УДК 004.891.2
DOI: 10.26102/2310-6018/2021.35.4.007

Сравнение методов кластеризации DBSCAN и модифицированного WrapDBSCAN для поиска аномальных перемещений пользователей в мобильной UBA системе

П.А. Савенков

Тульский государственный университет,
Тула, Российская Федерация

Резюме. Одной из актуальных проблем в имеющихся системах анализа поведения является извлечение признаков аномальной активности деятельности пользователей из больших массивов входных данных. Проблема, решаемая в данном исследовании, основана на невозможности поиска аномальной активности пользователей по их перемещениям в связи с высокой вариативностью входных данных. Целью исследования является разработка модифицированного метода плотностной кластеризации для применения в мобильной системе поведенческого анализа с использованием методов и алгоритмов машинного обучения для нахождения отклонений в поведении пользователей по их перемещениям. В статье осуществляется сравнительный анализ методов плотностной кластеризации, применяемых в разрабатываемом программном комплексе поиска аномалий в поведенческих биометрических характеристиках пользователей системы. Осуществляется сглаживающая интерполяция входных данных. Описывается результат поиска аномалий модифицированным методом пространственной кластеризации с различными входными параметрами и осуществляется сравнение результатов с базовым методом. Благодаря использованию разработанного метода пространственной кластеризации достигнуто повышение качества анализа аномальной активности в деятельности пользователей по их перемещениям. Нахождение отклонений в собранных данных обеспечит своевременное реагирование администратора системы на отклонения от поведенческого профиля пользователя.

Ключевые слова: машинное обучение, большие данные, наука о данных, программное обеспечение, информационная система, неструктурированные данные, поведенческий анализ, поведенческая биометрия, биометрические характеристики, искусственный интеллект

Для цитирования: Савенков П.А. Сравнение методов кластеризации DBSCAN и модифицированного WrapDBSCAN для поиска аномальных перемещений пользователей в мобильной UBA системе. *Моделирование, оптимизация и информационные технологии.* 2021;9(4). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=977 DOI: 10.26102/2310-6018/2021.35.4.007

Comparison of clustering methods DBSCAN and modified WrapDBSCAN to find abnormal user movements in the mobile UBA system

P.A. Savenkov

Tula State University, Tula, Russian Federation

Abstract: One of the urgent problems in the existing systems of behavior analysis is the extraction of signs of anomalous activity of user activity from large arrays of input data.The problem solved in this study is based on the impossibility of searching for anomalous activity of users by their movements, due to the high variability of the input data. The aim of the study is to develop a modified density clustering

method for application in a mobile system of behavioral analysis using machine learning methods and algorithms to find deviations in user behavior based on their movements. This article provides a comparative analysis of the density clustering methods used in the developed software package for searching for anomalies in the behavioral biometric characteristics of system users. Smoothing interpolation of the input data is performed. The results of searching for anomalies by the modified method of spatial clustering with different input parameters are described and the results are compared with the basic method. Thanks to the use of the developed method of spatial clustering, an increase in the quality of the analysis of anomalous activity in the activities of users on their movements has been achieved. Finding deviations in the collected data will ensure a timely response of the system administrator to deviations from the user's behavioral profile.

Keywords: machine learning, big data, data science, software, information system, unstructured data, behavioral analysis, behavioral biometrics, biometric characteristics, artificial intelligence

For citation: Savenkov P.A. Comparison of clustering methods DBSCAN and modified WrapDBSCAN to find abnormal user movements in the mobile UBA system. *Modeling, Optimization and Information Technology*. 2021;9(4). Available from: https://moitvivt.ru/ru/journal/pdf?id=977 DOI: 10.26102/2310-6018/2021.35.4.007(In Russ).

Введение

Целью исследования является применение и сравнение методов плотностной кластеризации в разрабатываемой системе поведенческого анализа с использованием методов и алгоритмов машинного обучения при детектировании отклонений от нормального поведения пользователей [1]. В последние годы особо активно проявляется интерес к анализу отклонений в деятельности пользователей для сохранения целостности данных информационных систем. Мировым сообществом отмечается возрастание количества внутренних угроз, наиболее часто утечки информации происходят по вине собственных сотрудников, имеющих доступ к информационной системе. Группу потенциально опасных для организации пользователей достаточно трудно идентифицировать [2]. Актуальные исследования показывают, что проходит от нескольких недель до нескольких месяцев для подготовки и хищения информации. Возможную потерю данных необходимо обнаруживать еще на ранних стадиях, до непосредственной утечки информации за контур предприятия.

По прошествии определенного промежутка времени пользователь начинает совершать не характерные для его предыдущей (нормальной) активности действия. Данная аномальная активность может продолжаться от двух недель до нескольких месяцев. Аномальное поведение может свидетельствовать о том, что пользователь не является тем, от имени кого он авторизовался [3]. Так же аномальное поведение может свидетельствовать тому, что пользователь перестал выполнять свои целевые обязанности должным образом и расходует рабочее время на решение иных задач. За последние несколько лет активное развитие получило направление анализа поведения пользователей для обнаружения отклонений от эталонного профиля.

Материалы и методы

В настоящее время сформировался самостоятельный класс систем, в основе которых лежат методы и алгоритмы машинного обучения. Данные методы и алгоритмы применяются для выявления отклонений от нормального поведения пользователей. Консалтинговая компания Gartner обозначает данные системы как UBA (англ. User Behavior Analytics – анализ поведения пользователей) [4]. UBA системы осуществляют мониторинг множества действий в деятельности пользователя и оповещают администратора системы в случае возникновения аномальной активности. Данные

системы предоставляют аналитические данные с описанием найденной аномалии для конкретного пользователя.

При формировании и дальнейшем анализе поведенческого профиля пользователя происходит обработка больших массивов разнородных данных, однако достаточно сложно предпринять какое-либо решение на их основе, в связи с тем, что обрабатываемый объем данных достаточно велик. Для обнаружения отклонений в поведении пользователя одним из анализируемых факторов является перемещение пользователя.

В реализуемой мобильной UBA системе используются методы машинного обучения. Данный подход позволяет сократить количество выходных параметров системы [5]. Сбор данных осуществляется при помощи разработанного мобильного приложения, устанавливаемого на мобильное устройство пользователя с операционной системой «Android» с его согласия.

Для нахождения отклонений от эталонного профиля пользователя в таких данных, как история перемещений сотрудника (GPS), применяется метод «DBSCAN» [6] и разработанный на его основе метод «WrapDBSCAN». «WrapDBSCAN» отличается от базового метода «DBSCAN» присутствием возможности поиска наиболее подходящего (оптимального) радиуса вхождения точек. Вводимый параметр «радиус» отсутствует в реализации данного метода. Добавляется новый параметр, определяющий количество итераций разбиения. Параметр показывает, насколько детализировано будет рассматриваться исходный набор данных. Чем больше значение параметра, тем более детализировано рассматривается набор данных.

В связи с особенностями работы (энергосбережение, проблемы с сетью, разряд батареи) различных мобильных устройств при сборе перемещений пользователей для корректной работы метода требуется интерполировать данные на анализируемом временном промежутке. Для создания массива, содержащего перемещения, время между которыми одинаково (создание результирующего массива данных, распределенного во временной области с указанным константным временным интервалом), используется многомерная кусочно-линейная интерполяция [7].

$$t = au + b \tag{1}$$

$$a = tg(\alpha) = \frac{\Delta t}{\Delta \mathrm{u}} = \frac{t_i - t_{i-1}}{u_i - u_{i-1}} \tag{2}$$

$$b = t_i - au_i, \tag{3}$$

где:

t – искомое, промежуточное значение;

a – коэффициент уравнения;

Δt, Δu – разница между известными граничными значениями параметра.

Интерполируются такие значения, как:

- Altitude (высота);
- Latitude (широта);
- Longitude (долгота);
- Speed (скорость).

После интерполяции данных GPS перемещений пользователей в базе данных формируется сглаженная выборка с значениями перемещений через каждые 2 минуты. Результат работы алгоритма представлен на Рисунке 1.

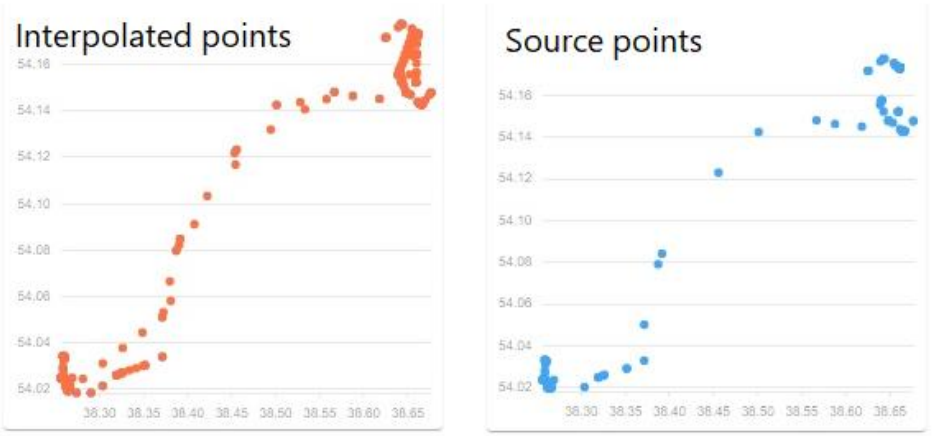

Рисунок 1– Результат интерполяции GPS координат пользователя
Figure 1– The result of the interpolation of the user's GPS coordinates

Рассмотрим экспериментальные данные, полученные при применении метода DBSCAN. Варьируя параметром радиуса сканирования и количества точек, попадающих в радиус, проведем анализ датасета 1 (Рисунок 2).

Анализ проводится со следующими вариациями параметров (eps – радиус сканирования, minPts – количество точек, попадающих в радиус):

- eps = 0.0001, minPts = 3;
- eps = 0.0015, minPts = 3;
- eps = 0.002, minPts = 3.

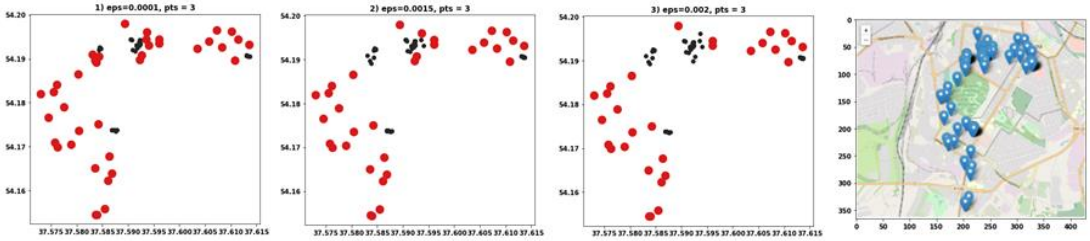

Рисунок 2 – Анализ датасета 1
Figure 2 – Dataset analysis 1

Из рисунка видно, что количество аномалий становится меньше при увеличении параметра eps.

Для второго набора данных, варьируя параметром радиуса сканирования (eps) и количества точек, попадающих в радиус (minPts), проведем анализ датасета 2 (Рисунок 3).

Анализ проводится со следующими вариациями параметров:

- eps = 0.0001, minPts = 3;
- eps = 0.0015, minPts = 3;
- eps = 0.002, minPts = 3.

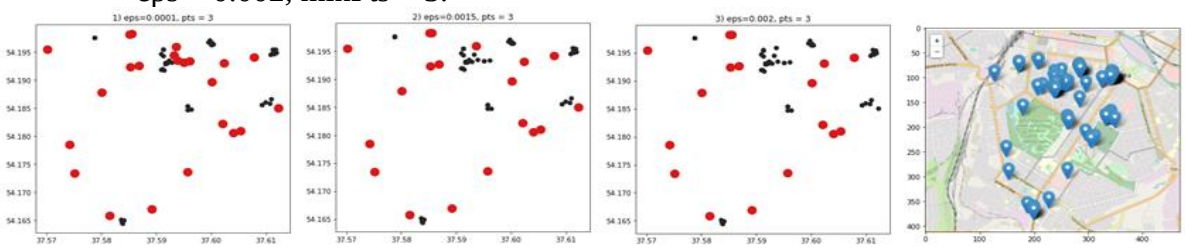

Рисунок 3 – Анализ датасета 2
Figure 3 – Dataset analysis 2

Из рисунка видно, что с увеличением радиуса поиска количество аномалий уменьшается, что логично. В силу того, что определить значение параметра minPts и результаты анализа при заданном конкретном его значении относительно возможно, можно выбирать его значение как жесткое (например, 50) или мягкое (например, 3). Однако данные о местоположении в контексте анализа алгоритмом DBSCAN крайне чувствительны к параметру eps. Его значение сильно влияет на результирующее количество найденных аномалий. В силу непредсказуемости собранных данных о местоположении значение параметра eps для каждого набора данных будет разным. Вручную подбирать этот параметр для каждого набора данных трудоемко.

Рассмотрим экспериментальные данные, полученные при применении модифицированного метода WrapDBSCAN. Варьируя параметрами количества итераций (iter) и минимального количества точек (min_pts), проведем анализ датасета 1.

- min_pts = 3, вариация iter=1..3 (см. Рис. 4);
- min_pts = 5, вариация iter=1..3 (см. Рис. 5);
- min_pts = 50, вариация iter=1..3 (см. Рис. 6).

На всех рисунках красным цветом выделены аномалии, черным – хорошие данные. Представлены результаты анализа с iter, равным 1, 2, 3.

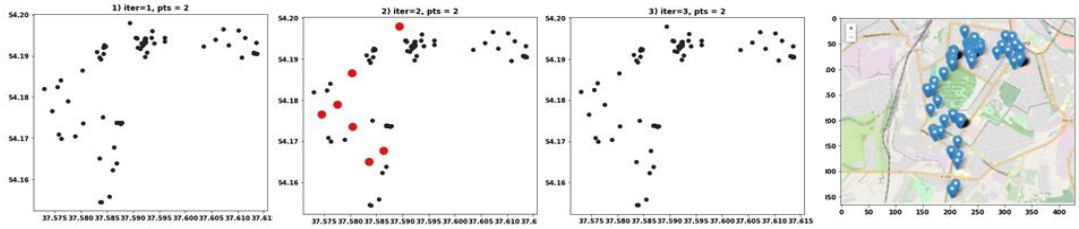

Рисунок 4 – Анализ датасета 1 (набор параметров 1)
Figure 4 – Dataset Analysis 1 (Option Set 1)

Из рисунка 20 видно, что при изменении жесткости детектирования аномалий за счет изменения iter они находятся не так грубо, как в предыдущих методах.

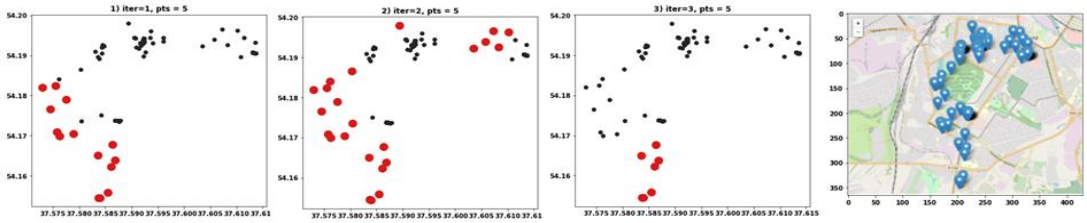

Рисунок 5 – Анализ датасета 1 (набор параметров 2)
Figure 5 – Dataset Analysis 1 (Option Set 2)

Из Рисунка 5 видно, что при изменении min_pts в большую сторону количество аномалий увеличивается по сравнению с результатами анализа из Рисунка 4.

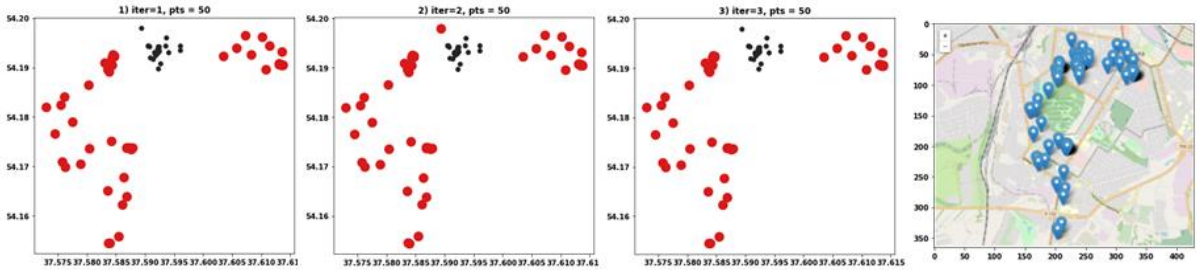

Рисунок 6 – Анализ датасета 1 (набор параметров 3)
Figure 6 – Dataset Analysis 1 (Option Set 3)

Из рисунка 6 видно, что при увеличении значения параметра min_pts аномалий становится еще больше.

Варьируя параметрами количества итерации и минимального количества точек, проведем анализ датасета 2.

- min_pts = 3, вариация iter=1..3 (см. Рис. 7);
- min_pts = 5, вариация iter=1..3 (см. Рис. 8);
- min_pts = 50, вариация iter=1..3 (см. Рис. 9).

На всех рисунках красным цветом выделены аномалии, черным – хорошие данные. Представлены результаты анализа с iter, равным 1, 2, 3. Во всех результатах анализа ситуация с аномалиями аналогична результатам анализа для датасета 1.

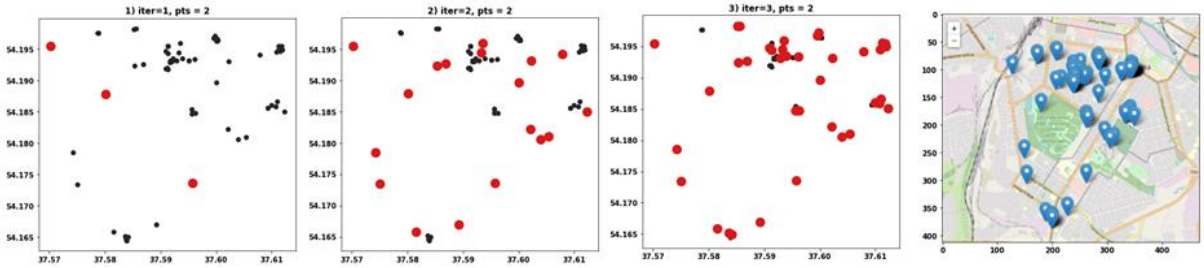

Рисунок 7 – Анализ датасета 2 (набор параметров 1)
Figure 7 – Dataset Analysis 2 (Option Set 1)

Манипулируя параметром количества итераций, можно регулировать степень разбиения на локальные окрестности.

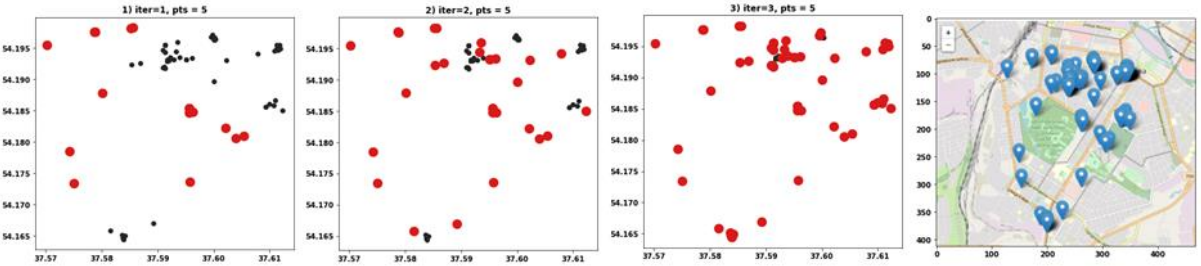

Рисунок 8 – Анализ датасета 2 (набор параметров 2)
Figure 8 – Dataset Analysis 2 (Option Set 2)

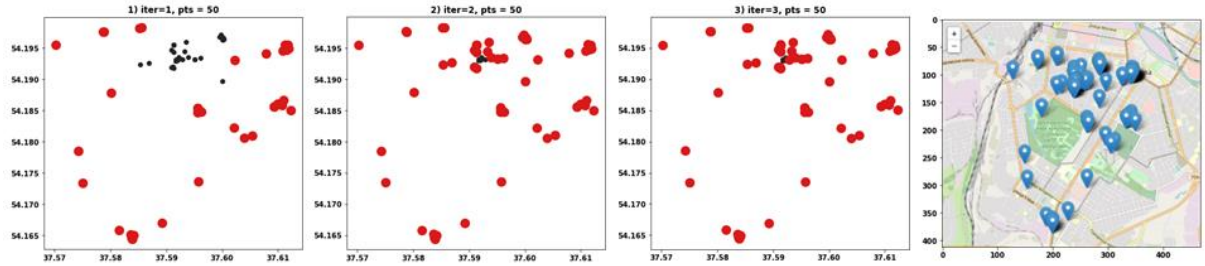

Рисунок 9 – Анализ датасета 2 (набор параметров 3)
Figure 9 – Dataset Analysis 2 (Option Set 3)

Из Рисунков 7, 8, 9 видно, что, варьируя всеми параметрами метода WRAPDBSCAN, можно регулировать чувствительность детектирования.

В зависимости от значения параметра iter мы более или менее локально оцениваем среднее расстояние между элементами датасета и применяем найденное общее среднее расстояние к алгоритму DBSCAN. Однако при большом количестве и плотности точек на маленькой локальной области датасета (что крайне вероятно в датасетах местоположения) среднее расстояние будет стремиться к минимальному значению, перекрывая значимые составляющие других средних локальных значений кластеров датасета.

Результаты

В ходе проведения исследования был разработан метод WrapDBSCAN, решающий проблему ручного поиска радиуса вхождения точек, в отличие от базового метода DBSCAN. Проблема решается при помощи автоматического подбора входных параметров для метода DBSCAN.

Проведены эксперименты по интерполяции GPS данных, и их анализа при помощи метода DBSCAN и модифицированного метода WrapDBSCAN. Благодаря интерполяции данных удалось произвести сглаживание собранных данных и, в конечном итоге, нормализовать их для корректной работы методов DBSCAN и WrapDBScan. Впервые собрана база перемещений пользователей.

Связка метода WrapDBSCAN с методом нормализации и методом временного детектирования может сформировать более точный результат детектирования аномалий местоположения в тех или иных сферах использования.

Применение методов машинного обучения позволило сократить количество результирующих данных, тем самым повысило их информативность [9].

Заключение

Методы и алгоритмы машинного обучения (методы пространственной кластеризации) применимы для анализа аномальной активности, однако требуют корректной настройки входных параметров. Использование методов кластеризации в задаче поиска аномалий обусловлено невозможностью использования методов классификации, так как отсутствует возможность сформировать обучающую выборку.

Актуальность UBA систем и машинного обучения в данный момент высока и будет расти по мере развития информационных систем и технологий [10].

СПИСОК ИСТОЧНИКОВ

1. Dhiman C., Vishwakarma D. K. A review of state-of-the-art techniques for abnormal human activity recognition. *Engineering Applications of Artificial Intelligence*. 2019;(77):21-45.
2. Shaw E. D. et al. Behavioral risk indicators of malicious insider theft of intellectual property: Misreading the writing on the wall. *White Paper, Symantec, Mountain View*. 2011.
3. Carmagnola F., Cena F. User identification for cross-system personalisation . *Information Sciences*. 2009;(179):16-32.
4. Shashanka M., Shen M. Y., Wang J. User and entity behavior analytics for enterprise security. *2016 IEEE International Conference on Big Data (Big Data). IEEE*. 2016:1867-1874.
5. Ivutin A. N., Savenkov P. A., Veselova A. V. Neural network for analysis of additional authentication behavioral biometrie characteristics *.2018 7th Mediterranean Conference on Embedded Computing (MECO*). – IEEE. 2018(8):1-3.
6. Mon T. L. L. Analysis of Trajectory Cleaning Based on DBSCAN and CB-SMOT Clustering Algorithms. *International Journal of Advanced Research in Technology and Innovation*. 2020(2):35-41.
7. Savenkov P. A., Ivutin A. N. Methods of Machine Learning in System Abnormal Behavior Detection *.International Conference on Swarm Intelligence. – Springer, Cham.* 2020:495-505.
8. Huang F. et al. Research on the parallelization of the DBSCAN clustering algorithm for spatial data mining based on the Spark platform. *Remote Sensing*. 2017(9):1301.
9. Akutota T., Choudhury S. Big data security challenges: An overview and application of user behavior analytics. *Int. Res. J. Eng. Technol*. 2017(4):1544-1548.
10. Touma M. et al. Framework for behavioral analytics in anomaly identification *.Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR VIII. – International Society for Optics and Photonics.* 2017(10190):101900H.

REFERENCES

1. Dhiman C., Vishwakarma D. K. A review of state-of-the-art techniques for abnormal human activity recognition. *Engineering Applications of Artificial Intelligence*. 2019;(77):21-45.
2. Shaw E. D. et al. Behavioral risk indicators of malicious insider theft of intellectual property: Misreading the writing on the wall. *White Paper, Symantec, Mountain View*. 2011.
3. Carmagnola F., Cena F. User identification for cross-system personalisation . *Information Sciences*. 2009;(179):16-32.
4. Shashanka M., Shen M. Y., Wang J. User and entity behavior analytics for enterprise security *.2016 IEEE International Conference on Big Data (Big Data).IEEE*. 2016:1867-1874.
5. Ivutin A. N., Savenkov P. A., Veselova A. V. Neural network for analysis of additional authentication behavioral biometrie characteristics *.2018 7th Mediterranean Conference on Embedded Computing (MECO*). IEEE. 2018(8):1-3.
6. Mon T. L. L. Analysis of Trajectory Cleaning Based on DBSCAN and CB-SMOT Clustering Algorithms. *International Journal of Advanced Research in Technology and Innovation*. 2020(2):35-41.
7. Savenkov P. A., Ivutin A. N. Methods of Machine Learning in System Abnormal Behavior Detection *.International Conference on Swarm Intelligence.Springer, Cham.* 2020:495-505.
8. Huang F. et al. Research on the parallelization of the DBSCAN clustering algorithm for spatial data mining based on the Spark platform. *Remote Sensing*. 2017(9):1301.
9. Akutota T., Choudhury S. Big data security challenges: An overview and application of user behavior analytics. *Int. Res. J. Eng. Technol*. 2017(4):1544-1548.

10. Touma M. et al. Framework for behavioral analytics in anomaly identification .*Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR VIII. International Society for Optics and Photonics.* 2017(10190):101900H.

ИНОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Савенков Павел Анатольевич, аспирант, кафедра вычислительной техники, ФГБОУ ВО "Тульский государственный университет" Институт прикладной математики и компьютерных наук, Тула, Российская Федерация.
e-mail: pavel@savenkov.net

Pavel A. Savenkov, Postgraduate Student, Department of Computer Engineering, Tula State University Institute of Applied Mathematics and Computer Science, Tula, Russian Federation.

Статья поступила в редакцию 30.06.2021; одобрена после рецензирования 23.09.2021; принята к публикации 28.10.2021.

The article was submitted 30.06.2021; approved after reviewing 23.09.2021; accepted for publication 28.10.2021.