

Кластеризация объектов со слабо формализуемыми признаками на основе нейронной сети в виде слоя Кохонена

Игорь А. Кубасов	^{1,4}	igorak@list.ru
Александр В. Мельников	²	meln78@mail.ru
Сергей А. Мальцев	¹	ucgudvo@mail.ru
Илья Р. Нарусhev	³	nar_i@bk.ru

¹ Вычислительный центр ФКУ «ГИАЦ МВД России»; ул. Новочеремушкинская, 67; г. Москва; 117418, Россия

² Воронежский государственный университет инженерных технологий, пр-т Революции, 19, г. Воронеж, 394036, Россия

³ Воронежский институт МВД России; пр-т Патриотов, 53, г. Воронеж, 394052, Россия

⁴ Академия управления МВД России; ул. Зои и Александра Космодемьянских 8; г. Москва; 125993, Россия

Реферат. Анализ анкет несовершеннолетних в социальных сетях показывает, что подростки указывают в них сведения, повышающие уровень своей социальной желательности. Такая информация зачастую не имеет соответствия с реальным поведением подростка. Для полноценного анализа уровня девиантности несовершеннолетнего нужны инструменты охватывающие весь спектр показателей. В отличие от обычного подхода к кластеризации объектов на основе их объединения в группы по критерию минимума расстояния в многомерном пространстве при кластеризации признаков целесообразно учитывать их близость по способам получения информации и методам обработки этой информации инспектором по делам несовершеннолетних. На первом этапе исследования проводится кластеризация признаков девиации, на втором определении весовых коэффициентов, показателя степени девиантности внутри каждой из групп признаков, на третьем используется методика кластерно-иерархического подхода при формировании интегрального показателя оценки девиантного поведения несовершеннолетних. Показатель обладает значительной гибкостью учета соотношений между группами признаков и парциальными признаками за счет введения соответствующих множеств весовых коэффициентов. Сделан вывод о предпочтительности методов, основанных на кластеризации объектов в двумерном пространстве целевых показателей или счетов метода главных компонент, а также необходимости дополнительного анализа графической картины взаимного расположения объектов. Из сопоставления разных подходов: 1) кластеризации на основе обобщенного показателя качества и признака обратной девиантности, 2) кластеризации по двум счетам метода главных компонент; 3) кластеризации по всем признакам экспертизы можно сделать следующие выводы. Все способы правильно распределяют объекты по кластерам. Однако при сохранении основных итогов (выделение наилучших и наихудших объектов) результаты несколько отличаются. Это объясняется различным объемом и формами представления исходной информации. Номера активных нейронов (кластеров) программа назначает произвольно, поэтому чтобы упорядочить номера кластеров по некоему признаку (например, качеству объектов), требуется воспользоваться дополнительной графической информацией. С практической точки зрения предпочтительными являются первые два способа, основанные на кластеризации объектов в двумерном пространстве, методе главных компонент и анализе графической картины взаимного расположения объектов.

Ключевые слова: кластеризация, нейронные сети, многокритериальный анализ

Clustering of objects with poorly formalizable features based on a neural network in the form of Kohonen layers

Igor A. Kubasov	^{1,4}	igorak@list.ru
Aleksandr V. Melnikov	²	meln78@mail.ru
Sergey A. Maltsev	¹	ucgudvo@mail.ru
Ilya R. Narushev	³	nar_i@bk.ru

¹ Center of the PKU "HIAC of the Ministry of Internal Affairs of Russia", Novocheremushkinskaya str., 67, Moscow, 11741, Russia

² Voronezh state university of engineering technologies, Revolution Av., 19 Voronezh, 394036, Russia

³ Voronezh Institute of the Ministry of the Interior of Russia, Patriotov Av., 53 Voronezh, 394052, Russia

⁴ Academy of Management of the Ministry of the Interior of Russia, st. Zoi & Alexandra Kosmodemyanskikh 8, Moscow, 125993, Russia

Summary. Analysis of profiles of minors in social networks shows that teenagers indicate in them information that increases the level of their social desirability. Such information often does not correspond to the real behavior of the teenager. For a full analysis of the level of deviance of a minor need tools covering the full range of indicators. In contrast to the usual approach to clustering objects based on their Association in groups by the criterion of the minimum distance in multidimensional space when clustering features it is advisable to take into account their proximity to the methods of obtaining information and methods of processing of this information by the inspector for minors. In the first phase of the study is the clustering of signs of deviation, the second the determination of the weighting factors of indicator of the degree of deviance within each group of signs, the third uses the method of cluster-hierarchical approach to forming integral indicator of assessment of deviant behavior of minors. The indicator has a considerable flexibility of the correlation between groups of symptoms and partial characteristics through the introduction of appropriate sets of weighting coefficients. The conclusion is made about the preference of methods based on clustering of objects in the two-dimensional space of targets or accounts of the principal components method, as well as the need for additional analysis of the graphical picture of the relative location of objects. From the comparison of different approaches: 1) clustering on the basis of the generalized indicator of quality and the sign of reverse deviance, 2) clustering on two accounts of the principal components method; 3) clustering on all signs of examination, the following conclusions can be drawn. All methods properly allocate the objects to clusters. However, when you save the main totals (highlighting the best and worst features), the results are slightly different. This is due to the different volume and forms of presentation of the source information. The program assigns numbers of active neurons (clusters) arbitrarily, so in order to arrange the cluster numbers by some feature (for example, the quality of objects), you need to use additional graphical information. From a practical point of view, the first two methods are preferred, based on clustering objects in two-dimensional space, the method of principal components and the analysis of the graphical picture of the mutual location of objects.

Keywords: clustering, neural networks, multicriteria analysis

Для цитирования

Нарусhev И.Р., Мальцев С.А., Мельников А.В., Кубасов И.А. Кластеризация объектов со слабо формализуемыми признаками на основе нейронной сети в виде слоя Кохонена // Вестник ВГУИТ. 2018. Т. 80. № 3. С. 86–91. doi:10.20914/2310-1202-2018-3-86-91

For citation

Narushev I.R., Malcev S.A., Melnikov A.V., Kubasov I.A. Clustering of objects with poorly formalizable features based on a neural network in the form of Kohonen layers. *Vestnik VGUIT* [Proceedings of VSUET]. 2018. vol. 80. no. 3. pp. 86–91. (in Russian). doi:10.20914/2310-1202-2018-3-86-91

Введение

Анализ анкет несовершеннолетних в социальных сетях показывает, что подростки указывают в них сведения, повышающие уровень своей социальной желательности. Такая информация зачастую не имеет соответствия с реальным поведением подростка. Для полноценного анализа уровня девиантности несовершеннолетнего нужны инструменты охватывающие весь спектр показателей, характеризующих уровень девиантности. Такие показатели могут быть получены в ходе детального изучения несовершеннолетнего, опроса родителей, соседей, педагогического состава образовательной организации, где обучается подросток и т.д.

В построенной модели девиантного поведения [1] присутствует более 30 критериев девиантного поведения классифицированных в соответствии с характером направленности и вносящие различный вклад в обобщенный показатель девиантного поведения. Следующим актуальным этапом в анализе является решение задачи кластеризации девиаций несовершеннолетних с дальнейшим принятием научно обоснованных решений.

В работе рассмотрим и сравним следующие возможные варианты кластеризации паттернов несовершеннолетних согласно критериям девиантности [1]:

1) разделение выборки пользователей социальных сетей на 3 кластера на основе векторов обобщенного показателя $J_{\text{общ}}$ и признака обратной девиантности \hat{J} ;

2) кластеризация на основе главных компонент референтной матрицы X ;

3) кластеризация на основе всех признаков объектов экспертизы.

Интегральный показатель девиантности несовершеннолетних

На первом этапе исследования проводится кластеризация признаков девиации, на втором определении весовых коэффициентов, показателя степени девиантности внутри каждой из групп признаков, на третьем используется методика кластерно-иерархического подхода при формировании интегрального показателя оценки девиантного поведения несовершеннолетних.

Определение интегральных показателей, характеризующих уровень девиантности несовершеннолетних, возможно с использованием кластерно-иерархического подхода. На первом этапе такого исследования проводится кластеризация

признаков девиации. В отличие от обычного подхода к кластеризации объектов на основе их объединения в группы по критерию минимума расстояния в многомерном пространстве при кластеризации признаков целесообразно учитывать их близость по способам получения информации и методам обработки этой информации инспектором по делам несовершеннолетних.

Согласно мультипликативной модели обобщенного показателя девиации J соответствует сумме произведений весовых коэффициентов v с тремя показателями: девиация J_1 , жертва J_2 , группа риска J_3 [2]. Выбор групповых весовых коэффициентов позволяет установить требуемое соотношение между вкладами оценок показателей отклонения в комплексный показатель девиантности. Показатель обладает значительной гибкостью учета соотношений между группами признаков и парциальными признаками за счет введения соответствующих множеств весовых коэффициентов [2].

Обобщенный комплексный показатель девиации $J_{\text{общ}}$ представим в следующем виде:

$$J_{\text{общ.}} = v_1 J_1 + v_2 J_2 + v_3 J_3, \quad (1)$$

где v_1, v_2, v_3 – межгрупповые весовые коэффициенты значений соответствующих типов отклонения J_1, J_2, J_3 .

Обобщенный показатель девиации $J_{\text{общ.}}$, полученный на основе кластерно-иерархического подхода, имеет вид

$$J = \sum_{j=1}^n \tilde{V}_j \times \left[\hat{V}_{\text{нал.}j} \frac{\sum_i V_{i,\text{нал.}j} \hat{x}_{ij}}{\sum_i V_{i,\text{нал.}j}} + \hat{V}_{\text{кач.пр.}j} \frac{\sum_l V_{l,\text{кач.пр.}j} \hat{x}_{lj}}{\sum_l V_{l,\text{кач.пр.}j}} \right]$$

где $V_{\text{нал.}}, V_{\text{кач.пр.}}, \tilde{V}_1, \tilde{V}_2, \tilde{V}_3$ – групповые весовые коэффициенты, определяющие предпочтительность качественных признаков (score), признаков наличия (existence), и весовые коэффициенты, относящиеся к группам критериев, \hat{x}_{ij} – нормированный признак, $J_{\text{общ.}}$ – обобщенная функция девиации.

Множества $\{V_j, V_l\}$ определяют относительный вклад отдельных признаков (частных критериев), n – количество типов поведения, весовые коэффициенты критериев оценки девиантности представлены в таблице 1.

Таблица 1.

Весовые коэффициенты критериев девиантности используемых для формирования обобщенного показателя девиации

Table 1.

Weight coefficients of deviance criteria for the adoption of the generalized deviation index

	Наименование критериев The name of the criteria	Вид кр. Type cr.	Вес приз. The weight of the sign	Вес групп Weight groups
Девиация Deviation	Алкоголь, курение табака Alcohol, tobacco smoking	нал.ех.	0,235	0,54
	Наркотики, одурманивающие вещества Drugs, intoxicants	нал.ех.	0,549	
	Криминальная субкультура Criminal subculture	нал.ех.	0,1	
	Нетрадиционные сексуальные отношения Perversion sexual relations	нал.ех.	0,116	
	Порнография Pornography	кач.сcore	0,108	
	Азартные игры (на деньги) Gambling (for money)	кач.сcore	0,064	
	Жестокость и насилие по отношению к сверстникам Vio- lence against peers	кач.сcore	0,329	
	Жестокое обращение с животными Cruelty to animals	кач.сcore	0,182	
	Экстремизм (дискриминация) Extremism (discrimination)	кач.сcore	0,316	
Группа риска Riskgroup	Игра не соответствует цензу The game is not age-appropriate	нал.ех.	0,196	0,163
	Видео не соответствует цензу The video is not age-appropriate	нал.ех.	0,493	
	Аудио не соответствует цензу Audio does not match age	нал.ех.	0,311	
	Нецензурная брань Foul language	кач.сcore	0,277	
	«Троллинг» ровесников «Trolling» peers	кач.сcore	0,095	
	«Троллинг» взрослых "Trolling" adults	кач.сcore	0,16	
	Порнографический контент Pornographic content	кач.сcore	0,467	
Жертва Victim	Суицид, вред здоровью Suicide, harm to health	нал.ех.	0,559	0,297
	Персональные данные Personal information	нал.ех.	0,089	
	Опасное «хобби» Dangerous «hobby»	нал.ех.	0,352	
	Общение со взрослыми Chat with strangers	кач.сcore	0,16	
	Демонстрация ценностей Demonstration of personal values	кач.сcore	0,095	
	Участие в розыгрышах, лотереях Participation in lotteries	кач.сcore	0,277	
	Нарушение / пренебрежение ПДД Violation / disregard of traffic rules	кач.сcore	0,467	

Распределение показателей девиантности $J_{общ}$ выборки из 9 исследуемых паттернов, проанализированных в работе [1] представлено на рисунке 1

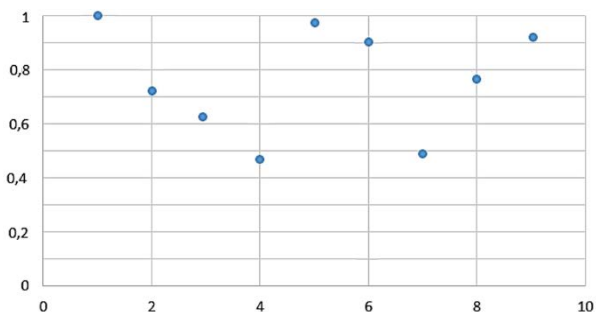


Рисунок 1. Обобщенные показатели девиантности девяти исследуемых паттернов

Figure1.The generalized deviation index of nine studied patterns

Разделение объектов экспертизы на 3 класса

Создадим слой Кохонена [3,4] с помощью разработанной программы (M-функции) на основе встроенной функции newsc (рисунок 2).

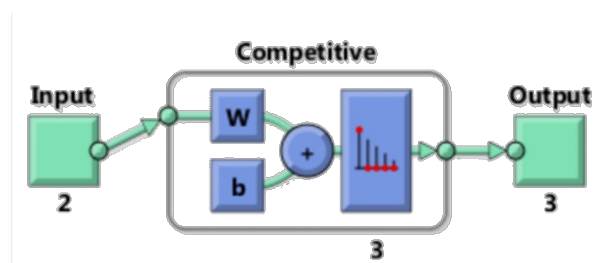


Рисунок 2. Самоорганизующаяся сеть с тремя нейронами

Figure2.Self-organizing network with three neurons

Это слой конкурирующего типа, поскольку в нем применена конкурирующая функция активации. Номер активного нейрона i определяет группу (кластер), к которой наиболее близок входной вектор.

В созданном M -файле языка MATLAB использованы, в частности, процедуры [3]

```
net = newc[-3 3; -3 3], c, 0, 1);
net.trainParam.epochs = 100;
net = train(net, Px);
w = net.IW(1); w = cell2mat(w),
```

определяющие: количество кластеров (c); количество итераций ($epochs$); процедуру обучения ($train$), веса w настроенной нейронной сети.

Пример кластеризации и после 500 итераций получим следующие результаты (рисунок 3).

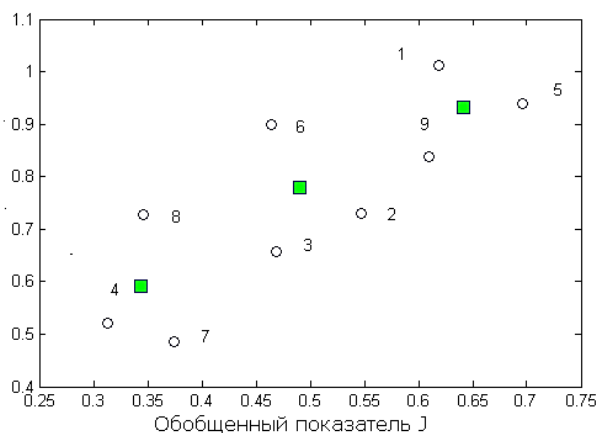


Рисунок 3. Кластеризация объектов экспертизы на 3 кластера

Figure 3. Clustering of examination objects for 3 clusters

На этом рисунке квадратными маркерами обозначены центры кластеров, а кружками – данные объектов экспертизы. Как видим, к первому кластеру (нормальное поведение) относятся 4, 7, 8 объекты, ко второму кластеру (приемлемое поведение) – 3, 2, 6 объекты, к третьему кластеру (девиантное поведение) – 1, 9, 5 объекты.

Чтобы исключить эффект «мертвых нейронов» и сделать все нейроны чувствительными к поступающим на вход векторам, используются положительные смещения ($biases$), которые позволяют нейрону стать конкурентным [4]. Поэтому после настройки все смещения положительны:

В начале процесса обучения параметр активности различных нейронов принимает значение

$$a_0 = \frac{1}{S} \quad (2)$$

где S – количество нейронов конкурирующего слоя, равное числу кластеров. По окончании обучения сети наиболее активным становится второй нейрон.

Кластеризация на основе выделения главных компонент

Существо метода главных компонент состоит в разложении матрицы референтных данных X на произведение матриц «счетов» T и транспонированной матрицы «нагрузок» P [4]:

$$X = TP^T + E, \quad (3)$$

где E – матрица остатков.

Как отражено в работе [2], ограничение только двумя первыми векторами счетов t_1, t_2 дает возможность качественно осуществить кластеризацию и представить взаимное расположение объектов графически.

Осуществим кластеризацию исследуемых объектов в два этапа: 1) на основе матрицы референтных данных X для 9 исследуемых паттернов, осуществим процедуру SVD (Singular Value Decomposition) для нахождения векторов счетов и векторов нагрузок; 2) применим разработанную программу (M -файл) конкурентного обучения к разделению множества первых двух счетов t_1, t_2 на 3 кластера.

В методе SVD обычным приемом нахождения главных компонент является предварительное центрирование и нормировка (шкалирование) векторов-строк матрицы референтных данных. Воспользовавшись встроенной функцией $prestd$ языка MATLAB, осуществим эти операции и получим

Для нахождения матрицы счетов T и матрицы нагрузок P применим встроенную функцию SVD языка MATLAB, согласно которой матрица \hat{X} разлагается в произведение матриц U, S, V :

$$\hat{X} = U \cdot S \cdot V \quad (4)$$

После нахождения матрицы T перейдем к задаче собственно кластеризации на основе разработанной программы на языке MATLAB, используя только первые два счета t_1, t_2 .

Для того, чтобы исключить эффект «мертвых нейронов» и сделать все нейроны чувствительными к поступающим на вход векторам, используются смещения ($biases$), которые позволяют нейрону стать конкурентным с нейронами-победителями. Поэтому все смещения положительны:

Выберем три кластера для счетов t_1, t_2 , характеризующих обобщенный комплексный показатель девиантности (1 – хорошее значение, 2 – промежуточное (нормальное) значение, 2 – неудовлетворительное значение).

На основе обучения слоя Кохонена определим центры кластеров рисунок 4.

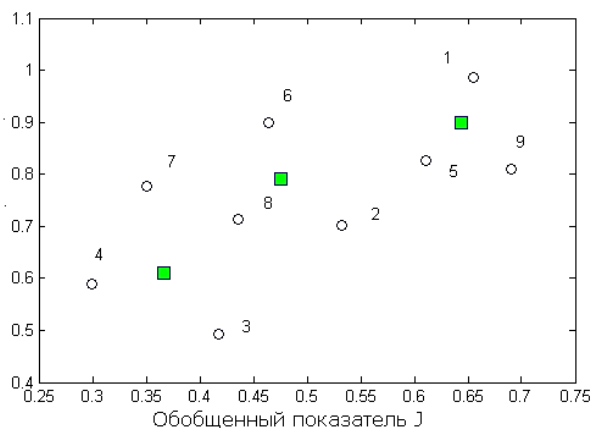


Рисунок 4. Кластеризация 9 объектов используя МГК
Figure 4. Clustering 9 objects using the PCA

На рисунке 4 квадратными маркерами обозначены центры кластеров, крестиками – значения счетов объектов экспертизы. Как видим, к объектам с хорошими характеристиками относятся 4, 7, 3 с нормальными характеристиками относятся 8, 6, 2, а к объектам с неудовлетворительными характеристиками – 1, 9, 5.

Сравнивая между собой результаты кластеризации, полученные на основе первого подхода (рисунок 3) и второго подхода (рисунок 4), можно видеть, что они несколько различаются. Этого и следовало ожидать, поскольку в первом случае кластеризация осуществлялась на основе обобщенного показателя Лобц., а во втором случае – на основе выделения методом МГК двух первых счетов t_1, t_2 референтной матрицы X. Однако, наиболее важные объекты в обоих случаях выявлены правильно, что позволяет на практике применять оба рассмотренных подхода.

Кластеризация на основе всех признаков объекта экспертизы

При построении рисунка 4 была использована кластеризация по двум счетам (4) метода МГК. Хотя известно [4], что главные компоненты учитывают до 85–95% полезной информации, содержащейся в матрице референтных данных X, такой метод является приближенным. Поэтому для проверки полученных результатов далее осуществим кластеризацию по всем 4 признакам (столбцам) матрицы X.

ЛИТЕРАТУРА

- 1 Нарушев И.Р., Мельников А.В., Денисенко В.В. Модели обобщенного показателя девиантного поведения несовершеннолетних // Вестник Воронежского института МВД России. 2018. № 1. С. 44–50
- 2 Мельников А.В., Мальцев С.А. Учет признаков инженерной оценки в экспертизе приемно-контрольных приборов охранно-пожарной сигнализации // Вестник Воронежского института МВД России. 2016. № 3. С. 51–57.
- 3 Хайкин С. Нейронные сети: полный курс. М.: Издательский дом Вильямс, 2008.

В созданном M-файле языка MATLAB использованы процедуры

```
net = newc([0 1; 0 1; 0 1; 0 1], 3);
wts = midpoint(3, p)
biases = initcon(3)
net.trainParam.epochs = 500;
net = train(net, p),
```

определяющие: количество кластеров (3); начальные значения элементов матрицы весов в центре интервала входных значений с помощью функции midpoint; начальные смещения с помощью функции initcon; количество итераций (epochs).

В первый кластер (хорошая характеристика) попадают 7, 4, 8 объекты; во второй кластер (нормальная характеристика) – 5, 3, 6 объекты; в третий кластер (неудовлетворительная характеристика) – 2, 1, 9 объекты.

Заключение

Из сопоставления разных подходов: 1) кластеризации на основе обобщенного показателя качества и признака обратной девиантности, 2) кластеризации по двум счетам t_1, t_2 метода МГК; 3) кластеризации по всем признакам экспертизы можно сделать следующие выводы.

1. Все способы правильно распределяют объекты по кластерам. Однако при сохранении основных итогов (выделение наилучших и наихудших объектов) результаты несколько отличаются. Это объясняется различным объемом и формами представления исходной информации.

2. Номера активных нейронов (кластеров) программа назначает произвольно. Таким образом, чтобы упорядочить номера кластеров по некоему признаку (например, качеству объектов), требуется воспользоваться дополнительной графической информацией.

3. С практической точки зрения предпочтительными являются первые два способа, основанные на кластеризации объектов в двумерном пространстве показателей $J_{\text{кач}}, \hat{P}$ или счетов t_1, t_2 МГК и анализе графической картины взаимного расположения объектов.

4 Zhang Z. Artificial neural network // Multivariate Time Series Analysis in Climate and Environmental Research. Springer, Cham, 2018. P. 1-35.

5 Witten I, Frank E., Hall M. Data Mining: Practical Machine Learning Tools and Techniques. N.Y.: Morgan Kaufmann, 2011. 664 p.

REFERENCES

- 1 Narushev I.R., Mel'nikov A.V., Denisenko V.V. Models of a generalized indicator of deviant behavior of minors. Vestnik Voronezhskogo instituta MVD Rossii [Bulletin of the Voronezh Institute of the Ministry of Internal Affairs of Russia] 2018. no. 1. pp. 44–50 (in Russian)

2 Mel'nikov A.V., Mal'cev S.A Consideration of signs of engineering assessment in the examination of control panels of fire alarm. *Vestnik Voronezhskogo instituta MVD Rossii* [Bulletin of the Voronezh Institute of the Ministry of Internal Affairs of Russia] 2016. no. 3. pp. 51–57. (in Russian)

3 Haikin S. Nejrionnye seti: polnyj kurs [Neural networks: full course] Moscow, Williams publishing house, 2008. (in Russian)

СВЕДЕНИЯ ОБ АВТОРАХ

Игорь А. Кубасов д.т.н., профессор кафедры информационных технологий, Академия управления МВД России, ул. Зои и Александра Космодемьянских 8; г. Москва; 125993, Россия, igorak@list.ru

Александр В. Мельников д.т.н., профессор, кафедра высшей математики и информационных технологий, Воронежский государственный университет инженерных технологий, пр-т Революции, 19, г. Воронеж, 394036, Россия, meln78@mail.ru

Сергей А. Мальцев старший инженер Вычислительного центра, ФКУ «ГИАЦ МВД России», «ГИАЦ МВД России», ул. Новочеремушкинская, 67; г. Москва, 117418, Россия, ucgudvdo@mail.ru

Илья Р. Нарусhev адъюнкт, Воронежский институт МВД России, пр-т Патриотов, 53, г. Воронеж, 394052, Россия, nar_i@bk.ru

КРИТЕРИЙ АВТОРСТВА

Игорь А. Кубасов написал рукопись, корректировал её до подачи в редакцию и несёт ответственность за плагиат
Александр В. Мельников консультация в ходе исследования

Сергей А. Мальцев предложил методику проведения эксперимента

Илья Р. Нарусhev обзор литературных источников по исследуемой проблеме, провёл эксперимент, выполнил расчёты

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

ПОСТУПИЛА 16.05.2018

ПРИНЯТА В ПЕЧАТЬ 27.07.2018

4 Zhang Z. Artificial neural network. Multivariate Time Series Analysis in Climate and Environmental Research. Springer, Cham, 2018. pp. 1-35

5 Witten I., Frank E., Hall M. Data Mining: Practical Machine Learning Tools and Techniques. New-York, Morgan Kaufmann, 2011. 664 p.

INFORMATION ABOUT AUTHORS

Igor A. Kubasov doctor of technical sciences, professor, department of information technology, Academy of Management of the Ministry of Internal Affairs of Russia, st. Zoi & Alexandra Kosmodemyanskikh 8, Moscow, 125993, Russia, igorak@list.ru

Aleksandr V. Melnikov Dr. Sci. (Engin.), professor, Higher Mathematics and Information Technologies department, Voronezh state university of engineering technologies, Revolution Av., 19 Voronezh, 394036, Russia, meln78@mail.ru

Sergey A. Maltsev senior engineer of the computing center, ИИАС, of the Ministry of Internal Affairs of Russia, Novocheremushkinskaya str., 67, Moscow, 11741, Russia, ucgudvdo@mail.ru

Ilya R. Narushev graduate student, Voronezh Institute of the Ministry of the Interior of Russia, Patriotov Av., 53 Voronezh, 394052, Russia, nar_i@bk.ru

CONTRIBUTION

Igor A. Kubasov wrote the manuscript, correct it before filing in editing and is responsible for plagiarism

Aleksandr V. Melnikov consultation during the study

Sergey A. Maltsev proposed a scheme of the experiment

Ilya R. Narushev review of the literature on an investigated problem, conducted an experiment, performed computations

CONFLICT OF INTEREST

The authors declare no conflict of interest.

RECEIVED 5.16.2018

ACCEPTED 7.27.2018