ATLANTIS PRESS
Advances in Computer Science Research (ACSR), volume 72
IV International Research Conference "Information Technologies in Science, Management, Social Sphere and Medicine"
(ITSMSSM 2017)

Clustering helps to determine the changes in telecom subscribers' behavior

Alexey Golubev, Anh Tuan Ngyuen, Maxim Shcherbakov, Tran Van Phu
Computer Aided Design,
Volgograd State Technical University, VSTU
Volgograd, Russia
maxim.shcherbakov@vstu.ru

***Abstract*— Telecom company tries to find out new ways for developing new personal-oriented services based on data analysis. Changes in telecom users behavior are the objects of investigation. If a user changes its service usage, the company should pay attention to the fact and minimize the negative outcomes (e.g. churn analysis). We propose a method, which allows identifying changes in user behavior based on analysis of Call Detail Records data. The main idea of the method is using clustering techniques to determine the clusters with typical user behavior. Since the clusters exist and all users are labeled in terms of cluster belonging, the new data about user behavior compare with a typical profile. Open data was used in use cases. The results show the outperform of the proposed method in comparison with benchmark model.**

Keywords—data analysis; CDR; telecommunication services; time series forecasting; MeanShift; KMeans; clustering

I INTRODUCTION

Providing of high-quality telecommunication services (TS) in the conditions of (1) the intensity increasing of the service using, especially mobile internet (2), the increasing risks associated with information security (3), the emerging need for personifying services for the needs of specific users is an actual strategic objective of modern telecommunications enterprise.

To ensure the continuity and high quality of TS provision in telecommunication companies (TC), corporate standards are developed that regulate the creation, implementation and management of TC core and supporting business processes. The international practice is the formalization of business processes in the format of the expanded process map of the telecommunications company eTOM (enhanced Telecom Operations Map), which is part of the program of the international consortium TM Forum NGOSS (Next Generation Operations Systems and Software). Within this framework, processes and systems for supporting the management of service provision are identified, for which corporate information systems and automation systems are proposed and implemented. Problems related to data analysis are usually solved by highly qualified specialists using business intelligence tools (Business Intelligence). Based on the results of the analysis, management decisions are made, including modifications of the business processes of providing services.

In the main and auxiliary processes, it is necessary to distinguish invariant operations in the processes associated with the personification of the provision of services, i.e. activities aimed at maximum satisfaction of communication needs of a particular subscriber. This is a new approach for TC, aimed at finding the optimal service plan for the user based on the activities and preferences of the latter (the strategy for developing personalized services to increase loyalty). To proceed to the provision of personalized services, it is necessary to study the behavior of [1] the subscriber, form a proposal for the provision of services in accordance with the company's development strategy and expectations for achieving the goals set, to realize the service. This requires the search for new processing methods [2] of the available data. In this regard, there is an urgent scientific task related to improving the methods of processing data in the processes of providing and managing personalized services to improve the efficiency of servicing the subscribers of a telecommunications enterprise.

There are a significant amount of work on modeling the behavior of subscribers in TC, as well as the formation of subscriber profiles [3-6]. As a result of the analysis, a number of topical problems are presented. One of them is the identification of changes in the behavior of subscribers. The solution of this problem plays an important role in the development of systems for the provision of personalized services by telecommunications enterprises, for example, (1) fraud detection systems; (2) systems for detecting and reducing outflow; and (3) consumer consumption forecasting systems.

II A METHOD

The purpose of the presented method is to identify a change in the behavior of subscribers on a predetermined forecast horizon. It should be noted that (1) typical approaches to identifying changes based on the forecast of the subscriber's behavior show unsatisfactory results due to the uncertainty in the time of making the call (using the service) and (2) the response time to the change in the subscriber's behavior depend on the data discretization. We considered various approaches to detect changes in the behavior of subscribers. In particular, methods based on user behavior forecasting on observation period. Autoregressive models were used for forecasting of subscriber's behavior (e.g. duration of calls in forecast horizon). At the same time, the method of searching for optimal

autoregressive model parameters was used with a brute force search based on grid search technique.

The proposed method is based on the following idea. If the subscriber's profile (usage of telecom services during the specific time interval) differs from typical profiles of subscribers of a similar cluster, then we suggest that there is a situation where subscriber's behavior change is observed. The new method includes two stages: the stage of clustering (labelling) and the stage of changes identification.

The stage of clustering consists of the following steps.

Step 1. Set the hyperparameters of clustering models: [*ts, tk*] – interval of observation, where *ts* – a time stamp of the beginning of observation, *tk* – a time stamp of the end of observation; *h* – is the length of the short observation interval within the interval [*ts, tk*]; *b* – the number of chunks on a short observation interval; *k* – the number of clusters. Set n_{cx}= 0.

We assume that the number of clusters is determined on the basis of the formula k = int (n/q), where int (·) is the rounding operation of the number to the integer, q is the empirically selectable coefficient, and we assume that $q<n$. A small number of clusters can lead to a situation where almost all subscribers profiles are grouped in the same cluster. Figure 1 shows examples of the distribution of profiles across clusters for a different number of clusters.

Step 2. The initial dataset has been modified based on hyperparameters initial values [*ts,tk*], *h*, *b*.

Step 3. Select the subscriber profiles with values equal to zero, and assign these profiles label of the Cluster id #0.

Step 4. Exclude subscribers with Cluster id #0 from the data sample.

Step 5. Build an ensemble of clustering models: k-means, MiniBatch K-Means, MeanShift (for the latter, the number k is not required).

Step 6. Search for the best variants of clustering models, using several launches and assessing the quality of clustering. The quality of each clustering model is evaluated in accordance with the Silhouette coefficient.

Step 7. Identify the centroids of clusters and assign the cluster ID to a user.

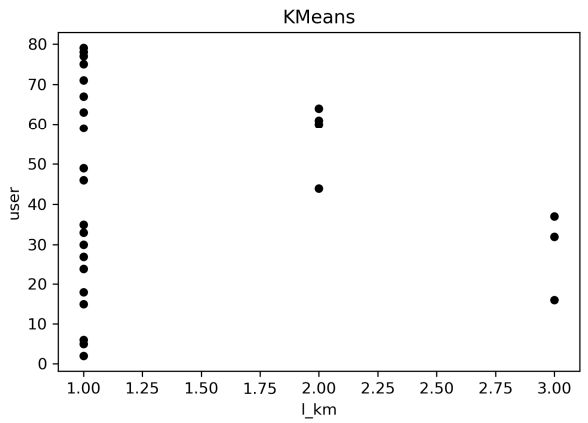
Stage of changes identification includes the following steps.

Step 1. Obtain the actual data for the analyzed subscriber. Data is modified according to the hyperparameters h, b for which the models are calculated.

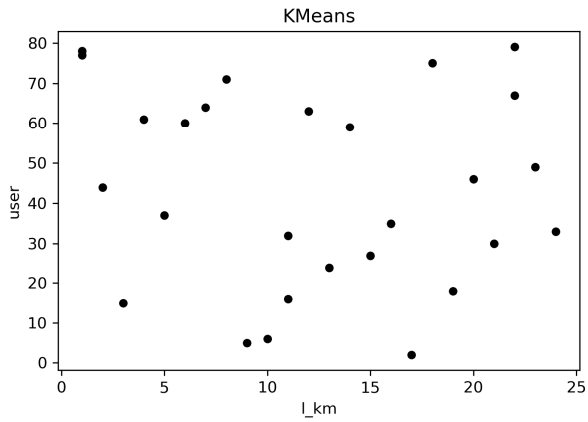
Step 2. Obtain the label of a cluster for analyzed subscribe #idc. Since profiles of subscriber were used in clustering, we assume, that all subscribers have its label.

Step 3. For all subscribers with the specified cluster label #idc select all the profiles included in the cluster #idc.

Step 4. For selected cluster id #idc calculate the lower bound, upper bound and average profile for each chunks in every short observation period. The lower bound is calculates according to the formula $x_{\mathrm{low}}^{(\mathrm{idc})}[b_i]=\min(x_j^{(\mathrm{idc})}[b_i])$. The upper bound is calculates according to the formula $x_{\mathrm{up}}^{(\mathrm{idc})}[b_i]=\max(x_j^{(\mathrm{idc})}[b_i])$. An average value is obtained using $x_{\mathrm{avg}}^{(\mathrm{idc})}[b_i]=\mathrm{avg}(x_j^{(\mathrm{idc})}[b_i])$. Note, that j=1,…, n_{idc}, n_{idc} – a number of profiles in cluster with id #*idc*.



a)



b)

Fig. 1. Distribution of subscribers among the clusters for (a) 3 clusters and (b) for 24 clusters.

Step 5. Check: if at least one current value in the profile is less than the lower bound, or greater than the upper bound, then assume that the behavior of the user has changed, and put $n_{cx} = n_{cx} + 1$.

To assess the quality of identification, a measure can be applied according to the formula: $ratio^{(\mathrm{idc})}= n_{cx} /(h\cdot \mathrm{b})$. If $ratio^{(\mathrm{idc})}$ we assume, that there is behavior change in a short observation interval.

III USE CASE AND RESULTS

The data used for the test from the open data sources Nodobo [13] (set CDR1) and Reality Commons (calls dataset) [14]. Thus, two sets of data were analyzed. For testing methods, software was developed in Python with Jupyter Notebook environment [15]. The main reason for this decision is the set of machine learning libraries used by researchers around the world, for example, the scikit-learn library [16]. The first data

source was used, including 63,824 entries, 5 columns and 72 unique subscribers for analysis (after purification).

We used a benchmark deviation detection model based on the analysis of the deviation of the obtained value from the previously allocated interval. In this paper, the following approach is used: all nonpositive values of the duration attribute are considered to be emissions, and positive ones are included in the interval [μ-aσ; μ + aσ], where μ is the mean and σ is the standard deviation, and the parameter a is given by the expert or determined experimentally (in the work, the value of a varied from 0.1 to 3).

Experiments were carried out on the daily data profiles. For testing, the converted sample of data was divided into the training (on which the clusters were formed) and the test: subscriber profiles that were not included in the sample for clustering. The evaluation of the quality of the method was carried out as follows: the cluster was selected, the quality of the identification of changes in the profiles of the cluster subscribers in the test sample was estimated and the quality of the identification of the changes on the subscriber profiles of other clusters of the test sample. In the experiment, the number of clusters k varied from 3 to 7. The tables show the results for clusters 3 and 4.

The illustrations to the work of the methods are presented in Figures 3-10. Figures 3 and 4 show the results in which cluster 1 is selected as the base one, i.e. Figure 3 shows that user 64 enters the group boundaries (minimum and maximum), and Figure 4 explicitly indicates that user 5 (from cluster 3) is clearly out of this range. The boundaries are specified by the dotted (minimal) and dash-dotted (maximum), and the change in the value itself is indicated by a solid one. In Figures 5, 6; 7, 8 and 9, 10 present similar results only with other values of k (the number of clusters) and m (the number of partitions), as well as by users.

TABLE I. RESULTS OBTAINED BY BENCHMARK MODEL FOR A NUMBER OF CLUSTER IS EQUAL TO 3.

m	1 1	1 2	1 3	2 1	2 2	2 3	3 1	3 2	3 3
2	0	0	0.19	0.45	0.01	0.5	0.25	0.15	0
4	0.03	0.5	0.5	0	0.02	0.23	0.03	0.17	0.07
8	0.08	0.5	0.5	0.04	0.11	0.44	0.03	0.29	0.08
16	0.25	0.57	0.9	0.5	0.08	0.5	0.5	0.23	0
24	0.17	0.5	0.21	0.5	0.29	0.26	0.5	0.5	0.27

TABLE II. RESULTS OBTAINED BY PROPOSED MODEL FOR A NUMBER OF CLUSTER IS EQUAL TO 3.

m	1 1	1 2	1 3	2 1	2 2	2 3	3 1	3 2	3 3
2	0	1	0.93	1	0	0.5	0.75	0.5	0
4	0	0.48	0.5	0.5	0	0.7	0.52	0.55	0
8	0	0.5	0.5	0.52	0	0.92	0.5	0.72	0
16	0	0.5	0.5	0.58	0	1	0.52	1	0
24	0	0.5	0.5	0.56	0	1	0.52	1	0

TABLE III. RESULTS OBTAINED BY BENCHMARK MODEL FOR A NUMBER OF CLUSTER IS EQUAL TO 4.

m	1 1	1 2	1 3	1 4	2 1	2 2	2 3	2 4
2	0.01	0.5	0.5	0.5	0	0	0.5	0
4	0.03	0.5	0.5	0.5	0.01	0.07	0.15	0.5
8	0.08	0.5	0.5	0.5	0.08	0.12	0.47	0.29
16	0.11	0.5	0.14	0.5	0.29	0	0.23	0.5
24	0.15	0.5	0.5	0.28	0.5	0.17	0.5	0.14
m	**3 1**	**3 2**	**3 3**	**3 4**	**4 1**	**4 2**	**4 3**	**4 4**
2	0.5	0.5	0	0.34	0.03	0.5	0.38	0
4	0	0.23	0	0.42	0.5	0.75	0.53	0.08
8	0.06	0.29	0.03	0.5	0.02	0.26	0.12	0.08
16	0.5	0.5	0.1	0.5	0.89	0.95	0.58	0.2
24	0.35	0.5	0.27	0.22	0.5	0.5	0.5	0.22

TABLE IV. RESULTS OBTAINED BY PROPOSED MODEL FOR A NUMBER OF CLUSTER IS EQUAL TO 4.

m	1 1	1 2	1 3	1 4	2 1	2 2	2 3	2 4
2	0	0.5	0.5	0.5	0.56	0	0.83	0.5
4	0	0.5	0.47	0.5	0.55	0	0.82	0.86
8	0	0.5	0.5	0.5	0.52	0	0.88	0.5
16	0	0.5	0.5	0.5	0.51	0	1	1
24	0	0.53	1	0.76	0.5	0	0.5	0.5
m	**3 1**	**3 2**	**3 3**	**3 4**	**4 1**	**4 2**	**4 3**	**4 4**
2	0.5	0.54	0	0.81	0.5	0.5	0.64	0
4	0.5	0.92	0	0.68	0.5	0.92	0.53	0
8	0.5	0.7	0	1	0.62	0.6	1	0
16	0.56	1	0	0.82	0	0.5	0.5	0
24	1	0.5	0	1	0.5	0	0.5	0

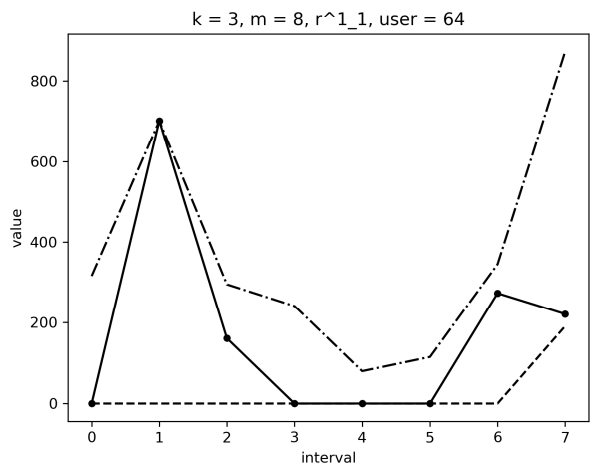


Fig. 2. Calls for the 64th user from the 1st cluster with the base cluster 1 (partition interval 8, number of clusters 3).

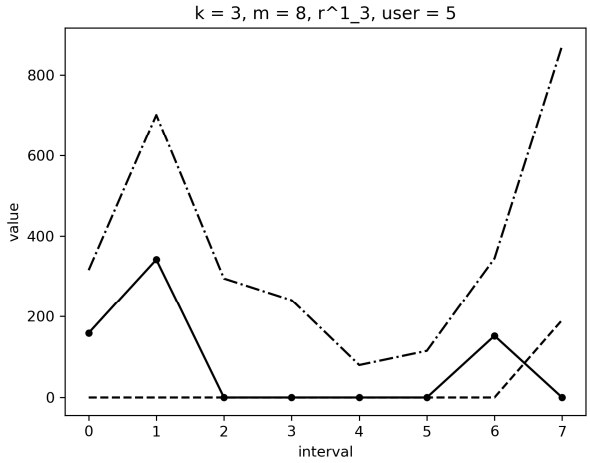


Fig. 3. Calls for the 5th user from the 3rd cluster with the base cluster 1 (partition interval 8, number of clusters 3).

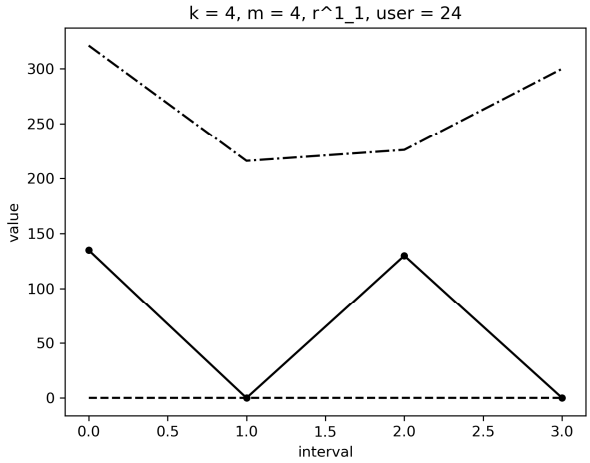


Fig. 4. Calls for the 24th user from the 1st cluster with the base cluster 1 (partition interval 4, number of clusters 4).

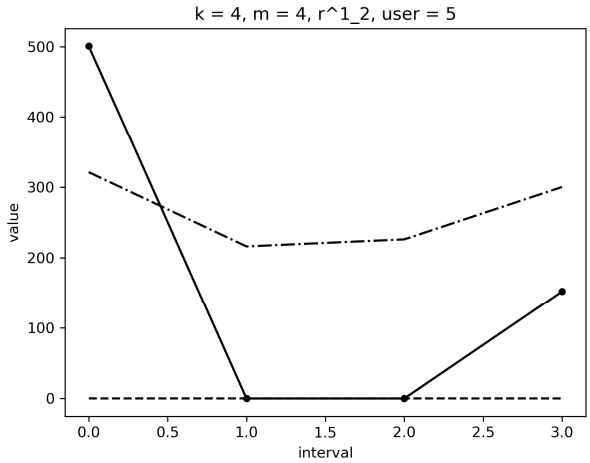


Fig. 5. Calls for the 5th user from the 2nd cluster with the base cluster 1 (partition interval 4, number of clusters 4).

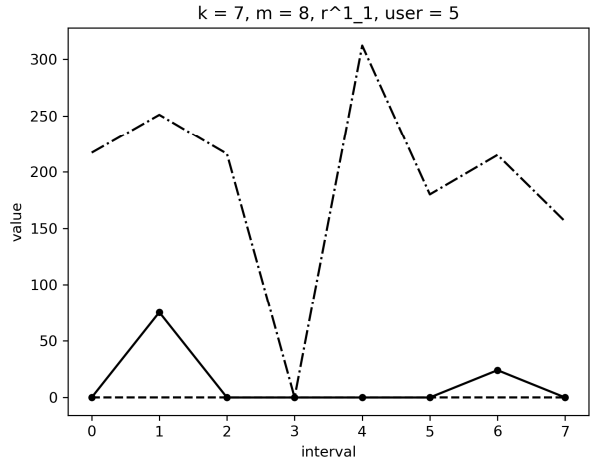


Fig. 6. Calls for the 5th user from the 1st cluster with the base cluster 1 (partition interval 8, number of clusters 7).

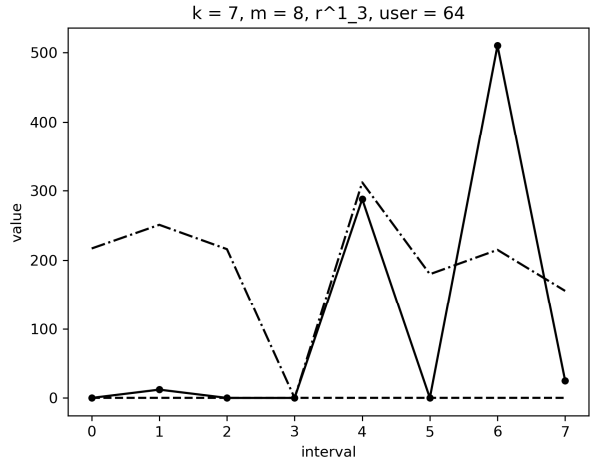


Fig. 7. Calls for the 64th user from the 3rd cluster with the base cluster 1 (partition interval 8, number of clusters 7).

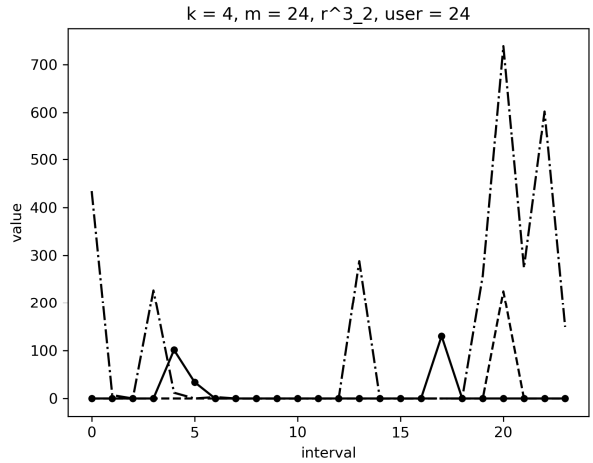


Fig. 8. Calls for the 24th user from the 2nd cluster with the base cluster 3 (partition interval 24, number of clusters 4).

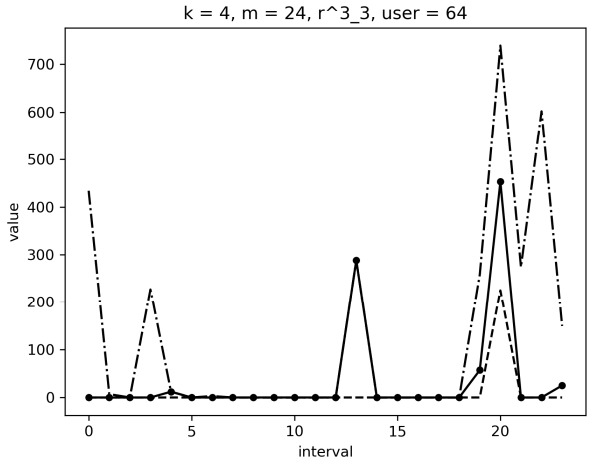


Fig. 9. Calls for the 64th user from the 3rd cluster with the base cluster 3 (partition interval 24, number of clusters 4).

IV CONCLUSION

We proposed a new method for detecting changes in the behavior of subscribers which allows to reveal changes in behavior without preliminary sampling of data, which differs in that it is built on formal methods of clustering behavior.

As a result of the clustering test, high identification scores were obtained on the test sample in comparison with benchmark model.

ACKNOWLEDGMENT

The reported study was partially supported by RFBR research projects 16-37-60066_mol_a_dk, Project MD-6964.2016.

REFERENCES

[1] S. Ustugova, D. Parygin, N. Sadovnikova, A. Finogeev, A. Kizim, “Monitoring of social reactions to support decision making on issues of urban territory management”, Procedia Computer Science, Proc. of the 5th International Young Scientist Conference on Computational Science, YSC 2016, Krakow, Poland, 26–28 October 2016, Elsevier, 2016., vol. 101, pp. 243–252.

[2] A. Golubev, I. Chechetkin, D. Parygin, A. Sokolov, M. Shcherbakov, “Geospatial data generation and preprocessing tools for urban computing system development”, Procedia Computer Science, Proc. of the 5th International Young Scientist Conference on Computational Science, YSC 2016, Krakow, Poland, 26–28 October 2016, Elsevier, 2016., vol. 101, pp. 217–226.

[3] S. F. Hinde, Call Record Analysis, Making Life Easier - Network Design and Management Tools (Digest No: 1996/217), IEE Colloquium on, 8/1 8/4, (1996).

[4] DEFINITY Enterprise Communication Server (ECS): Technical Articles and Technical Tips, available online: http://esearch.avaya.com/

[5] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official Journal L 281/31 of 23.11.95, pp 31-50, available online at: http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX: 31995L0046:EN:NOT (1995).

[6] Directive 2002/58/EC. Directive on privacy and electronic communications. Official Journal L 201, 31.7.2002, 3747, available online at: http://eurlex.europa.eu/LexUriServ/LexUriServ.do? uri=CELEX: 32002L0058:EN:NOT (2002)

[7] Hong, R. Clustering analysis of telecommunication customers / R. Hong, Z. Yan , W. Ye–rong // The Journal of China Universities of Posts and Telecommunications, 2009, vol.16, I. 2, pp. 114–116.

[8] Ye, L. Customer Segmentation for Telecom with the k–means Clustering Method / L. Ye, C. Qiuru, X. Haixu, L. Yijun и Z. Guangping // Information Technology Journal, 2013, vol. 12, no 3, pp. 409–413.

[9] Bascacov, A. Using Data Mining for Mobile Communication Clustering and Characterization / A. Bascacov, C. Cernazanu и M. Marcu // Conference: Applied Computational Intelligence and Informatics (SACI), 2013 IEEE 8th International Symposium on. – DOI: 10.1109/SACI.2013.6609004

[10] Guo, Z. Telecommunications User Behaviors Analysis Based on Fuzzy C–Means Clustering/ Z. Guo and F. Wang // FGIT 2010, LNCS 6485, pp. 585–591.

[11] Jakub Konecny. Mini-Batch Semi-Stochastic Gradient Descent in the Proximal Setting/ Jakub Konecny, Jie Liu, Peter Richtarik, MartinTakac. https://arxiv.org/pdf/1504.04407.pdf (access date: 20.10.2017).

[12] Kaufman L., and Rousseeuw P.J., Finding groups in data: an introduction to cluster analysis, New York: John Wiley & Sons, Inc. (1990)

[13] Data Source (DCR – Detailed Call Record). http://realitycommons.media.mit.edu/ (access date: 20.02.2015).

[14] Data Source (DCR – Detailed Call Record). http://nodobo.com/release.html (access date: 20.02.2015).

[15] Interactive shell for Python language. http://jupyter.org. (access date: 25.10.2017)

[16] Machine Learning library for Python language. http://scikit–learn.org. (access date: 25.10.2017)