Analysis Of Mobile Users' Activities Using Mean-Normalization Method

Sandhya B S* and Rohini Deshpande

Reva University, India

Corresponding author. E-mail: sandhya.navaneetham@gmail.com

Received: Oct. 06, 2022; Accepted: May. 25, 2023

In recent decades, with the attracting features of mobiles including 4G and 5G, world is getting more connected to mobile communications. This results in the accumulation of large amount of data in the mobile network. The analysis of the network data is very complex but is essential in terms of resource and cost management. The network data analytics include detection of unusual network behaviour due to traffic created by the mobile users and Short Message Service (SMS) spammers. Research to an approach with the same impulsion is creating a new interest in the field of mobile network data analytics using machine learning tools. To attain this, Call Detail Record (CDR) provided by the telecom network industry is utilized. The timely analysis of CDR helps to understand the behaviour of the network due to various activities of mobile users. To analyse CDR, it has to be pre-processed to convert it from the raw data into machine understandable form. The proposed method is mean-normalization pre-processing which is suitable in understanding the behaviour of mobile users' individual activities like incoming-outgoing calls, incoming-outgoing SMS and internet activity. Later, machine learning tools can be applied to analyse and predict the network anomalies like network traffic and Short

Keywords: Call Detail Record, machine learning tool, mobile user activity, network anomalies, pre-processing

http://dx.doi.org/10.6180/jase.202402_27(2).0012

1. Introduction

With the increased population of mobile users and the emergence of new technology there is an immense accumulation of data in the network. This big data must be analysed timely and effectively for the better network resource management and customer service. The big data accumulated in the network is provided by the service companies as Call Detail Record-CDR. The CDR gives the transactions of telecom department and details of all the event details that occur in the network. CDR also provides detailed information of the telecom transactions like incoming and outgoing calls, incoming and outgoing SMSs and internet activities of the customers including time stamps. CDR includes the beginning of a call or SMS or internet activity and termination of the event. At this point other elements of the network are also connected. At the later level CDR has to enter into another phase where conversion has to happen from raw form to processed form [1]. Then the processed form of CDR can be used for further analysis with the application of machine learning tools. It is also very challenging to transfer and store the CDR data in the network and data centres [2]. However, the analysis of CDR data gives the knowledge of mobile user behaviour and network behaviour as well. To analyse the CDR data, it is required to convert it into required processed form. There are various pre-processing techniques that can be applied to the big data. In this regard, a suitable technique must be selected to apply for CDR to process it and make it ready for the further required analysis.

However, while pre-processing the characteristics of the CDR must also be considered. There are 4 important characteristics. They are volume, velocity, variety and value [3].

Volume: CDR is a huge volume data of order of Pico

bytes and is continuously increasing.

Velocity: It is the speed of the data entering and exiting the mobile devices and mobile network.

Variety: It refers to the different forms of data like structured, unstructured and semi structured.

Value: The values are the numerical features that should be extracted from the CDR data using suitable design proposed by the big data technologies group.

The rest of the paper is organised as follows. Initial sections give the related work, description of call details record and pre-processing challenges. Next formatting and sampling of CDR data is explained. Also, briefed about quality assessment and features selection of CDR data. In the later section pre-processing of CDR is explained along with results and discussions. Finally, the paper is concluded with the future work.

2. Related work

Distributed clustered based analytics in Hadoop system is proposed by Nirmal Ghotekar for pre-processing the CDR. The required CDR field is selected to perform extract, transform and load operations in the Hadoop system to obtain the normalized monthly traffic of each base station to its maximum [1].

Chenhan Xu et al. designed a big data analytics model to solve big data redundancy problem in data centres. Classification accuracy is increased using several inputs. Also, the proposed model saves the computational resources [2].

Kashif Sultan et al. proposed a model to predict the future traffic in the network which serves users. In this CDR data is pre-processed by removing the irregularities such as missing entries or noise which misleads the patterns [3].

Md Salik Parwez et al. combined activities of all the calls and texts by the users at every grid and every time stamp in the final pre-processed CDR data set [4].

Bilal Hussain et al. carried out the data pre-processing by removing irrelevant parameters. Initially the CDR data is cleaned by filling entries. Also, the total activities of the network are combined and transformed into the format acceptable by the algorithm for the further analysis [5].

Ramin Sharif et al. carried out pre-processing using Horton works sandbox. All the empty data fields are set to zero. All the CDR activities are summed up into one single activity. Also, the time stamps are summed to 1 hour to save the memory and processing power [6].

Kashif Sultan et al. stated that data has to be prepared well before performing any analysis. It helps in the reduction of processing overhead and improves the quality of the output. The missing values are replaced by the average values of the previous and next time stamp values [7].

Muna Al-Saadi et al. stated that to analyse traffic patterns pre-processing is done statistically using teptrace which includes data cleaning, normalization and reduction. Principal Component Analysis (PCA) tool is used to remove the unwanted parameters and Z score normalization to scale the parameters [8].

3. Call detail record and its description

A call detail record is the data set produced by the network operator. The CDR used in the work is issued by Telecom, Italia in Milan, Trentino. The CDR data set type is grid type. The area of Milan is a grid of 1000 squares and the area of Trentino is a grid of 6575 squares. Each square is an area of about 235 square meters [3].

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.

4. Data pre-processing challenges

There are many challenges involved while pre-processing the big data. The data collected usually will be having incompletely filled fields and contains noise as well. It may include invalid data and personal information. Data cleaning and pre-processing helps to remove this irrelevant information [9]. After the elimination of the unwanted data the accuracy and the reliability must be ensured [10]. There are basic steps which must be adapted while preprocessing the data. However not all the steps have to be followed in the pre-processing. According to the requirement pre-processing steps can be adapted. The following are the basic steps [11].

A. Quality test of the data set

The CDR data set must be investigated for any limitations and flaws during the collection of the data and measurement. The collected CDR data must be tested for the human errors. The CDR data set must be checked for any missing values. The entire rows of the missing values must be eliminated or estimated with suitable methods. The CDR data set must also be examined for the irrelevant values. For example, in the place of phone number of the mobile user the address of the user could be entered. Also, the repeated entries must be removed. Thus, before starting the actual procedure of pre-processing, the missing values, irrelevant values and repetitive values must be eliminated.

B. Accruement of the data set characteristics

To get a better preview of the collected CDR data set,

it is very important to group and accumulate all the characteristics in the data set timely manner, monthly, bi-yearly, or yearly. This helps in the better account management. Also, it helps in the less memory consumption in a system, decreases the processing time and complexity.

C. Sampling of the data set
Sampling of the collected CDR data set is a vital step in the pre-processing procedure. In the huge data set, the sample which is a subset has to be selected carefully. The sample must be the representative of the whole data set. It must have almost all the properties of the entire data set for the effective pre-processing. While sampling the data set, the other attributes like cost, memory and time must also be accounted. Usually, random sampling is the common strategy used in big data pre-processing [12]. Selection of the sample as a representative of the entire data set is also challenging considering less computational time and cost [13]. In this regard, sampling algorithm must be developed to get a small replicative training set and to estimate the required parameters [14, 15].

D. Reduction of features in the data set
This is an optional step in the pre-processing of the data set but very effectual if implemented. The CDR data set can be checked for any features which can be reduced or even removed if they are not used at all for the future analysis. This step removes noise in any data set and gives the better visualisation for the data analysis. If the data set size is reduced, then the analysis will be easier with less complexity.

E. Encoding of the data set
The encoding of the CDR data set is the final step in the pre-processing. Here the features of the data set are converted into machine understandable form while reserving the originality. There are many methods available for encoding the data set and the best suitable method to be chosen.
During this encoding process we need to understand and consider the following rules.
There are two types of encoding depending on the relationship among the variables.

- Nominal variable encoding: This encoding is used when there is a set of discrete variables but they are independent of each other.
- Ordinal variable encoding: This encoding is used when there is a set of discrete variables but they are dependent variables. In ordinal coding, one-hot encoding can also be considered if any variable is removed after encoding and a new variable is added.
  There are two important points while we are handling numerical variables in encoding the data set.
- If there is an interval where in the difference between the consecutive values are co-related, then we need to consider the below equation.
  present value = previous value*constant A + constant B
- If there are mathematical variables having the meaningful ratios, then we need to consider the below equation.
  present value = previous value*constant A

After encoding the data set, it must be divided into 2-3 parts before applying any machine learning algorithm. They are training data, validation data and test data.

- Training data: Using this data set, the machine learning algorithms are trained. The model learns the features and characteristics of the data set. The model can also analyse whether the data is over fitting or under fitting the model.
- Validation data: In this part, the model can validate itself by understanding the other parameters which do not change the performance of the model except the speed and quality of the learning process.
- Test data: Using this part of the data set, the hypothesis model is tested.

The data set must be split into specific ratio according to the required model. In most of the scenarios much training is required therefore, the training data set is kept in a larger ratio, then the remaining is split into validation data and test data. If the other parameters which are used to improve the performance are more in number, then the validation data proportion can be increased accordingly. Usually, the data split ratio recommended is 70% training data and remaining 30% of the data set into validation data and test data.

There are also other challenges in pre-processing which cannot be ignored [16]. They are:

- Data pre-processing has to be conducted in real time.
- Suitable technique is required to pre-process the data to understand the concealed problems.

- Proper platform has to be chosen for data pre-processing as per the requirement.
- Data pre-processing might be a repetitive process case by case which consumes time.
- Fig. 1 shows the framework of the proposed pre-processing technique i.e., meannormalization method after data collection, formatting and sampling. Later the preprocessed data can be analysed using suitable ML tools.

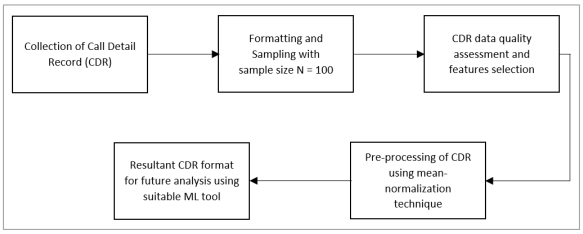

Fig. 1. Framework of the proposed technique

5. Cdr data set formatting and sampling

The available and collected CDR data sets of 5 different dates of Trentino grid is in text format with the size of about 300MB. The initial view of the data is in a very scattered manner and is in the non-understandable manner. In this concern the primary goal is to convert the CDR data set into comprehensible form. The data set is checked for the errors and are removed while formatting. After this step, the data is viewed using suitable viewer without changing the data format to understand the hidden numerical features.

In the next step, the sample of the CDR data is selected for the pre-processing. The sample size is chosen as N = 100 for all the 5 different data sets. Now the size of each data set is reduced to 8 KB, and it is appropriate for the testing purpose.

6. Cdr data quality assessment and features selection

Initially all the 5 sampled CDR data sets of different dates are checked for the human errors. The CDRs are made to undergo the primary scanning process which is the first step to remove all the repeated values. In the next step, it is the removal of the unwanted features which may not be useful for the aimed analysis. In the CDR, there are 8 numerical features. Not all parameters are required for the analysis. Hence, the irrelevant features can be removed. Table 1 shows the details of the numerical features that are retained for the analysis and about the features that are not used in the pre-processing and further analysis.

7. Pre-processing of the call detail records

Pre-processing is carried out to convert the raw CDR data set into the processed form. After pre-processing, the CDR will be ready to use for the future required analysis and will be in the machine understandable form. Here the pre-processing is done by using the meannormalization algorithm which is run by the python code. In this method, the mean value for all the retained parameters is estimated and all the missing values in the CDR are normalized. Thus, the data now can be used to understand the mobile users' activities. Below Table 2. shows the mean values of all the attributes of 5 CDRs that are used in the pre-processing.
Algorithm for the proposed pre-processing technique is given below.

Step 1: Collect the Call Detail Record (CDR) data set.

Step 2: Sample the CDR with sample size $N = 100$.

Step 3: Identify the CDR activity columns of Call-in, Call-out, SMS-in, SMS-out and Internet activity.

Step 4: Remove the duplicate or repetitive values in all columns.

Step 5: Estimate the mean value for every identified column.

Step 6: Normalize the empty fields using the mean value.

Step 7: Analyse the distribution of the pre-processed CDR activities using suitable machine learning tool.

8. Results and discussion

The distribution of the mobile users' s activities is simulated after pre-processing for all the 5 sets of CDRs. Each activity is separately considered with all the country codes available in the data set. For example, below are the graphs obtained for the CDR dated 05/07/2013 for all the activities of mobile users i.e., call-in activity, call-out activity, SMS-in activity, SMS-out activity and internet activity versus country codes available in the original data set. For detailed explanation, in Fig. 2 the graph shows the distribution of the incoming call activity by the mobile users over all the country codes available in the data set. It evidences variations from the lowest value 0.01 to the highest value 1.01 for the incoming call activity. Similarly, Figs. 3 to 6 describe the outgoing calls, incoming -outgoing SMS and internet activity.

Table 1. Numerical features of CDR with requirement factor for pre-processing.

SL. No	Call Detail Record attributes	Requirement factor (Used/Removed)
1	Call in activity	Used
2	Call out activity	Used
3	Internet activity	Used
4	Square grid ID	Removed
5	Country code	Used
7	Time stamp information	Removed
8	SMS in activity	Used

Table 2. Mean values of all the activities for 5 days.

Date **2013**	Call-in activity	Call-out activity	SMS-in activity	SMS-out activity	Internet activity
01/07	0.1	0.1	0.1	0.1	6.9
02/07	0.1	0.1	0.1	0.1	5.8
03/07	0.1	0.1	0.1	0.1	5.6
04/07	0.2	0.2	0.2	0.2	5.8
05/07	0.2	0.1	0.2	0.12	6.2

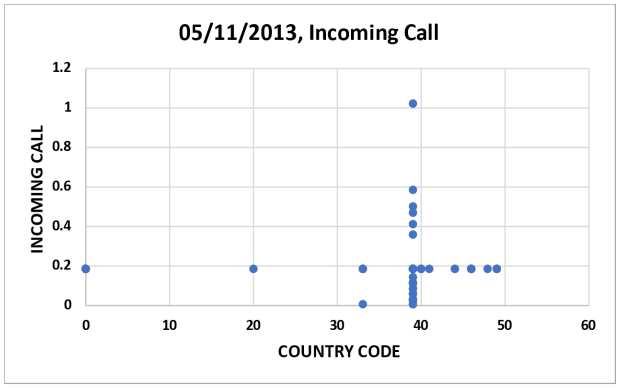

Fig. 2. Incoming calls activity on 05/11/2013.

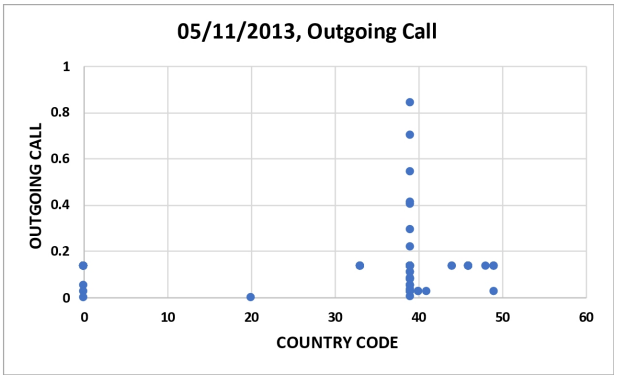

Fig. 3. Outgoing calls activity on 05/11/2013.

With the above tabulations in Table 3 we can get details of the lowest and highest records examined in the scattering of in-coming and out-going calls, in-coming and out-going SMS and internet activity.

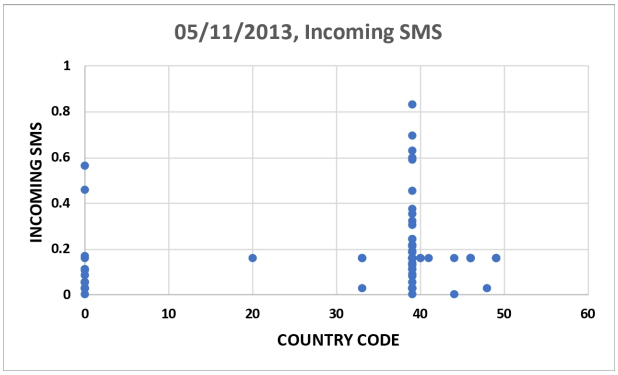

Fig. 4. Incoming SMS activity on 05/11/2013.

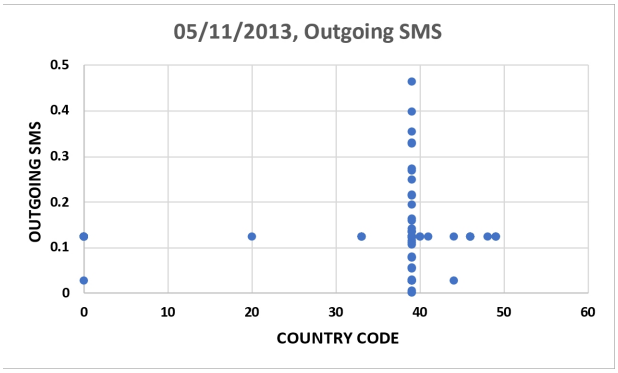

Fig. 5. Outgoing SMS activity on 05/11/2013.

In the previous work [3], only high traffic density of users is analysed. The proposed preprocessing mean-normalization method is suitable for analysing both high and low traffic densities using the same CDR data set.

The pre-processed data is simulated to understand the

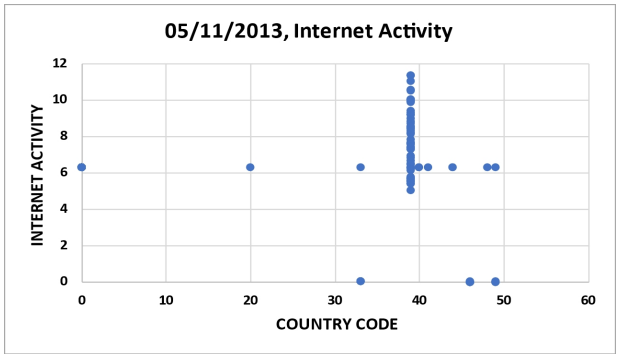

Fig. 6. Internet activity on 05/11/2013.

Table 3. Lowest and highest value records for the CDR activities.

Activity/ Parameter	Lowest value record	Highest value record
Call in activity	0.01	1.01
Call out activity	0.01	0.84
SMS in activity	0.01	0.83
SMS out activity	0.01	0.46
Internet activity	0.01	11.37

density of the mobile CDR activity. The minimal and the maximal value of the CDR event recorded for all the mobile activities are shown in the below graphs.

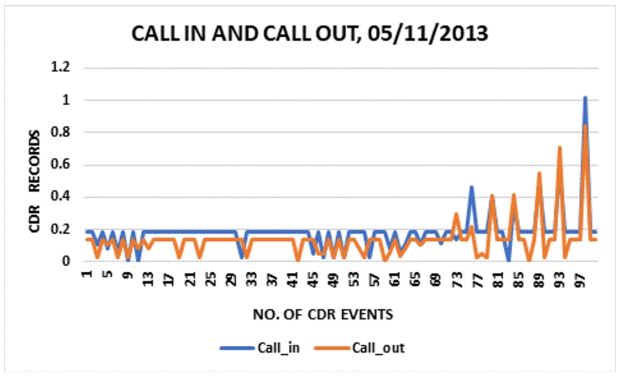

Fig. 7. Call in and call out density distribution analysis.

9. Conclusion and future work

The paper presents the pre-processing of CDR using mean-normalization method. This helps in analysing network behaviour due to incoming calls, outgoing calls, incoming SMS, outgoing SMS and internet activity. The analysis also interprets the density of every CDR activity. The pre-processed data is ready for further analysis to understand the unusual behaviour in the network.

In future, with the application of machine learning tools to the pre-processed data, the mobile network anomalies

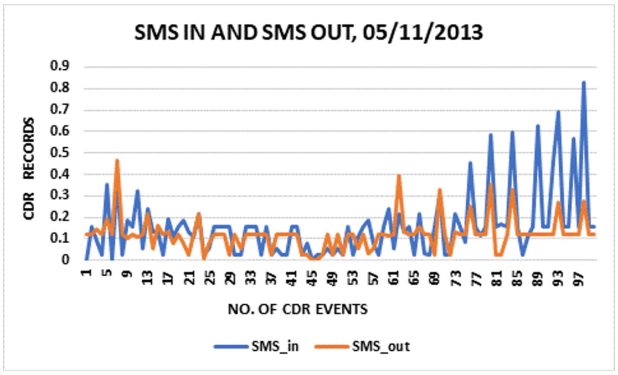

Fig. 8. SMS in and SMS out density distribution analysis.

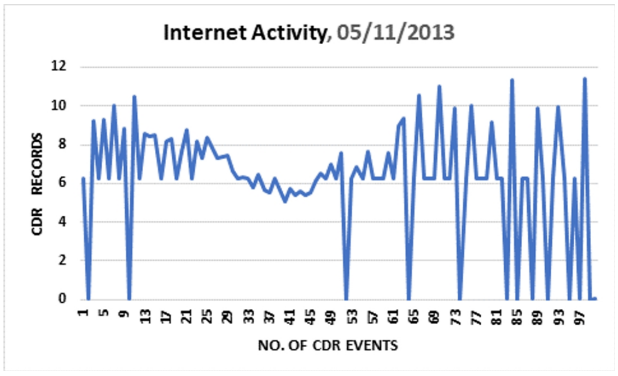

Fig. 9. Internet activity density distribution analysis.

of the different mobile users' activities can be analysed. Specifically, the future traffic in the network can be predicted. This serves the mobile network service providers and the customers in terms of resource and cost management.

Acknowledgements

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

References

[1] N. Ghotekar, (2016) *"Analysis and Data Mining of Call Detail Records using Big Data Technology"* **IJARCCE** *5*(12): 280–283.

[2] C. Xu, K. Wang, Y. Sun, S. Guo, and A. Y. Zomaya, (2018) *"Redundancy avoidance for big data in data centers: A conventional neural network approach"* **IEEE Transactions on Network Science and Engineering** *7*(1): 104–114.

[3] K. Sultan, H. Ali, and Z. Zhang, (2018) *"Call detail records driven anomaly detection and traffic prediction in mobile cellular networks"* **IEEE Access** *6*: 41728–41737.

[4] M. S. Parwez, D. B. Rawat, and M. Garuba, (2017) *"Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network"* **IEEE Transactions on Industrial Informatics *13***(4): 2058–2065.

[5] B. Hussain, Q. Du, and P. Ren, (2018) *"Semi-supervised learning based big data-driven anomaly detection in mobile wireless networks"* **China Communications *15***(4): 41–57.

[6] R. Sharifi, M. M. Majdabadi, and V. T. Vakili. "Mobile user-activity prediction utilizing LSTM recurrent neural network". In: *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*. IEEE. 2019, 1–7.

[7] K. Sultan, H. Ali, A. Ahmad, and Z. Zhang, (2019) *"Call details record analysis: A spatiotemporal exploration toward mobile traffic classification and optimization"* **Information *10***(6): 192.

[8] M. Al-Saadi, B. V. Ghita, S. Shiaeles, and P. Sarigiannidis. "A novel approach for performance-based clustering and management of network traffic flows". In: *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*. IEEE. 2019, 2025–2030.

[9] L. Kesheng, N. Yikun, L. Zihan, and D. Bin. "Data mining and feature analysis of college students' campus network behavior". In: *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*. IEEE. 2020, 231–237.

[10] S. Qin, Y. Zuo, Y. Wang, X. Sun, and H. Dong. "Travel trajectories analysis based on call detail record data". In: *2017 29th Chinese Control And Decision Conference (CCDC)*. IEEE. 2017, 7051–7056.

[11] P. Pandey. *Data Preprocessing: Concepts*. 2019. URL: https : / / towardsdatascience . com / data - preprocessing - concepts - fa946d11c825 (visited on 11/25/2019).

[12] M. S. Mahmud, J. Z. Huang, S. Salloum, T. Z. Emara, and K. Sadatdiynov, (2020) *"A survey of data partitioning and sampling methods to support big data analysis"* **Big Data Mining and Analytics *3***(2): 85–101.

[13] S. Salloum, J. Z. Huang, and Y. He, (2019) *"Random sample partition: a distributed data model for big data analysis"* **IEEE Transactions on Industrial Informatics *15***(11): 5846–5854.

[14] M. Li, H. Wang, and J. Li, (2019) *"Mining conditional functional dependency rules on big data"* **Big Data Mining and Analytics *3***(1): 68–84.

[15] A. Alim and D. Shukla. "A Parameter Estimation Model of Big Data Setup Based on Sampling Technique". In: *2nd International Conference on Data, Engineering and Applications (IDEA)*. IEEE. 2020, 1–5.

[16] R. Jony et al. *"Preprocessing solutions for telecommunication specific big data use cases"*. (mathesis). 2014.