

Original Research

Received 13 June 2026

Revised 25 July 2026

Accepted 08 Aug 2026

Natural Resources for Human
Health 2026, Vol. 6 (15s): 1252–1260

KEYWORDS:

Call Detail Record, Traffic clusters,
Within Cluster Sum of Squares,
Arm method

Mobile traffic clusters analysis using Within Cluster Sum of Squares optimization

Sandhya B S¹, Pavana H²

¹Dept. of Electronics and Communication Engineering, KNS Institute of Technology, Bengaluru, India

²Dept. of Artificial Intelligence and Data Science, K. S. School of Engineering and Management, Bengaluru, India

ABSTRACT: The increasing utilization of mobiles, encompassing 4G and 5G technologies, generates a substantial accumulation of data within the cellular network. Proper management of resources is vital to provide service without any interruption to customers with cost effectiveness. In this regard, one of the major challenges is the detection of the cellular traffic caused due to various mobile phone activities like in-out calls, in-out Short Message Service (SMS), internet activity etc. In the proposed work, mobile traffic clusters are detected using Within Cluster Sum of Squares optimization method. Firstly, unprocessed Call Detail Record (CDR) is to be converted into a structured format. Next, Within Cluster Sum of Squares (WCSS) optimization method identifies both high-volume and low-volume traffic clusters in the mobile network. Finally, the work is verified using arm method subsequent the formation of the clusters.

1. INTRODUCTION

The increased usage of mobile phones has created huge accumulation of data traffic in mobile network. The effective detection and analysis of the data traffic is required for proper management of resource. The transaction details of all the mobile activities are recorded as Call Detail Record (CDR) by the service providers. The events are in-out calls, SMS and internet transactions. Every activity's start and end lapses are mentioned in the CDR.

The key goal of mobile network is to provide uninterrupted phonic and internet activities with larger range and cost effectiveness. The objective is not about detection of any radio source. However, it is possible to find active radio sources using CDR [1]. To analyze the CDR precisely good amount of data from the network are required with geographical information [2]. The CDR records every mobile user's activity and analysis of the same helps to provide good service [3]. The challenging part for network providers is to move and store the huge data in the designated centers [4]. Prior to any analysis the unstructured data must be converted into machine understandable structured data [5]. This important step is called pre-processing which includes quality analysis of 4Vs. They are volume, velocity, variety and value [6].

Before any analysis is done, the unstructured data is converted to structured data. Detection of mobile traffic and fault in network, spam in SMS is crucial analysis to be conducted. It helps with improving the performance of the mobile network in terms of many aspects like reliability and capacity. The end results of analysis help in reducing delays and errors in the mobile network [7].

Machine learning tools would be a possible solution to analyze the big, structured CDR data with less human intervention. Before applying, it is necessary to widely understand machine learning tools and suitable approaches. As per the intended result, right tool can be selected. Hence, it is better to understand machine learning in detail and its suitability in the mobile network.

Machine learning approach can replace conventional methods by understanding the required features to develop an algorithm. The machine learning tool must be selected suitably for the intended data analysis. The different types of learning are supervised, unsupervised and reinforcement learning. If the input data points are mapped to output data points while training an algorithm, then it is known as supervised learning. Naïve Bayes, Random Forest, Logistic Regression and Support Vector Machine (SVM)

are few examples. If an algorithm is trained by mapping unknown data points of input to output, then it is unsupervised learning. Principal Component Analysis (PCA), Independent Component Analysis (ICA) and K-means clustering are familiar examples. Reinforcement learning is a compromised learning which gives an option to retrain the algorithm for better analysis.

In the view of understanding and analysing, mobile network traffic it is suitable to incorporate unsupervised learning integrated with the verification method. Within Cluster Sum of Squares (WCSS) is the metric that can be used by K-means clustering to detect traffic clusters in the mobile network. This method detects both low-volume and high-volume traffic clusters and can be verified with the graph of arm shape. The sections in the paper are arranged in following fashion. Related works in section 2, followed by formatting of the CDR in section 3. Identification of traffic clusters using WCSS optimal method is detailed in section 4. Results and conclusion are detailed in sections 5 and 6 respectively.

2. RELATED WORKS

The structured data obtained after pre-processing includes details of calls and SMS with digital record of time and details of each grid [9]. The removal of irrelevant features is very important in the structured data to avoid unnecessary overhead in CDR analysis and make machine understandable [10]. The structured data can be obtained using sandbox Horton functions. Not applicable fields are taken as zero, and the CDR activities are treated as one single activity with a record of 60 minutes. This increases efficiency and handles memory overhead as mentioned in the work [11]. The preparedness of CDR data to be analyzed also involves exclusion of unusualness, misleading datapoints or formats and disturbing samples as discussed in the paper [12].

The unsupervised methods for converting analog data points to digital points are implemented in cellular network. The approach used to achieve the same is peak value scattering record of the data points [13]. To integrate seasonality in the detection of the anomalies, a model called Seasonal Auto Regressive Integrated Moving Average (SARIMA) is proposed [14]. It is also important to determine window size in the anomaly detection. This is based on experimental analysis using network data which helps to build a convolutional model [15]. The CDR data gives essential information and the behaviour of mobile users. Efficient analysis in terms of both space and time can be conducted. Thus, it is possible to detect cellular traffic and can be grouped using suitable tools in machine learning which are experimented and mentioned in the paper [16]. Still there are many real time constraints in the implementation of cellular traffic detection systems [17].

With the thorough study it is notable that meticulous research is been conducted in terms of human free network management. It emphasizes on break free services including future generation. Initially it is important to convert the raw big CDR data into structured form making it ready for intended analysis. The former works identifies only high-rate traffic clusters and low-rate traffic clusters are not accounted. The identification and detection of low traffic data is also crucial for the management of resources and network. In this regard, K-means clustering algorithm integrated with optimization method Within Cluster Sum of Squares (WCSS) is proposed to detect both the types of traffic in the network which are accumulated due to various services like calls in – out, SMS in – out and internet activity.

3. FORMATTING OF THE CDR

The mobile service providers account all the details of the customers in the form of Call Detail Record. It comprises of all the users' activities and events with the time stamp. The CDRs are usually very private and concealed raw data. However, for the research purpose few CDRs are made available to analyse and solve the persistent and anticipated issues of mobile network with present and future generations. In this work, the CDR used is from area of Italia, Milan, Trentino with grid areas 235-meter squares of 1K and 6.575 K squares each [6]. It is a very huge data of about 300MB. In the initial stage of research, it is better to implement sampling. The conversion of unstructured data to formatted data involves cleaning, integration and transformation [18]. It is not necessary to trace all the steps, only required steps in suitable methods must be incorporated to increase the performance metrics making the data suitable for the needed analysis [19].

4. IDENTIFICATION OF TRAFFIC CLUSTERS USING WCSS OPTIMAL METHOD

The basic intent of network management in the proposed research is to ensure uninterrupted service to the customers [20], [21]. One among the tasks to achieve the same is the detection of mobile network with the investigation of formatted CDR. There are many types of methodologies available, however machine learning in this problem gives an emerging practical solution to the telecom network unlike ports and payloads methods [23]. This approach provides less/no human intervention with better metrics of performance.

Unsupervised method of K-means clustering detects the patterns of traffic in the network. Various methods are available for grouping the similar data points. However, a suitable one must be chosen basis the purpose of the objective. K-means clustering considers all data points with same preference. To minimize characteristic parameters, entropy approach is suggested as it facilitates minimum computations [24]. The same can be conducted in a semi-supervised way if minimal sudden changes happen in data points. This supports regularities in the clustering and reduces mathematical complexities which in turn saves time [25]. When there is ambiguity in the dataset points Within Cluster Sum of Squared optimization can be implemented. It is a parameter used in clustering to analyse compactness in the clusters. It helps to understand right number of clusters to be formed [26]. It computes sum of squared errors value before clustering.

The proposed work uses WCSS method to analyse mobile traffic caused by the customers including both low-volume and high-volume clusters. It is based on the dissemination of the data points in the CDR [27]. The metric chosen detects the optimal number of clusters in the network without any external parameter features [28]. If WCSS is low, then data points are closer to the centroid which helps in well integration of the clusters. As the number of clusters increases WCSS decreases which results in smaller and specific clusters. Thus, WCSS in K-means clustering is very powerful and suitable method to identify number of traffic clusters with correctness. Steps in the proposed algorithm is as follows:

1. The unformatted dataset CDR is converted into formatted structure.
2. The optimal number of traffic clusters is determined by WCSS.
3. The chosen optimal value is verified with subsequent formation of the clusters.

5. RESULTS AND DISCUSSION

The first step in detection of network traffic is to format the raw CDR dataset. It gives diffused and detailed events of the customers. The structured format is machine understandable form. Figure 1 shows the distribution information of the total SMS activities which conveys the number of SMS events recorded with activity score. It includes 100 events of both incoming and outgoing SMS. Activity score gives the detailed information of every event with time stamp. Similarly, Figure 2 shows the distribution information of the total calls' activities, Figure 3 shows the distribution information of the internet activities and Figure 4 shows the distribution information of the aggregated activities.

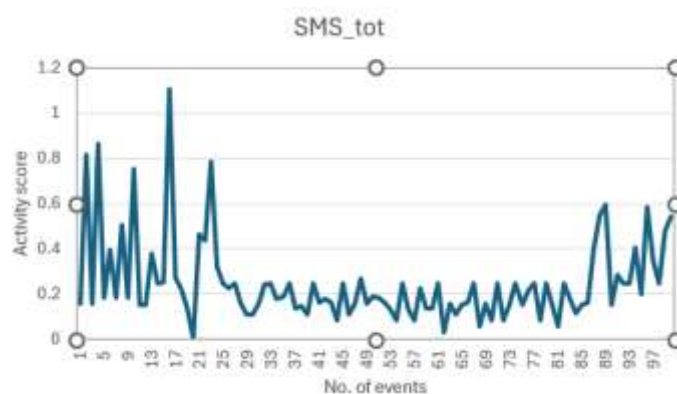


Figure 1 Total SMS activity/day

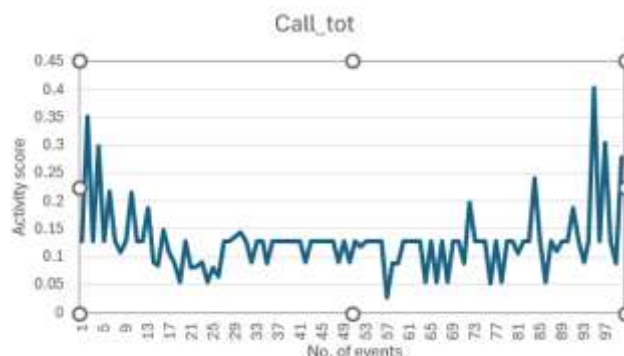


Figure 2 Total Calls activity/day

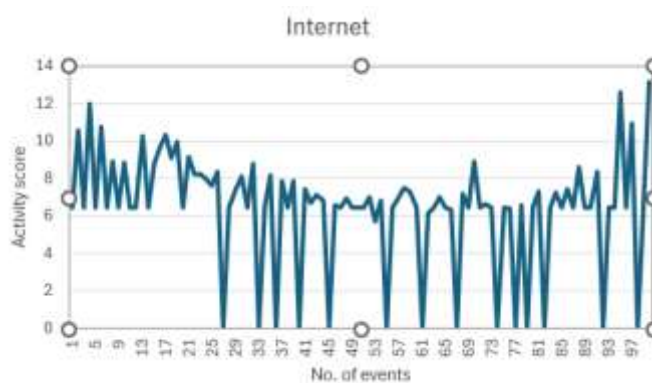


Figure 3 Internet activity/day

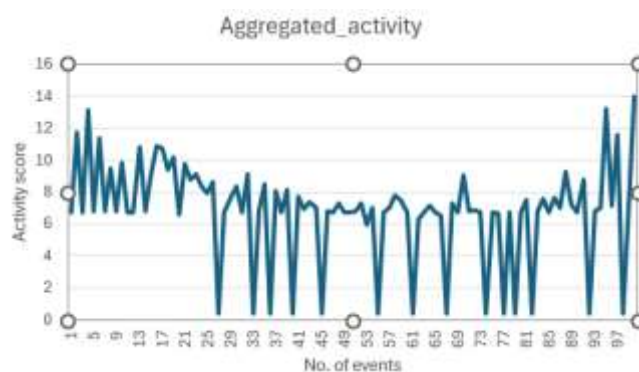


Figure 4 Aggregated activity/day

In addition to understanding of distribution of each activity, it is very important to find the number of traffic clusters in the mobile network. Within Cluster Sum of Squares optimization method is used to find the number of expected clusters. The method is verified by a plot of arm shape which represents the allowable inertia, if the number of clusters is assumed to be exceedingly more from practical value, then inertia is distorted. So, it is ideal to stop the iteration of cluster number.

Previous works detect heightened events in terms of traffic. The originality of the proposal is to detect flattened events as well, i.e., both high-volume and low-volume traffic events. The reasons for flattened or low-volume traffic could be like cell outages or sleeping cells, but they must be accounted. Figures 5 to 8 show both high-volume and low-volume clusters of traffic. Figure 5 shows traffic clusters of total SMS activity. Larger clusters are formed at 39 area code comprising both the heightened and flattened traffic data. Similarly, Figure 6-8 shows traffic clusters of total calls activity, internet activity and aggregated activity respectively. Each activity has 4 traffic clusters including both high and low volumes. WCSS is calculated for every event and is plotted against the number of clusters using the following equation.

$$WCSS = \sum_{i=1}^k \sum_{x \in C_i} ||x - \mu_i||^2$$

Figure 9 shows WCSS optimization method which verifies the clusters numbers. The shape of the graph like an arm prompts the number of traffic clusters to be chosen.

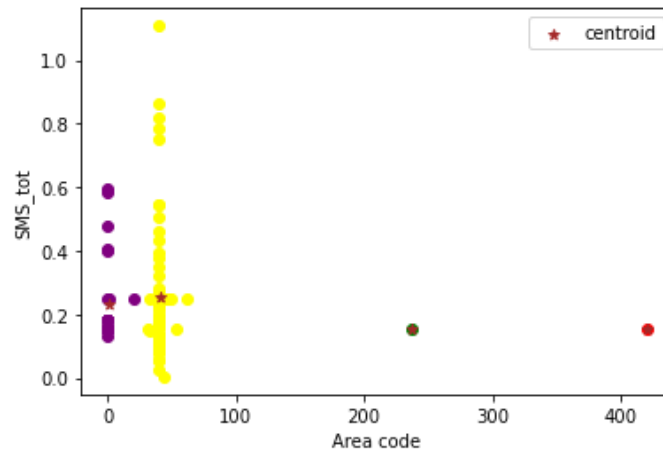


Figure 5 Total SMS clustering/day

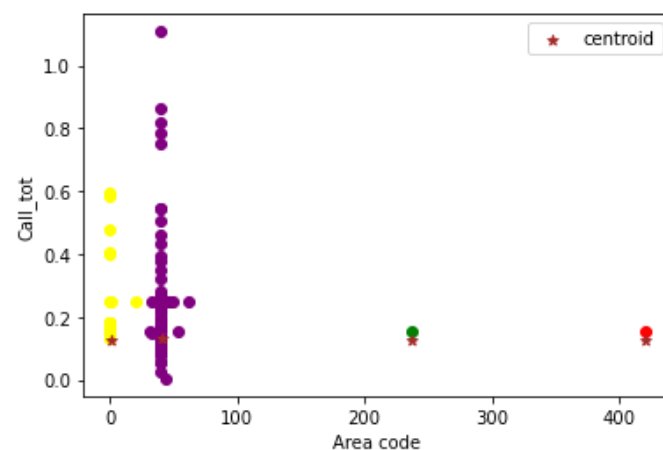


Figure 6 Total Calls clustering/day

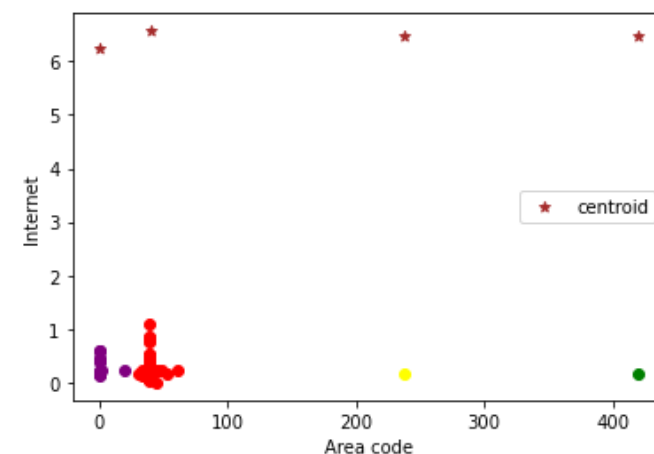


Figure 7 Internet clustering/day

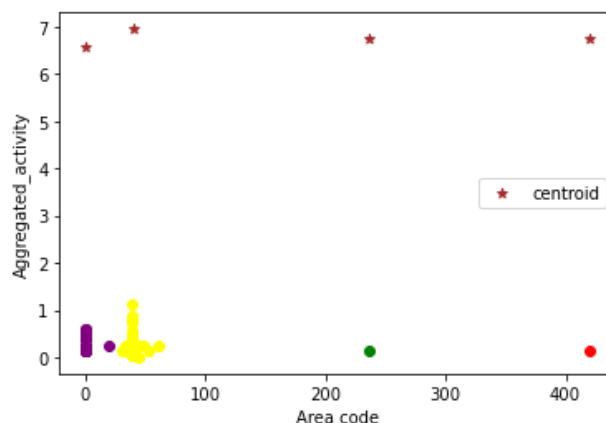


Figure 8 Aggregated activity clustering/day

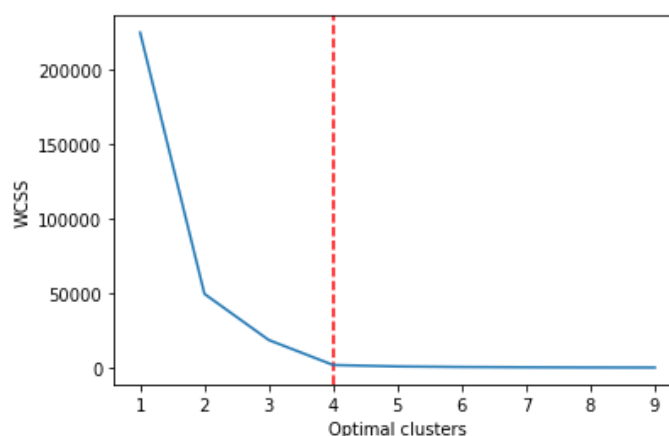


Figure 9 WCSS optimization method for clusters verification

The proposed WCSS optimization method is compared with the other methods as well, like, Hierarchical clustering with Principal Component Analysis (PCA). Table 1 gives the details of the evaluation parameters used for traffic clusters analysis in the mobile network in comparison with WCSS optimization method. It strongly clarifies that the proposed method is the best possible way to meet the objective irrespective of any volume of traffic.

Table 1. Comparative analysis of WCSS with other combination of tools

Evaluation parameters	WCSS optimization method	Hierarchical clustering with PCA
Traffic volume detection	Detects high-volume and low-volume traffic clusters	Detects only high-volume traffic clusters
Individual traffic analysis	Individual events detected.	No individual activities are detected but gives statistical relation among the events
Cost & Complexity	Cost effective with less complexity	Expensive with complexity

The previous work of detecting network traffic clusters is executed using Hierarchical clustering with PCA. An analysis based on various evaluation parameters is conducted to compare the former method along with the proposed optimization method. Hierarchical clustering with PCA detects high-volume traffic without any individual event identification. The analysis is very complex and not cost-effective. Whereas the WCSS optimization method detects both high-volume and low-volume traffic in the network. Analysis is much suitable for larger CDR datasets and is budget friendly. The method is also verified with graph of an arm shape obtained with the chosen number of clusters and inertia value.

6. CONCLUSION AND FUTURE WORK

The paper reveals about formatting the CDR dataset and detection of traffic in the network using WCSS optimization method. The method is suitable for individual events like total SMS, total call, internet activity and for aggregated activity. It shows that every event has 4 active clusters of traffic in the network which include both high and low-volume traffic. Along with the detection of traffic clusters, it allows to verify the iteration value. This is so much important for the mobile network services to ensure continuity of connection to the customers with minimal usage of resources.

Future work can be designed to understand the reason behind the cause of low-volume traffic in the network. It can also be extended to classify the cause as either cell outages or sleeping cells or inactive users in the network.

ACKNOWLEDGEMENTS

The authors express gratitude to the affiliated institutions for the facilities afforded to conduct the research.

AUTHOR CONTRIBUTIONS

Sandhya B S: Conceptualization, Investigation, Laboratory Analysis, Methodology, Data Curation.

Pavana H: Formal Analysis, Writing—Review & Editing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Ethical approval was not required for this study.

DATA AVAILABILITY STATEMENT

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

- [1] Larry E. Daniel, Lars E. Daniel, "Cellular system evidence and call detail records", ScienceDirect, pp. 225-237, year: 2012, doi: <http://dx.doi.org/10.1016/B978-1-59749-643-8.00033-X>.
- [2] Larry E. Daniel, Lars E. Daniel, "Discovery of call detail records", ScienceDirect, pp. 163-165, year: 2012, doi: <http://dx.doi.org/10.1016/B978-1-59749-643-8.00024-9>
- [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, doi: <https://doi.org/10.1016/B978-0-12-805467-3.00003-X>.
- [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, doi: <https://doi.org/10.1109/TNSE.2018.2843326>.
- [5] Nirmal Ghotekar, "Analysis and Data Mining of Call Detail Records using Big Data Technology", IJARCCCE, vol. 5, Issue 12, pp. 280-283, December 2016, doi: 10.17148/IJARCCCE.2016.51264.
- [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, doi: <https://doi.org/10.1109/ACCESS.2018.2859756>.
- [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, doi: 10.1109/iwcmc.2019.8766728.

- [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, doi: <https://doi.org/10.48550/arXiv.1808.02342>.
- [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, doi: <https://doi.org/10.1109/TII.2017.2650206>.
- [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, doi: <https://doi.org/10.1109/CC.2018.8357700>.
- [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, doi: <https://doi.org/10.1109/PACRIM47961.2019.8985068>.
- [12] Kashif Sultan, Hazrat Ali, Zhongshan Zhang, "Big Data Perspective and Challenges in Next Generation Networks", MDPI, year: 2018, doi: <https://doi.org/10.3390/fi10070056>.
- [13] Pekka Kumpulainen, Kimmo Hätönen, Pekko Vehviläinen, "Automatic Discretization In Preprocessing For Data Analysis In Mobile Network", Research gate, pp. 813-816, year: 2003, doi: https://www.researchgate.net/publication/228956812_Automatic_Discretization_in_Preprocessing_for_Data_Analysis_in_Mobile_Network.
- [14] Anish Nediyanath, 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, doi: <https://doi.org/10.1109/WCNCW48565.2020.9124843>.
- [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, doi: <https://doi.org/10.1609/aaai.v33i01.33011409>.
- [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, doi: <https://doi.org/10.3390/info10060192>.
- [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, doi: <https://doi.org/10.1109/COMST.2019.2924243>.
- [18] 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, doi: <https://doi.org/10.1109/ICBDA49040.2020.9101257>.
- [19] Siyang Qin, Youchen Zuo, Yaguan Wang, Xuan Sun Honghui Dong, "Travel Trajectories Analysis Based on Call Detail Record Data", IEEE Chinese Control And decision Conference, pp. 7051-7056, year: 2017, doi: <https://doi.org/10.1109/CCDC.2017.7978454>.
- [20] 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, doi: <https://doi.org/10.1109/ACCESS.2020.3010637>.
- [21] Dainotti, A. Pescapé and K. Claffy, "Issues and future directions in traffic classification", IEEE paper, vol. 26, Issue 1, pp. 35-40, year: 2012, doi: <https://doi.org/10.1109/MNET.2012.6135854>.
- [22] 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, doi: https://link.springer.com/chapter/10.1007/978-3-540-31966-5_4.
- [23] 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, doi: <https://doi.org/10.1109/SURV.2013.100613.00161>.
- [24] 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, doi: <https://doi.org/10.1109/ACCESS.2021.3077622>.

- [25] 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, doi: <https://doi.org/10.1109/TKDE.2019.2952596>.
- [26] 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, doi: <https://doi.org/10.1109/TFUZZ.2020.2966182>.
- [27] 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, doi: <https://doi.org/10.1109/CHICC.2006.4347538>.
- [28] Kristina P. Sinaga, Miin-Shen Yang, "Unsupervised K-Means Clustering Algorithm", IEEE paper, vol. 8, pp. 80716-8-727, year: 2020, doi: <https://doi.org/10.1109/ACCESS.2020.2988796>.