. The values of k represent the initial values of the centroids of k clusters.
2) Compare the distance of each of the objects to each of the centroid, and assign the object to the cluster with closest centroid.
3) Recalculate the centroid of clusters by finding the mean value of the new cluster formed in step 2).
4) Repeat steps 2) and 3) until the centroids become stationary.

After clustering through k-means, the clusters containing the fewest number of objects are considered to be anomaly. There are many other ways to detect an anomaly after grouping them into k clusters. Since the activity of the user at a grid is always changing based on events and occasion, date, and time, anomaly will be detected only when there is unexpectedly high rise in their activity and such activities are observed to be fewer and will be grouped into a different cluster by k-means. Thus the cluster containing only few objects, we identify it to be anomalous.

B. Hierarchical Clustering

The Hierarchical clustering is also a partition-based clustering method as k-means. It makes a hierarchy of clusters for which it takes either bottom-up (agglomerative) approach or top-down (divisive) approach. The agglomerative approach treats each object as a singleton clusters and then successively merges pairs of clusters until all clusters are merged into a single cluster that contains all objects. On the other hand, the divisive clustering considers the whole objects a single cluster and then splits them recursively until all individual objects are a single luster. The agglomerative is more frequently used than the divisive approach.

Authorized licensed use limited to: University of North Texas. Downloaded on October 17,2024 at 20:18:40 UTC from IEEE Xplore. Restrictions apply.

The Hierarchical clustering approach can be described in following few steps.

1) We start by considering each object into its own cluster. Therefore, for n number of objects, there will be n number of clusters.
2) The distance between those individual clusters are calculated and the closest (most similar) pair of clusters are merged into a single cluster. So now we have one less cluster.
3) Then we compute the distances (similarities) between the new made clusters and each of the old clusters.

We repeat steps 2) and 3) until all items are clustered into a single cluster. However, step (3) is performed in different ways that are called *single-link*, *complete-link,* and *average-link* type of hierarchical clustering. We utilized average-link clustering, which consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster. For more on single-link, complete-link, and average-link clustering, please refer to [23]–[26].

C. Complexity Analyses of K-means and Hierarchical Clustering

As we know, the time complexity of k-means is linear in the number of data objects, i.e., $O(kn)$, where k is number of cluster, n is number of data objects. However, the time complexity of the hierarchical clustering algorithms is quadratic, i.e., $O(n^2\log n)$. Therefore, for the same amount of data, hierarchical clustering will take quadratic amount of time.

The space complexity of k-means algorithm is $O(k + n)$, where again k is number of cluster and n is number of data objects. The space complexity of hierarchical clustering is $O(n^2)$, which is quite higher than k-means clustering. Thus, space complexity of hierarchical clustering is a limiting factor for large dataset processing.

From above discussion, we conclude that k-means has lesser time as well as space complexity than hierarchical clustering, especially for large datasets. However, k-means starts with a random choice of cluster centers, therefore it may yield different clustering results on different runs of the algorithm. Thus, the results may not be repeatable and lack consistency. Hierarchical clustering on the other hand definitely results in same clusters after several repetitions, thus maintains consistency. Therefore, there exists tradeoff in their performance. In other words, for lesser complexity, k-means clustering is better, whereas for better performance, hierarchical clustering is better. In our experiments, both performances were almost same with higher than 90% accuracy.

V. Investigation and Verification of Anomalies

In this section, we present the anomalies detected through machine learning algorithms explained above. We first investigate the data by plotting in its original form to check if there are anomalies occurring at any grid on any date and time. After this visual check, we find anomalies through the algorithms which we verify by matching the anomalies detected in a grid for a given date and time.

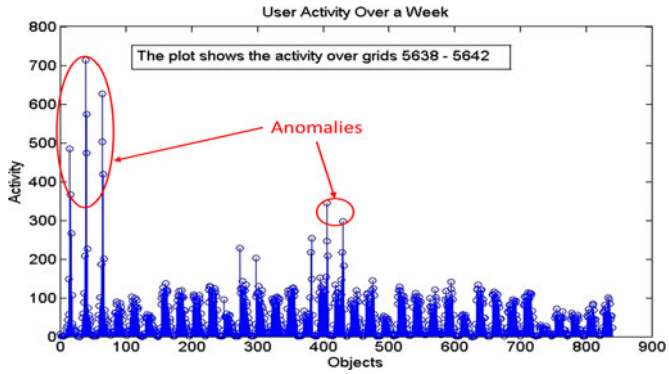

Fig. 4. Activity plot and anomaly investigation.

TABLE I
Anomalous Activity and the Corresponding Grids

Date	Square_ID	Time	Activity
12/1/2013	5638	14:00	458.258
12/2/2013	5638	15:00	365.670
12/3/2013	5639	14:00	713.012
12/3/2013	5639	15:00	573.832
12/3/2013	5640	14:00	627.117
12/3/2013	5640	15:00	501.914
12/3/2013	5640	16:00	419.929
12/4/2013	5639	21:00	344.277
12/4/2013	5640	21:00	183.173

A. Investigating the Anomalies Through Ground Truth Data

Fig. 4 shows the activity of the user at the grids 5638–5642 collected over a week. Looking at the user activity trend over a week, it can be noticed that the activity level are quite higher in the beginning and slightly higher toward the middle of the plot. Since those activity level exhibit abnormal behavior of the network, they are counted to be anomalies. While investigating those anomalies corresponding to their grids, it is found that it was Sunday and the anomalies occurred between 2 PM and 4 PM, as shown in Table I. The grids at which the anomalies were observed were nearby a stadium in which at 3 PM, a huge soccer match was held. This is not surprising that when there is such a big event, there is presence of thousands of people, each of them texting, calling, and sharing multimedia contents. Thus it leads to surge in traffic flow showing an abnormal behavior in the network. Anomaly detection in such a case helps in identifying ROI which can be stadiums, hospitals, universities, exhibition centers, etc., for which network can be planned and resource can be allocated in advance for uninterrupted provision of the service. The other anomalies in the small circle have been observed when it was 9 PM at few of the grids, which is relatively higher and that was due to busy hour. As compared to the anomalies in the first case, this may or may not be considered anomaly depending upon types of data and the expected behavior.

Authorized licensed use limited to: University of North Texas. Downloaded on October 17,2024 at 20:18:40 UTC from IEEE Xplore. Restrictions apply.

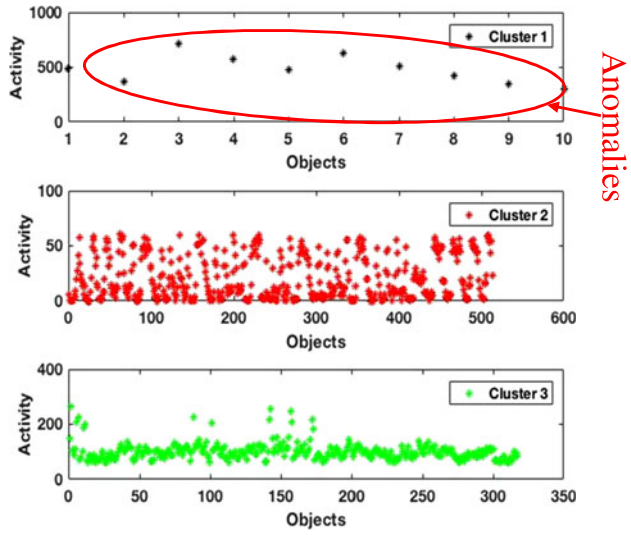

Fig. 5. Anomaly detection through k-means clustering.

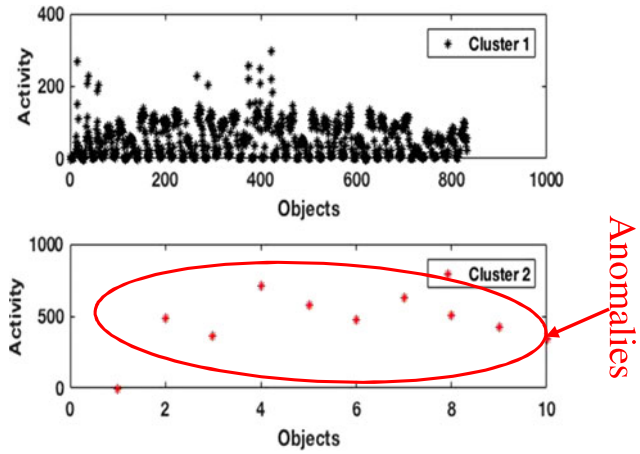

Fig. 6. Anomaly detection through Hierarchical clustering.

B. Anomalies Detected Through Machine Learning Algorithms

The anomalies found through k-means and hierarchical clustering methods are shown in Figs. 5 and 6, respectively. We considered that the anomalous activity objects are unique than the objects exhibiting normal activity, thus will be grouped into separate cluster. Since normal operation exhibits common/normal characteristics, they will be grouped into a cluster against the abnormal activities that are mostly few in number lie in a separate cluster. Thus the clusters with fewest members are considered to be containing anomalous activities. As shown in Fig. 5, there are three clusters, each containing activities with most of the objects lying within an activity range, e.g., objects of activity level 50–250 are grouped into cluster 1, activity level 1–50 are grouped into cluster 2, and the rests in cluster 3. From the activity level, we observe that most of the user's activities are represented by clusters 1 and 2, whereas cluster 3 contains less number of users with higher activity level. This is against the trend shown by (most) users in clusters 1 and 2. Intuitively, this indicates an abnormal trend and thus, they have been categorized into anomalies. Similarly, cluster 2 found through hierarchical clustering has fewest members as shown in Fig. 6. When the objects in those clusters were verified against the anomalies found through ground truth data, their activity level matched with those of the anomalies.

In case of user-centric, proactive, and automated 5G networks operation, user and network behavior analysis as well as its exact prediction are the backbone of efficient network operation. To enable that, intelligent prediction models are being developed that heavily depends upon precise information and correct data—also known as veracity. Anomalous data can pose minor to major threat on wireless network operation. For example, there is an intelligent model developed to forecast/understand user's future demand in a network. If the model has not been trained with clean and anomaly-free data, it will not forecast the future demand correctly, but this pose a minor threat as the user will still be served with at least average quality of service. Thus, anomalous data might result in minor threat. However, an intelligent model developed to predict/detect fault occurrence (or a sleeping cell [25]) in the network does not perform correctly, then there may occur network outage affecting large number of customers, eventually resulting in churn. In the next, we observe the effect of anomalous and anomaly-free data by passing it through a neural-network-based prediction model and observing the error difference.

VI. Forecasting Activity Levels

After detecting and locating the anomalies (anomalous activities), we replace those anomalous activities by average activities of all the users to make the data anomaly-free. We do this for the sake of performance evaluation. Now we have earlier data (which is anomalous) and the latest data, the data we just made anomaly-free.

Using these data, we train a neural-network-based forecasting model to observe the difference in error. We passed activity data of users corresponding to six days at one of the square grid. In the case of this grid, we already have information about the anomalies and its whereabouts through ground truth observation. Thus we understand that utilizing data corresponding to this grid will help in assessing the performance. We observed the effect on MSE caused in prediction of the model after passing anomalous and anomaly-free data. Figs. 7 and 8 presents the MSE observed in the prediction model using data with and without anomalies, respectively. It is observed that, the MSE for training data, validation data as well as test data are significantly higher when the prediction model acted upon the data with anomalies. However, when the same model acted upon anomaly-free data, the overall MSE significantly decreased. Moreover, there are many differences observed in MSE among the training, validation, and test data when the prediction model was passed anomalous data. Additionally, the error difference among the training, validation, and test data after removing the anomalies decreased.

Further, the impact of the MSE on the performance of the model can be severe depending upon the target of the model.

Authorized licensed use limited to: University of North Texas. Downloaded on October 17,2024 at 20:18:40 UTC from IEEE Xplore. Restrictions apply.

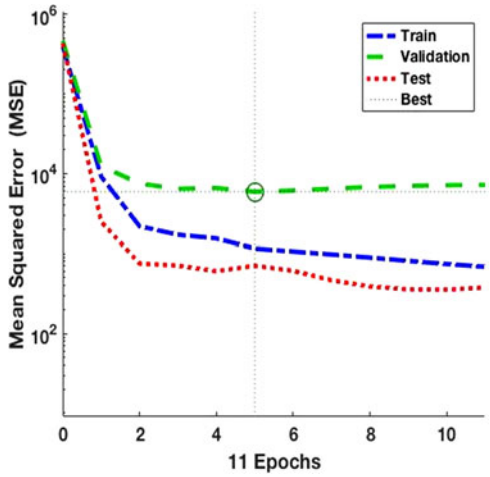

Fig. 7. MSE of the prediction using training data with anomaly.

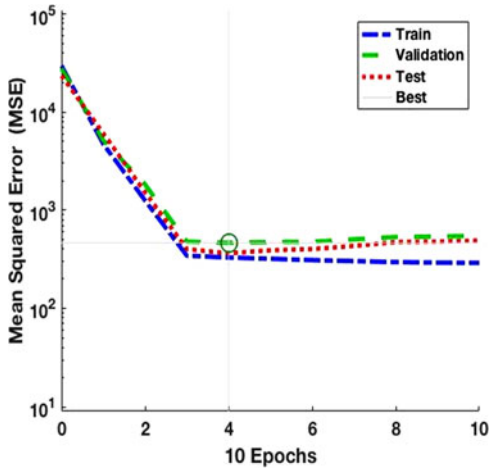

Fig. 8. MSE in the prediction using training data without anomaly.

For example, consider there is a model designed to detect a sleeping cell in the network. If the model is unable to detect the sleeping cell because of anomalous data it was trained with, then the network service will be heavily affected eventually leading to churn. In the case of 5G networks and beyond, there will be thousands of nodes, many of them co-operating with each other. In such a case, along with intelligent design of system, accuracy, and speed of the data and its processing will play vital role. Thus ensuring the data anomaly-free is extremely important

VII. Conclusion and Future Work

In this paper, we presented anomaly detection in mobile network Big Data (*Call Detail Record*: CDR) using machine-learning (clustering) technique. We analyzed user activities at different time and location from the spatio-temporal information contained within CDR using k-means and hierarchical clustering techniques. The user activities that were unusually high caused unusual traffic demand and thus were categorized into anomalies. Further, after verifying these anomalies with ground truth information, we found that the location where there was an unusual high traffic, was a stadium and at the time when the network experienced high demand, there was a soccer match going on and thus the traffic experienced was higher than on usual days. Thus with the help of anomaly detection, ROI can be identified, for which proper action (e.g., proper resource allocation) can be taken in advance to meet the requirements. We also discussed the effect of anomalous and anomaly-free data by experimenting on a prediction model. We found that training the model with anomaly-free data resulted in less MSE than with anomalous data.

This paper can be extended to understand the dynamics of user in envisioned smart cities. Big Data analytics approach can be utilized to understand users' contextual information, such as mobility pattern, traffic pattern, their choice of contents, their social network, ties, etc. By extracting these insightful information, smart and intelligent resource allocation algorithms can be developed for efficient resource utilization. Additionally, data analytics can also be performed on data collected from energy meters in every home to understand and appropriately determine energy utilization pattern in smart cities.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments on this paper. The authors would also like to thank Dr. A. Imran for a preliminary discussion on the topic presented in this paper. Any opinion, finding, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation.

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update 2015-2020," White Paper, 2016. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.pdf

[2] N. Baldo, L. Giupponi, and J. Mangues-Bafalluy, "Big data empowered self organized networks," in *Proc. 20th Eur. Wireless Conf.*, 2014, pp. 1–8.

[3] J. Lloret, A. Canovas, S. Sendra, and L. Parra, "A smart communication architecture for ambient assisted living," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 26–33, Jan. 2015.

[4] D. Naboulsi, R. Stanica, and M. Fiore, "Classifying call profiles in large-scale mobile traffic datasets," in *Proc. IEEE Conf. Comput. Commun.*, 2014, pp. 1806–1814.

[5] M. Ahmed, A. Anwar, A. N. Mahmood, Z. Shah, and M. J. Maher, "An investigation of performance analysis of anomaly detection techniques for big data in SCADA systems," *EAI Endorsed Trans. Ind. Netw. Intell. Syst.*, vol. 15, no. 3, pp. 1–16, 2015

[6] V. Soto and E. Frías-Martínez, "Automated land use identification using cell-phone records," in *Proc. 3rd ACM Int. Workshop MobiArch*. 2011, pp. 17–22.

[7] M. Amer, "Comparison of unsupervised anomaly detection techniques," Bachelor Thesis, 2011. [Online]. Available: http://www.madm.eu/_media/theses/thesis-amer.pdf

[8] A. Zoha, A. Saeed, A. Imran, M. A. Imran, and A. Abu-Dayya, "A SON solution for sleeping cell detection using low-dimensional embedding of MDT measurements," in *Proc. IEEE 25th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun.*, 2014, pp. 1626–1630.

[9] G. Münz, S. Li, and G. Carle, "Traffic anomaly detection using k-means clustering," in *Proc. GI/ITG Workshop MMBnet*, 2007, pp. 1–8.

[10] M. F. Lima, B. B. Zarpelao, L. D. H. Sampaio, J. J. P. C. Rodrigues, T. Abrao, and M. L. Proença Jr., "Anomaly detection using baseline and K-means clustering," in *Proc. Int. Conf. Softw. Telecommun. Comput. Netw.*, 2010, pp. 305–309.

Authorized licensed use limited to: University of North Texas. Downloaded on October 17,2024 at 20:18:40 UTC from IEEE Xplore. Restrictions apply.

[11] B. Cici, M. Gjoka, A. Markopoulou, and C. T. Butts, "On the decomposition of cell phone activity patterns and their connection with urban ecology," in *Proc 16th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2015, pp. 317–326.
[12] I. A. Karatepe and E. Zeydan, "Anomaly detection in cellular network data using big data analytics," in *Proc. 20th Eur. Wirel. Conf.*, 2014, pp. 1–5.
[13] Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of things and big data analytics for smart and connected communities," *IEEE Access*, vol. 4, pp. 766–773, 2016. [Online]. Available: https://doi.org/10.1109/ACCESS.2016.2529723
[14] D. B. Rawat and S. R. Reddy, "Software defined networking architecture, security and energy efficiency: A survey," *IEEE Commun. Surveys Tuts.*, vol. PP, no. 99, pp. 1–22, 2016.
[15] R. K. Sharma and D. B. Rawat, "Advances on security threats and countermeasures for cognitive radio networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 1023–1043, Apr.–Jan., 2015.
[16] X. Xiong *et al.*, "Empirical analysis and modeling of the activity dilemmas in big social networks," *IEEE Access*, vol. PP, no. 99, pp. 1–9, 2016.
[17] C. Nicola *et al.*, "Distributed and adaptive resource management in cloud-assisted cognitive radio vehicular networks with hard reliability guarantees," *Veh. Commun.*, vol. 2, no. 1, pp. 1–12, 2015.
[18] M. Shojafar, N. Cordeschi, and E. Baccarelli, "Energy-efficient adaptive resource management for real-time vehicular cloud services," *IEEE Trans. Cloud Comput.*, vol. PP, no. 99, pp. 1–14, 2016.
[19] W. Li and H. Song, "ART: An attack-resistant trust management scheme for securing vehicular ad hoc networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 960–969, Apr. 2016.
[20] H. Song and M. Brandt-Pearce, "Range of influence and impact of physical impairments in long-haul DWDM systems," *J. Lightw. Technol.*, vol. 31, no. 6, pp. 846–854, Mar. 2013.
[21] (2014). [Online]. Available: https://dandelion.eu
[22] [Online]. Available: https://pig.apache.org
[23] M. Yan, "Methods of determining the number of clusters in a data set and a new clustering criterion," Ph.D. dissertation, Dept. Statist., Virginia Tech, Blacksburg, VA, USA, 2005.
[24] D. B. Rawat and C. Bajracharya, Vehicular cyber physical systems: Adaptive connectivity and security. New York, NY, USA: Springer, 2016.
[25] S. Chernov, M. Cochez, and T. Ristaniemi, "Anomaly detection algorithms for the sleeping cell detection in LTE networks," *2015 IEEE 81st Veh. Technol. Conf.*, May 2015, pp. 1–5.
[26] M. G. Pineda, S. Sendra, C. T. Ribalta, and J. Lloret, "Users macro and micro-mobility study using WLANs in a university campus," *Int. J. Adv. Int. Technol.*, vol. 4, no. 1, pp. 37–46, 2011.

Md Salik Parwez (S'13) received the B.S. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2008, the M.E. degree in information science from Nara Institute of Science and Technology, Nara Japan in 2014, and the M.S. degree in electrical and computer engineering from the University of Oklahoma, Norman, OK, USA, in 2016. He is currently working toward the Ph.D. degree in the Department of Electrical Engineering and Computer Science, Howard University, Washington, DC, USA.

His research interests include next generation wireless networking, 5G mobile networks, big data analytics, and cybersecurity.

Danda B. Rawat (M'10–SM'13) received the bachelor's degree in computer engineering in 2002, the master's degree in information and communication engineering from the Tribhuvan University, Kathmandu, Nepal, in 2005, and the PhD degree in electrical and computer engineering from Old Dominion University, Norfolk, VA, USA, in 2010.

He is currently an Associate Professor in the Department of Electrical Engineering and Computer Science, Howard University, Washington, DC, USA. Prior to Howard University, he was with the College of Engineering and Information Technology, Georgia Southern University (GSU), Statesboro, GA, USA, as a faculty member. He is the Founder and Director of the Cyber-security and Wireless Networking Innovations Research Laboratory, Howard University. His research interests include wireless communication networks, cyber security, cyber physical systems, Internet of Things, big data analytics, wireless virtualization, software-defined networks, smart grid systems, wireless sensor networks, and vehicular/wireless ad hoc networks.

Dr. Rawat received the NSF Faculty Early Career Development (CAREER) Award in 2016, and the Outstanding Research Faculty Award (Award for Excellence in Scholarly Activity) 2015, Allen E. Paulson College of Engineering and Technology, GSU, among others. He has been serving as an Editor/Guest Editor for more than 15 international journals. He serves as a technical program committee (TPC) Vice-Chair (Information Systems) for IEEE INFOCOM 2018, served as a Web-Chair for IEEE INFOCOM 2016/2017 and as a Student Travel Grant Co-chair of IEEE INFOCOM 2015, a Track Chair for Wireless Networking and Mobility of IEEE CCNC 2016, a Track Chair for Communications Network and Protocols of IEEE AINA 2015, and more. He served as a program chair, general chair, and session chair for numerous international conferences and workshops, and served as a TPC member for several international conferences including IEEE INFOCOM, IEEE GLOBECOM, IEEE CCNC, IEEE GreenCom, IEEE AINA, IEEE ICC, IEEE WCNC, and IEEE VTC conferences. He is a member of the ACM and ASEE. He served as a Vice Chair of the Executive Committee of the IEEE Savannah Section and Webmaster for the section from 2013 to 2017.

Moses Garuba (M'13) received the B.Sc. degree in computer science, the M.Sc. degree in information technology, both from the University of London, London, U.K. in 1992 and 1993, respectively, Master of Computer Science degree from Howard University, Washington, DC, USA, in 2000, and Ph.D. degree in computer science and infomation security from the University of London, in 2000, and the Master of Laws degree from the University of Strathclyde, Glasgow, United Kingdom, in 2015.

He is currently a Professor of computer science at Howard University. He has authored numerous research articles. His research interests include multilevel database security, secure transaction processing, quantum cryptography, integrity of digital evidence, electronic privacy law, and e-commerce.

Dr. Garuba received many awards from the National Science Foundation, Department of Defense and the Defense Advanced Research Projects Agency. He is an Editor of the *Journal of Information Technology Impact*, an Associate Editor and program committee member of the International Conference on Information Technology: Next Generations, a member of the IEEE-USA Working Group on Bioterrorism, and the Co-founder of the International Federation for Information Processing (IFIP) Working Group 9.6 IT Misuse and the Law.

Authorized licensed use limited to: University of North Texas. Downloaded on October 17,2024 at 20:18:40 UTC from IEEE Xplore. Restrictions apply.