Моделирование и анализ данных
2022. Том 12. № 1. С. 5–15
DOI: https://doi.org/10.17759/mda.2022120101
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2022. Vol. 12, no. 1, pp. 5–15
DOI: https://doi.org/10.17759/mda.2022120101
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

АНАЛИЗ ДАННЫХ

УДК 004.855

Кластеризация водителей по степени опасности вождения с использованием алгоритмов машинного обучения

Баданина Н.Д.*
Московский авиационный институт (МАИ)
г. Москва, Российская Федерация
ИПМ им. М.В.Келдыша РАН
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-5301-1526
e-mail: natashabadanina99@gmail.com

Судаков В.А.**
Московский авиационный институт (МАИ)
г. Москва, Российская Федерация
ИПМ им. М.В.Келдыша РАН
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-1658-1941
e-mail: sudakov@ws-dss.com

Для цитаты:
Баданина Н.Д., Судаков В.А. Кластеризация водителей по степени опасности вождения с использованием алгоритмов машинного обучения // Моделирование и анализ данных. 2022. Том 12. № 1. С. 5–15. DOI: https://doi.org/10.17759/mda.2022120101

***Баданина Наталья Дмитриевна,** студент магистратуры, Московский авиационный институт (национальный исследовательский университет) (МАИ (НИУ)), программист, Федеральное государственное учреждение «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук» (ИПМ им. М.В.Келдыша РАН), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-5301-1526, e-mail: natashabadanina99@gmail.com

****Судаков Владимир Анатольевич,** доктор технических наук, профессор, Московский авиационный институт (национальный исследовательский университет) (МАИ (НИУ)), ведущий научный сотрудник, Федеральное государственное учреждение «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук» (ИПМ им. М.В.Келдыша РАН), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-1658-1941, e-mail: sudakov@ws-dss.com

В работе проведено исследование по определению опасности вождения транспортного средства посредством анализа сигналов, полученных во время поездки. Был применен функционал ряда современных моделей кластеризации для разбиения множества водителей на классы, соответствующие степени опасности вождения. Разработан новый подход агрегации данных с целью кластеризации с использованием гистограмм распределения сигналов. Полученные результаты могут быть использованы в промышленных системах мониторинга качества поведения водителей в ретроспективе.

Ключевые слова: кластеризация, гистограммы распределения, опасное вождение, системы мониторинга, машинное обучение.

1. ВВЕДЕНИЕ

Системы мониторинга безопасности вождения транспортных средств (ТС) призваны снижать аварийность на дорогах; контролировать общее соблюдение правил дорожного движения. Частично разработки нацелены на заблаговременное предотвращение аварийных ситуаций. Такие системы безусловно важны, по причине того, что от них в прямом смысле может зависеть безопасность на дороге, жизни водителей, пассажиров и пешеходов. При должном контроле поведения водителя, существует возможность применения санкционных мер, с целью улучшить его поведение за рулем, что в перспективе может снизить риск дорожно-транспортного происшествия.

Методы анализа данных широко применяются для обработки больших данных высокой размерности. Кластеризация – разбиение множества объектов на группы (кластеры), основываясь на свойствах этих объектов. Кластер представляет собой группу объектов, имеющих общие признаки. Целью алгоритмов кластеризации является создание классов, которые максимально связаны внутри себя, но различны друг от друга [1]. Таким образом, в системах мониторинга опасности вождения можно применить алгоритмы кластеризации, с целью решения задачи группировки водителей на некоторое количество классов, соответствующее различным типам вождения.

2. РЕЛЕВАНТНЫЕ ИССЛЕДОВАНИЯ И РАЗРАБОТКИ

На рынке присутствуют как государственные реализации системы мониторинга, так и частные. К государственным относится система камер фото- и видео-фиксации ГИБДД. В их основе лежит одна из технологий искусственного интеллекта – компьютерное зрение. Эта система имеет ряд преимуществ, однако, не имеет возможности сделать вывод о безопасности вождения в любом моменте времени – необходимо обязательное наличие камер или иных средств фиксации нарушения. У органов регулировки дорожно-транспортной ситуации нет возможности обеспечить стопроцентное покрытие всех автомагистралей, дорог специализированным оборудованием, и нет возможности осуществлять мониторинг ТС на протяжении всего пути. Во многом операционная система Яндекс.Навигатор и подобные «помогают» водителям избежать нарушений, предупреждая о наличии средств контроля заранее. Это ведет

к недостоверной статистике о поведении водителя. Третьим недостатком является зависимость от технологии компьютерного зрения, которая не способна работать при ненадлежащем качестве изображения, при поломке фото- или видеорегистраторов.

Альтернативные способы контроля качества вождения представлены частными компаниями. Яндекс.Про считывает данные акселерометра устройства, на которое он установлен, – смартфона или планшета. Если водитель совершает агрессивные манёвры – резкий старт и торможение, многократную смену полос, несоблюдение дистанции, – акселерометр фиксирует перегрузку и считает нарушением при систематичности [3]. Недостатком является факт того, что по каждому ТС показатели собираются множество раз за все время поездки, что усложняет анализ и может вести к ложным выводам из-за ошибки оборудования или из-за факта разовых нарушений.

В исследовательской области есть ряд работ по определению опасного вождения – кластеризация водителей по стилям торможения [2] с использованием нейронной сети Кохонена; определение поведения водителей на основе мониторинга движения в реальном времени [5]; детекция несосредоточенности водителя с использованием изображения машины внутри и снаружи [6].

Таким образом, на рынке мониторинга безопасности вождения транспортных средств присутствует возможность внедрения новых технологий, использующих машинное обучение. В данной работе описана методика построения моделей кластеризации водителей для задачи обнаружения паттернов опасного вождения.

3. ГИСТОГРАММЫ РАСПРЕДЕЛЕНИЯ КАК СПОСОБ СОСТАВЛЕНИЯ ОБУЧАЮЩЕГО МНОЖЕСТВА

В качестве данных в работе был использован датасет, собранный из информации о сигналах с ТС. К сигналам относятся значения скорости; значения ускорений по трем осям x, y, z.

Дано множество водителей $D = \{d_1, d_2, \ldots, d_n\}$. Для каждого $d_i \in D$, $i = 1, \ldots, n$ дан вектор скоростей $V = (v_1, v_2, \ldots, v_m, \ldots)$, где каждое значение соответствует скорости ТС. Аналогично заданы вектора ускорений $A_x = (a_{x_1}, a_{x_2}, \ldots, a_{x_m}, \ldots)$, $A_y = (a_{y_1}, a_{y_2}, \ldots, a_{y_m}, \ldots)$, $A_z = (a_{z_1}, a_{z_2}, \ldots, a_{z_m}, \ldots)$ по осям x, y, z соответственно. Требуется разделить множество D на K непересекающихся кластеров $C = \{c_1, c_2, \ldots, c_k\}$, $k = 1, \ldots, K$.

Данные о сигналах поступали с некоторой периодичностью на протяжении всей поездки на ТС, даже при условии того, что фактическая скорость могла быть нулевой, то есть факт движения отсутствовал. На их основе были обучены и протестированы модели кластеризации, реализованные с помощью методов машинного обучения без учителя, то есть в данных отсутствовала разметка о принадлежности того или иного водителя к некоему классу, отвечающему за степень опасности вождения.

Данные по трекам были агрегированы. Для каждого водителя была собрана вся доступная информация по сигналам, записанным во время всех поездок. За каждую

поездку могло поступить множество сигналов с разным временным промежутком. Проблема обучения моделей на таких данных состоит в том, что для каждого водителя количество собранных сигналов велико и для разных водителей их может быть разное количество, и, следовательно, вектор признаков будет иметь различную длину. С векторами различной длины нельзя составить обучающее множество. Вариант дополнения каждого вектора до максимальной длины нецелесообразен, так как может вести к существенному увеличению времени обучения модели и ложным выводам.

Решением описанной проблемы является составление частотных гистограмм распределения значений сигналов по интервалам. Гистограммы распределения указывают насколько часто встречаются те или иные значения. Рассмотрим алгоритм построения для задачи кластеризации водителей.

Выделяются сигналы одной природы. В рассматриваемой задаче описано множество сигналов $S = \{V,\ A_x,\ A_y,\ A_z\}$. Для каждого вектора из множества S выделяются интервалы значений. Не допускается перекрытие промежутков и наличие пропущенных значений.

- Интервалы задаются экспертом;
- Весь диапазон значений разбивается на равное количество частей с некоторым шагом;
- Оптимальное количество интервалом определяется математически, исходя из мощности выборки. Применяется формула Стерджесса,

$$m = 1 + 3,322 * \lg(n) \tag{1}$$

где n – количество наблюдений. После подсчета коэффициента m весь диапазон значений от минимального до максимального разбивается на равные части. Ширина интервала определяется по формуле

$$w = \frac{X_{max} - X_{min}}{m} \tag{2}$$

Тогда первый интервал начинается в X_{min}, а последний (с номером m) заканчивается в X_{max}. Интервалы составляются следующим образом:

$$(X_{min},\ X_{min} + w), \ldots, (X_{max} - w,\ X_{\max}).$$

После определения интервалов осуществляется подсчет частоты попадания значений сигналов в соответствующие интервалы. Частота считается явно: сколько из всех значений находятся внутри интервала. Пусть a_i и b_i – левая и правая граница интервала i соответственно, тогда считается количество значений x_j, для которых выполняются неравенства $x_j \geq a_i$ и $x_j < b_i$.

Таким образом, можно подсчитать частоту попадания значений сигналов в определенный интервал и составить обучающее множество. После выполнения вышеописанного алгоритма был составлен датасет следующего вида: колонки соответствуют интервалу сигнала из S, строки – водителям из D, а цифры – количеству вхождений значений сигнала в интервал.

4441	634	442	209	251	231	310	326	263	42	...	21
21786	2642	4962	3853	3861	4174	3764	2443	1024	181	...	0
4565	1196	923	777	1362	1391	1908	1814	2254	466	...	0
4846	731	681	381	377	503	477	503	168	61	...	0
78748	17543	10267	3995	3309	3303	2567	919	86	1	...	0

Рис. 1. Пример полученного датасета

Интервалы были заданы одинаковой длины: для скорости с шагом 10, а для ускорений с шагом 100. Каждому водителю соответствует единственная строка, в которой указано количество попадания значений сигнала в заданные промежутки значений. Такой подход позволяет:

- учесть данные по всем поездкам, совершенным водителем ТС;
- создать вектор определенной неизменной длины для каждого водителя;
- объединять интервалы при необходимости сбора более обобщенной статистики;
- добавлять или исключать водителей без необходимости пересчета всего обучающего множества;
- собрать наглядную статистику по каждому доступному водителю за весь период активности;
- исключить введения штрафных санкций за разовые нарушения, которые не свойственны определенному водителю;
- перевести целочисленные значения частотности в долевые значения, указывающие какой процент значений попадает в определенный интервал по отношению ко всему множеству значений сигналов, которое было разбито на промежутки.

4. ПОСТРОЕНИЕ МОДЕЛИ КЛАСТЕРИЗАЦИИ

Цель моделей кластеризации – определить K количество групп для n объектов на основании метрик сходства таким образом, чтобы максимизировать схожесть между объектами одного класса, одновременно минимизировав схожесть между объектами разных классов.

Пусть $X = \{x_i\}$, $i = 1, \dots n$ – множество из n точек некоторой размерности d, которое необходимо разбить на K кластеров $C = \{c_k\}$, $k = 1, \dots, K$.

Алгоритм K-means стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров. Пусть μ_k – центр кластера c_k. Среднеквдаратическая ошибка между μ_k и точками кластера c_k определяется как

$$J(c_k) = \sum_{x_i \in c_k} x_i - \mu_k^2 \tag{3}$$

Цель модели K-means – минимизировать суммарную среднеквадратическую ошибку для всего множества кластеров K [4].

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} x_i - \mu_k^2 \rightarrow min \tag{4}$$

В работе были протестированы различные методы кластеризации, для каждого из которых была вычислена метрика силуэт. Коэффициент силуэт вычисляется с помощью среднего внутрикластерного расстояния и среднего расстояния до ближайшего кластера по каждому наблюдению. Силуэт вычисляется по формуле

$$s = \frac{b - a}{\max(a, b)} \tag{5}$$

где a – среднее расстояние от данного объекта до объектов из того же кластера, b – среднее расстояние от данного объекта до объектов из ближайшего кластера (отличного от того, в котором лежит сам объект). Эта метрика позволяет оценивать качество работы моделей кластеризации, обученных без учителя.

Таблица 1

Вариации данных

	Интервалы составлены с шагом 10 для скорости и с шагом 100 для ускорения
Частотные долевые значения в %	А
Частотные целочисленные значения	Б

Таблица 2

Результаты работы алгоритмов

Название модели с указанием основных параметров	**Метрика силуэт**	
	А	**Б**
KMeans n_clusters=5	0.459	0.505
KMeans А, Б: n_clusters=2, init='k-means++', max_iter=300, algorithm= 'auto'	0.686	0.801
AgglomerativeClustering n_clusters=5	0.449	0.434
AgglomerativeClustering А: n_clusters=2, affinity='l1', linkage='complete' Б: n_clusters=2, affinity='euclidean', linkage=' single'	0.682	0.892
AffinityPropagation damping=0.9	0.221	0.344
AffinityPropagation А: damping=0.8, max_iter=200, affinity='euclidean' Б: damping=0.85, max_iter=200, affinity='euclidean'	0.224	0.342

Название модели с указанием основных параметров	**Метрика силуэт**	
	А	**Б**
Birch threshold=0.01, n_clusters=5	0.446	0.434
Birch А: n_clusters=2, threshold=0.36, branching_factor=50 Б: n_clusters=2, threshold=0.1, branching_factor=50	0.686	0.804
DBSCAN А: eps=0.30, min_samples=9 Б: eps=5, min_samples=4	0.384	-0.281
DBSCAN А: eps=0.5, min_samples=8, algorithm='auto' Б: eps=9, min_samples=3, algorithm='auto'	0.473	-0.277
MiniBatchKMeans n_clusters=5	0.442	0.497
MiniBatchKMeans А: n_clusters=2, init='k-means++', max_iter=100, batch_size=4, reassignment_ratio=0.01 Б: n_clusters=2, init='k-means++', max_iter=100, batch_size=64, reassignment_ratio=0.01	0.686	0.802
MeanShift()	0.537	0.394
MeanShift max_iter=300, bandwidth=0.8	0.683	0.002
OPTICS А: eps=0.8, min_samples=10 Б: eps=10, min_samples=9	-0.643	-0.624
OPTICS А, Б: eps=1, min_samples=2, metric= 'euclidean', cluster_method= 'xi', algorithm= 'auto'	-0.291	-0.131
SpectralClustering n_clusters=5	0.444	-0.322
SpectralClustering А: n_clusters=2, eigen_solver='arpack', gamma=0.2, affinity='rbf', assign_labels='kmeans' Б: n_clusters=2, eigen_solver='lobpcg', affinity='nearest_neighbors', assign_labels='discretize'	0.686	0.352
GaussianMixture n_components=2	0.376	0.099
GaussianMixture А: covariance_type='spherical', n_components=2, init_params='kmeans' Б: covariance_type='tied', n_components=2, init_params='kmeans'	0.682	0.805

Из Таблицы 2 видно, что метрику силуэт удалось улучшить для большинства моделей. От вида данных, подаваемых в модель зависят полученные результаты, поэтому при разработке стоит учитывать разнообразие существующих алгоритмов, уделить внимание их подбору.

Рассмотрим подробнее модель KMeans на данных типа Б, так как она показала одну из наилучших метрик, а также проста в понимании. Построим гистограммы для центров кластеров, на которые модель разделила данные. Центр можно считать средним значением по кластеру. Из Рис. 2 видно, что 3 класс можно считать классом, относящимся к опасному вождению из-за распределения значений в интервалах с высокими скоростями.

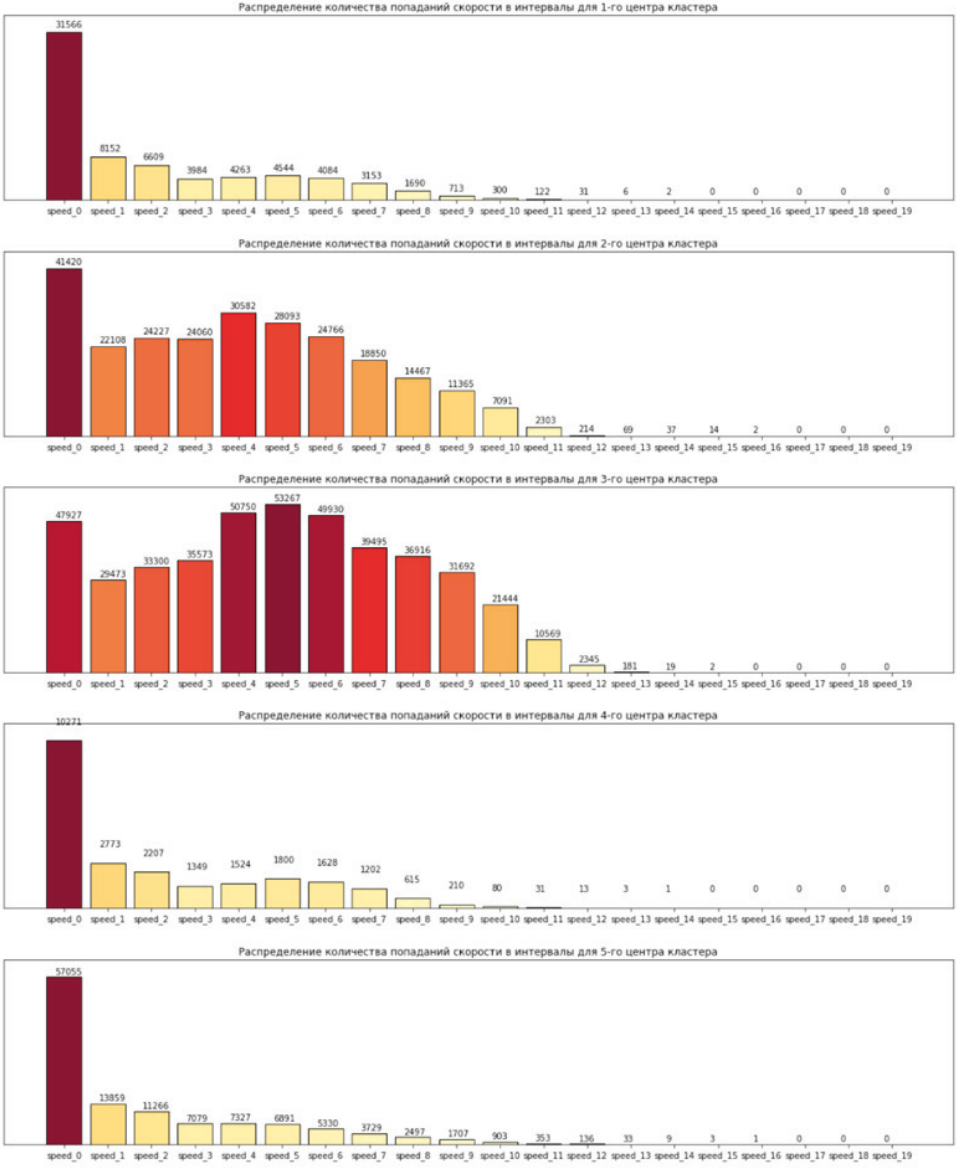

Рис. 2. Гистограммы для центров классов

Для выполнения расчетов и построения моделей был выбран язык программирования Python3, с использованием среды разработки Jupyter Notebook. Выбор обусловлен наличием большого количество библиотек с готовыми реализациями математического аппарата, а также наличием мощных инструментов для визуализации готовых результатов. В ходе выполнения работы были использованы готовые библиотеки, в том числе feather для хранения больших данных. Формат feather был использован как альтернатива csv, так как позволяет уменьшить объем сохраняемых файлов, увеличивает скорость чтения и записи данных. Были обучены модели кластеризации из широко используемого пакета Scikit Learn.

5. ЗАКЛЮЧЕНИЕ

Целью работы являлась реализация методов кластеризации в решении задачи разбиения множества водителей на классы, соответствующие степени опасности вождения. В качестве данных были использованы записи о сигналах водителей за некоторый временной период. Для агрегации датасета были использованы гистограммы распределения, позволяющие привести данные в формат, пригодный для обучения моделей машинного обучения, также, с помощью этого подхода была собрана собирательная статистика за все время активности водителя ТС.

В ходе выполнения работы был обучен ряд моделей кластеризации и подсчитала метрика силуэт. При принятии решения о качестве модели кластеризации, обученной без учителя, можно опираться на значение этой метрики в том случае, когда результаты работы модели совпадают с ожиданиями от поставленной задачи.

Использование описанного метода кластеризации и работы с гистограммами распределения сигналов в системах мониторинга качества вождения может улучшить контроль за поведением водителей, поможет выявлять и предотвращать поведение, способное привести к дорожно-транспортному происшествию.

Литература

1. *Кутуков Д.С.* Применение методов кластеризации для обработки новостного потока // Технические науки: проблемы и перспективы: материалы I Междунар. науч. конф. Санкт-Петербург: Реноме, 2011. С. 77–83.
2. *Дик Д.И.* Кластеризация водителей по стилям торможения // Курганский государственный университет. Вестник КГУ. 2012. № 2(24). С. 17–20.
3. Мониторинг манеры вождения [Электронный ресурс]: https://pro.yandex/ru-ru/moskva/knowledge-base/taxi/safety/monitoring-driving
4. *Jain A.K.* Data clustering: 50 years beyond K-means // Pattern Recognition Letters, 31(8). 2010. pp. 651–666.
5. *Huaikun Xiang , Jiafeng Zhu , Guoyuan Liang and Yingjun Shen.* Prediction of Dangerous Driving Behavior Based on Vehicle Motion State and Passenger Feeling Using Cloud Model and Elman Neural Network // Frontiers in Neurorobotics. April 2021. Vol. 15. p. 16.
6. *Omerustaoglu Furkan, Sakar C. Okan, Kar Gorkem.* Distracted driver detection by combining in-vehicle and image data using deep learning // Applied Soft Computing 96(6). 2020.
7. *J. Hartigan.* Clustering Algorithms // New York: Wiley, 1975.

Driver Clustering According to the Ratio of Dangerous Behavior Using Machine Learning Algorithms

Natalya D. Badanina*
Moscow Aviation Institute (MAI), Moscow, Russia
Keldysh Institute of Applied Mathematics, Moscow, Russia
ORCID: https://orcid.org/0000-0002-5301-1526
e-mail: natashabadanina99@gmail.com

Vladimir A. Sudakov**
Moscow Aviation Institute (MAI), Moscow, Russia
Keldysh Institute of Applied Mathematics, Moscow, Russia
ORCID: https://orcid.org/0000-0002-1658-1941
e-mail: sudakov@ws-dss.com

The paper conducts the research of defining dangerous driving of a vehicle using signals collected during the ride. A number of modern clustering models for drivers segmentation on classes based on the ratio of dangerous driving was used. New approach of data aggregation aiming to cluster data by signal distribution histograms was developed. Achieved results could be used in commercial systems that monitor the quality of drivers behavior in retrospective.

Keywords: clustering, distribution histograms, dangerous driving, monitoring systems, machine learning.

For citation:
Badanina N.D., Sudakov V.A. Driver Clustering According to the Ratio of Dangerous Behavior Using Machine Learning Algorithms. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2022. Vol. 12, no. 1, pp. 5–15. DOI: https://doi.org/10.17759/mda.2022120101 (In Russ., abstr. in Engl.).

References

1. Kutukov D.S. Primenenie metodov klasterizacii dlya obrabotki novostnogo potoka // Tekhnicheskie nauki: problemy i perspektivy: materialy I Mezhdunar. nauch. konf. Sankt-Peterburg: Renome, 2011. pp. 77–83. (In Russ.).
2. Dik D.I. Klasterizaciya voditelej po stilyam tormozheniya. Kurganskij gosudarstvennyj universitet. Vestnik KGU. 2012. № 2(24). pp. 17–20. (In Russ.).
3. Monitoring manery vozhdeniya [URL]: https://pro.yandex/ru-ru/moskva/knowledge-base/taxi/safety/monitoring-driving. (In Russ.).

***Natalya D. Badanina,** Master Student, Moscow Aviation Institute (National Research University), Programmer, Keldysh Institute of Applied Mathematics (Russian Academy of Sciences), Moscow, Russia, ORCID: https://orcid.org/0000-0002-5301-1526, e-mail: natashabadanina99@gmail.com

****Vladimir A. Sudakov,** Doctor of Technical Sciences, Professor, Moscow Aviation Institute (National Research University), Leading Researcher, Keldysh Institute of Applied Mathematics (Russian Academy of Sciences), Moscow, Russia, ORCID: https://orcid.org/0000-0002-1658-1941, e-mail: sudakov@ws-dss.com

4. Jain A.K. Data clustering: 50 years beyond K-means // Pattern Recognition Letters, 31(8). 2010. pp. 651–666.
5. Huaikun Xiang, Jiafeng Zhu , Guoyuan Liang and Yingjun Shen. Prediction of Dangerous Driving Behavior Based on Vehicle Motion State and Passenger Feeling Using Cloud Model and Elman Neural Network // Frontiers in Neurorobotics. April 2021. Vol. 15. p. 16.
6. Omerustaoglu Furkan, Sakar C. Okan, Kar Gorkem. Distracted driver detection by combining in-vehicle and image data using deep learning // Applied Soft Computing 96(6). 2020.
7. J. Hartigan. Clustering Algorithms // New York: Wiley, 1975.

Получена 04.03.2022
Принята в печать 14.03.2022

Received 04.03.2022
Accepted 14.03.2022