DOI: 10.24143/2072-9502-2019-4-106-114
УДК 004.855

ПОВЫШЕНИЕ КАЧЕСТВА КЛАССИФИКАЦИИ ОБЪЕКТОВ НА ОСНОВЕ ВВЕДЕНИЯ НОВОЙ МЕТРИКИ КЛАСТЕРИЗАЦИИ

Р. Ю. Демина[1], И. М. Ажмухамедов[2]

[1] Астраханский государственный технический университет, Астрахань, Российская Федерация

[2] Астраханский государственный университет, Астрахань, Российская Федерация

Кластеризация объектов является одной из основных задач машинного обучения. Она нашла широкое применение в различных предметных областях: маркетинге, социологии, психологии и пр. В основе алгоритмов кластеризации, как правило, лежит метрика, отражающая расстояние между объектами. Однако в ряде случаев пользоваться расстоянием между объектами нецелесообразно. В определенных ситуациях можно говорить о том, что один объект похож на второй, притом что второй объект не похож на первый. Такими объектами могут являться, например, оригинал картины и ее копия. Для подобных случаев в работе предложена мера схожести объектов, которая отражает, какая часть признаков одного объекта содержится в другом. На основании данной меры строится матрица схожести, анализ которой позволяет выявлять кластеры взаимно схожих объектов. При проведении апробации предложенного метода кластеризации индекс Рэнда (доля корректно связанных или не связанных между собой объектов) составил 0,93. Предложен алгоритм, позволяющий формировать множество максимально различающихся между собой объектов. Множество объектов, сформированное подобным образом, может в дальнейшем стать обучающим множеством для классификаторов и повысить верность их распознавания.

Ключевые слова: кластеризация, метрика, сравнение, мера схожести, обучающее множество, признаки объекта, индекс Рэнда.

Для цитирования: *Демина Р. Ю., Ажмухамедов И. М.* Повышение качества классификации объектов на основе введения новой метрики кластеризации // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2019. № 4. С. 106–114. DOI: 10.24143/2072-9502-2019-4-106-114.

Введение

Кластеризация – одна из важнейших задач машинного обучения, которая позволяет выявить внутри множества такие кластеры (подмножества объектов), что объекты внутри них будут более похожи друг на друга, нежели на объекты других кластеров. Кластеризация относится к классу задач обучения без учителя и позволяет получить новое знание, которое эксперты выявить не смогли. Кластеризация широко применяется в разнообразных предметных областях: сегментации изображений, маркетинге, прогнозировании [1, 2]. Данная задача является фундаментальной, и ее решением занимаются ученые всего мира [3–5].

При решении задач кластеризации, как правило, используется метрика. Множество X является метрическим пространством, если каждой паре элементов x и y этого множества поставлено в соответствие неотрицательное число $p(x, y)$, называемое расстоянием между элементами x и y, такое, что для любых элементов x, y, z множества X выполняются следующие условия [6]:

1. $p(x, y) = 0 \Leftrightarrow x = y$;
2. $p(x, y) = p(y, x)$;
3. $p(x, y) \le p(x, z) + p(z, y)$, $z \in R$, $z = (z_1, z_2, \ldots, z_n)$.

Рассмотрим подробнее второе свойство: расстояние от x до y равно расстоянию от y до x. Для большинства задач данное свойство органично и не вызывает вопросов. Если от города

N до города *M* 100 км, то и от города *M* до города *N* столько же. Однако существуют задачи, при решении которых возникают некоторые нюансы, например, если мы рассматриваем задачу определения степени схожести объектов в некоторых предметных областях.

Предположим, что имеется оригинал картины и ее копия. При том, что полотна могут быть неразличимы невооруженным взглядом, говорят только о похожести копии на оригинал. О том, что оригинал похож на копию, речи идти не может: при всей точности копии оригинал все равно представляет бóльшую ценность, поскольку содержит больше информации, хотя мы в данный момент можем и не знать, какой именно. Получается, что $p(x, y) \neq p(y, x)$, где x – оригинал, а y – копия.

Другой случай. Жилой дом и собачья будка являются архитектурными сооружениями. Допустимо сказать, что будка – это дом для собаки. Однако сравнение человеческого дома с собачьей конурой некорректно. В данном случае также получается, что «расстояние» от будки до дома отличается от «расстояния» от дома до будки.

Остановимся также на примере с компьютерными файлами. Рассмотрим два исполняемых файла. Функционал одного из них ограничивается только подключенными к нему библиотеками, второй же, кроме всего этого, содержит и свой код. Очевидно, что в данном случае «расстояния» между ними не совпадают, поскольку один из файлов является частью другого.

В подобных рассмотренным случаях полезно иметь возможность оперировать понятием, которое могло бы отразить приведенные нюансы сравнения двух объектов.

Для предметных областей, описанных выше, и им подобных алгоритм кластеризации не может опираться на классическую метрику. Необходимо ввести некую меру схожести, которая бы учитывала описанный выше нюанс, и на ее основе разработать метод кластеризации. Исходя из этого нами была сформулирована основная задача исследования – разработать метод кластеризации объектов, на основании которого в дальнейшем будет происходить отбор объектов в обучающее множество.

Методы и результаты исследования

Мера схожести. Для решения поставленной задачи введем понятие *меры схожести* двух объектов: мера схожести объекта A с объектом B есть отношение числа уникальных значений признаков объекта A, которые встречаются в наборе уникальных значений признаков объекта B, к количеству уникальных значений признаков объекта A [7]:

$$\rho(A, B) = \frac{\left|\tilde{A} \cap \tilde{B}\right|}{\left|\tilde{A}\right|},$$

где $\tilde{A}$ – множество значений признаков объекта A; $\tilde{B}$ – множество значений признаков объекта B.

Величина $\rho(A, B) \in [0, 1]$ и характеризует долю уникальных значений признаков объекта A, входящих в множество уникальных значений признака объекта B.

Рассмотрим свойства введенной меры схожести:

– если ни одно из значений признака объекта A не входит в множество значений признака объекта B, то $\rho(A, B) = 0$;

– если все значения признаков объекта A входят в объект B, то $\rho(A, B) = 1$;

– в общем случае $\rho(A, B) \neq \rho(B, A)$;

– объекты A и B являются взаимно схожими в равной степени, если $\rho(A, B) \approx \rho(B, A) \approx 1$.

Рассмотрим также свойства меры схожести как бинарного отношения:

– рефлексивность: $\forall x \in M(xRx)$. Для любого объекта A из обучающего множества можно рассчитать меру схожести $\rho(A, A)$, которая при этом всегда будет равна 1;

– антисимметричность: $\forall x, y \in M(xRy \& Rx \Rightarrow x = y)$. Если для объектов A и B $\rho(A, B) = \rho(B, A)$, то в рамках данной методики файлы A и B считаются идентичными (взаимно схожими).

Для примера рассмотрим ситуацию со сравнением жилых строений Д1 и Д2 (табл. 1).

Таблица 1

Пример расчета меры схожести для жилых строений Д1 и Д2

Д1	**Д2**
	
Признак: архитектурная деталь	
Список деталей	
Стена Крыша Окно Дверь Труба Балкон Веранда Ступеньки Гараж	Стена Крыша Окно Дверь Ступеньки Декоративный кирпич
Всего деталей	
9	6
Список общих деталей	
Стена Крыша Окно Дверь Ступеньки	
Всего общих деталей	
5	
ρ (Д1, Д2)	ρ (Д2, Д1)
5/9 ≈ 0,56	5/6 ≈ 0,83

Второй дом почти полностью состоит из деталей, которые есть в первом доме. При этом первый дом «похож» на второй лишь наполовину. Таким образом, если необходимо на примере одного дома продемонстрировать максимум архитектурного разнообразия, то, основываясь на рассчитанных мерах схожести, целесообразно выбрать первый дом.

Аналогично меру схожести можно применить для сравнения двух исполняемых файлов. Это позволит проанализировать, насколько два исполняемых файла схожи по своему функционалу, по подключенным библиотекам и пр. Это поможет определить, чем «занимается» неизвестный исполняемый файл: обрабатывает фото/видеоконтент, является ли текстовым редактором или администраторской утилитой.

Такая возможность актуальна в сфере антивирусного эвристического анализа на этапе формирования обучающего множества, которое должно состоять из максимально разнообразных элементов.

В рамках данной задачи сравнение двух исполняемых файлов будет проходить следующим образом. В качестве признака может выступать, например, n-грамма – последовательность рядом стоящих n байт. Каждый файл необходимо считать побайтово, составить перечень уникальных n-грамм, выявить перечень общих n-грамм для данных двух файлов и вычислить соотношение общих n-грамм к количеству всех n-грамм каждого файла.

Пример расчета представлен в табл. 2 для файлов $F1$ и $F2$.

Таблица 2

Пример расчета меры схожести для файлов *F*1 и *F*2

***F*1**	***F*2**
Файлы считаны побайтово	
123, 56, 87, 38, 176, 56, 73, 145	123, 56, 87, 38, 176, 68, 4, 56, 73, 145
Признак: 3-грамма	
Список 3-грамм	
123, 56, 87 56, 87, 38 87, 38, 176 38, 176, 56 176, 56, 73 56, 73, 145	123, 56, 87 56, 87, 38 87, 38, 176 38, 176, 68 176, 68, 4 68, 4, 56 4, 56, 73 56, 73, 145
Всего 3-грамм	
6	8
Список общих 3-грамм	
123, 56, 87 56, 87, 38 87, 38, 176 56, 73, 145	
Всего общих 3-грамм	
4	
ρ (*F*1, *F*2)	ρ (*F*2, *F*1)
4/6 ≈ 0,67	4/8 = 0,5

Подобные расчеты в дальнейшем помогают выбрать наиболее информативные для обучения классификаторов исполняемые файлы.

Матрица схожести. В ситуации, когда имеется некое множество, которое необходимо проанализировать на степень оригинальности представленных в нем объектов, имеет смысл провести попарное сравнение элементов и представить собранную информацию в виде матрицы, которую назовем матрицей схожести (МС).

Матрица схожести – квадратная матрица, состоящая из элементов ρ_{ij} – мер схожести i-го объекта с j-м [8]. Матрица схожести может быть симметрична, если $\forall A, B$: $\rho(A, B) = \rho(B, A)$. Но такая ситуация исключительна и возможна скорее только теоретически; на практике же матрица асимметрична, т. к. $\rho(A, B)$, как правило, отличается от $\rho(B, A)$.

На главной диагонали МС всегда расположены единицы, т. к. $\rho(A, A) = 1$.

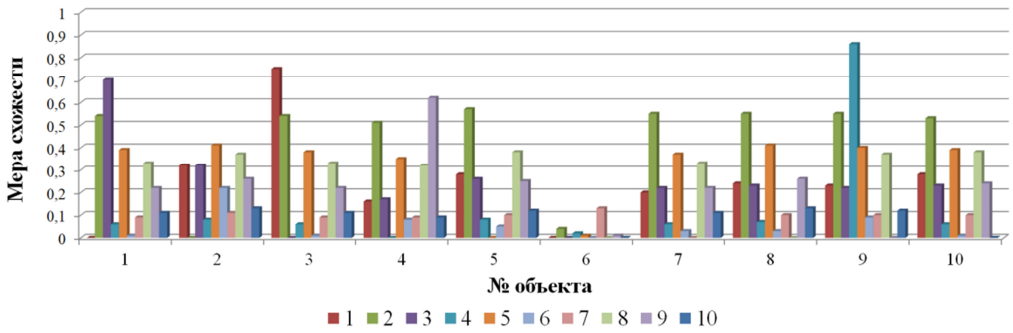
Рассмотрим пример МС (табл. 3).

Таблица 3

Пример матрицы схожести

№ объекта	**Мера схожести для объекта №**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1	0,32	0,75	0,16	0,28	**0,00**	0,20	0,24	0,23	0,28
2	0,54	1	0,54	0,51	0,57	**0,04**	0,55	0,55	0,55	0,53
3	0,70	0,32	1	0,17	0,26	**0,00**	0,22	0,23	0,22	0,23
4	0,06	0,08	0,06	1	0,08	**0,02**	0,06	0,07	**0,86**	0,06
5	0,39	0,41	0,38	0,35	1	**0,01**	0,37	0,41	0,40	0,39
6	0,01	0,22	0,01	0,08	0,05	1	0,03	0,03	0,09	0,01
7	0,09	0,11	0,09	0,09	0,10	**0,13**	1	0,10	0,10	0,10
8	0,33	0,37	0,33	0,32	0,38	**0,00**	0,33	1	0,37	0,38
9	0,22	0,26	0,22	0,62	0,25	**0,01**	0,22	0,26	1	0,24
10	0,11	0,13	0,11	0,09	0,12	**0,00**	0,11	0,13	0,12	1

Для наглядности МС может быть также представлена графически в виде гистограммы. На ней можно изобразить графически, насколько каждый объект схож с другими объектами (чтобы не загромождать рисунок, сравнение каждого объекта с самим собой опущено) (рис).



Графическое изображение матрицы схожести

Графическое представление более наглядно и позволяет проанализировать заданное множество объектов визуально при отборе объектов в обучающее множество. Так, на рис. видно, что наибольшую меру схожести имеют объекты № 9 и № 4. Также видно, что с объектом № 2 все остальные файлы имеют примерно равную меру схожести (0,5–0,6). А объект № 6 в наименьшей степени схож с остальными объектами.

Анализ МС позволяет кластеризовать множество объектов $F=\{f_i\}$: разбить его на подмножества взаимно схожих фалов. А дальнейший отбор объектов из этих подмножеств позволит сформировать $F'(|F| \geq |F'|)$ – множество максимально различающихся между собой объектов, что может быть полезно для формирования обучающего множества кластеризатора.

Кластеризация и отбор максимально различающихся между собой объектов. Для формирования множества максимально различающихся между собой объектов необходимо выполнить следующий алгоритм, состоящий из двух этапов:

1. *Этап кластеризации F*:

1.1. выбирается множество объектов – F;

1.2. строится МС для множества F;

1.3. исходя из условий задачи задается $M = |F'|$;

1.4. задается пороговое значение меры схожести K. Объекты считаются взаимно схожими, если их мера схожести больше K;

1.5. выделяются группы взаимно схожих объектов. Для этого $\forall f_i$ рассчитывается L_i – количество f_j $(i \neq j)$: $\rho_{ij} \geq K$. Признаком уникальности f_i в F является условие $L_i = 0$.

2. *Этап формирования F'*:

2.1. выбирается случайным образом любой объект из каждой совокупности схожих между собой объектов и добавляется в F';

2.2. объекты, отобранные на шаге 1.4, включаются в F' в порядке убывания L_i, пока объем F' не достигнет M;

2.3. если $|F'| = M$, то алгоритм заканчивает свою работу и в качестве результата выдается множество F'. Иначе понижается значение K и повторяются пункты 1.2–2.2.

Разработанный алгоритм, применительно к сфере антивирусного эвристического анализа, может быть полезен на этапе формирования обучающего множества. Из множества вредоносных объектов можно сформировать множество файлов, где каждый отличался бы от остальных по своим злонамеренным действиям. Это позволит более качественно обучить эвристический классификатор и в большей степени обезопасить конечного пользователя.

Продемонстрируем работу алгоритма на следующем примере.

Дано множество F, которое включает в себя 10 объектов. На его основе необходимо сформировать F', такое, что $M = 6$. Матрица схожести для F представлена в табл. 4.

Таблица 4

Матрица схожести

Объект	Мера схожести для объекта №										
	L*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	*1*	1	0,714	**0,983**	0,615	0,040	0,040	0,577	0,588	0,009	0,040
2	*5*	**0,946**	1	**0,946**	**0,850**	0,019	0,019	**0,829**	**0,837**	0,004	0,019
3	*1*	**0,983**	0,715	1	0,615	0,040	0,040	0,577	0,588	0,009	0,040
4	0	0,198	0,198	0,198	1	0,013	0,013	0,146	0,153	0,003	0,013
5	*2*	0,302	0,230	0,304	0,043	1	**0,978**	0,010	0,017	0,095	**0,982**
6	*2*	0,302	0,230	0,305	0,043	**0,978**	1	0,010	0,016	0,095	**0,979**
7	0	0,009	0,009	0,009	0,008	0,002	0,002	1	0,016	0,002	0,002
8	0	0,043	0,041	0,043	0,041	0,012	0,012	0,050	1	0,002	0,012
9	0	0,066	0,042	0,066	0,059	0,079	0,079	0,002	0,006	1	0,079
10	*2*	0,302	0,230	0,305	0,043	**0,981**	**0,979**	0,010	0,016	0,095	1

* *L* – количество мер схожести больше *K* в ячейках по горизонтали.

Зададим $K = 0{,}8$. В табл. 4 выделены значения мер схожести больше K. Высчитаем для каждого объекта параметр L (см. п. 1.5 алгоритма). В столбце L табл. 4 курсивом выделено, для каких объектов имеются схожие с ними f_j.

Можно выделить две группы взаимно схожих объектов: 1-й и 3-й, а также 5-й, 6-й и 10-й файлы.

Из анализа табл. 4 также следует, что имеется объект 2, похожий на объекты 1, 4, 7 и 8, которые, в свою очередь, на него не похожи.

Из каждой группы остается только один объект (табл. 5).

Таблица 5

Матрица схожести после исключения схожих объектов

Объект	Мера схожести для объекта №							
	L	**1**	**2**	**4**	**7**	**8**	**9**	**10**
1	0	1	0,714	0,615	0,577	0,588	0,009	0,040
2	*4*	**0,946**	1	**0,850**	**0,829**	**0,837**	0,004	0,019
4	0	0,198	0,198	1	0,146	0,153	0,003	0,013
7	0	0,009	0,009	0,008	1	0,016	0,002	0,002
8	0	0,043	0,041	0,041	0,050	1	0,002	0,012
9	0	0,066	0,042	0,059	0,002	0,006	1	0,079
10	0	0,302	0,230	0,043	0,010	0,016	0,095	1

Исключив данный объект, получаем 6 максимально различающихся между собой объектов, которые и становятся искомым множеством F'.

Практическая проверка

На тестовом примере оценим, насколько корректно предложенный алгоритм разобьет исходное множество файлов на классы, из которых в дальнейшем будут взяты экземпляры для обучающего множества.

Кластеризатор выделил следующие группы файлов: {1, 3}, {5, 6, 10}, {{2, 1}, {2, 4}, {2, 7}, {2, 8}}, {9}, причем первые две группы взаимно схожих, а третья – нет. Посмотрев на функционал данных файлов, эксперт разделил их на следующие группы взаимно схожих объектов: {1, 2, 3}, {5, 6, 10}, {4}, {8}, {7}, {9}.

Эксперт составил 12 пар точно связанных между собой файлов (1 и 2, 2 и 1, 1 и 3, 3 и 1, ….) и 78 обязательно разделенных (1 и 5, 5 и 1, 1 и 6, 6 и 1, …). Кластеризатор выделил 12 пар точно связанных между собой файлов (1 и 3, 3 и 1, 5 и 6, 6 и 5, ….) и 78 обязательно разделенных (1 и 2, 1 и 4, 4 и 1, …). При этом кластеризатор согласился с экспертом относительно 9 точно связанных между собой пар. По результатам эксперимента была составлена таблица сопряженности (табл. 6) [9].

Таблица 6

Таблица сопряженности

Файлы	Связаны	Разделены	Сумма
Должны быть связаны	9	3	12
Должны быть разделены	3	75	78
Сумма	12	78	90

Индекс Ренда [10] (доля корректно связанных/несвязанных пар) для данного примера составляет (9 + 75) / 90 = 0,93.

Заключение

Кластеризация, будучи одной из важнейших задач машинного обучения, опирается, как правило, на метрику, т. е. на расстояние между объектами. Однако для ряда предметных областей необходима более гибкая величина, отражающая, насколько один объект является частью другого объекта. Для этого была предложена мера схожести, которая не является метрикой. На основании ее была построена матрица схожести, анализ которой позволяет кластеризовать объекты и отобрать максимально разнообразные для, например, дальнейшего обучения классификатора. Экспериментальная проверка показала, что разбиение экспертом множества на кластеры почти совпало с разбиением, произведенным предложенным кластеризатором. Индекс Ренда при этом составил 0,93.

СПИСОК ЛИТЕРАТУРЫ

1. *Kumar A., Kuppusamy K. S., Aghila G.* A learning model to detect maliciousness of portable executable using integrated feature set // Journal of King Saud University Computer and Information Sciences. King Saud University. 2017. URL: https://reader.elsevier.com/reader/sd/pii/S1319157817300149?token=A3C9C0EE7EC36B541 CE5E 33E71C3C5383BA752D2A973FBD698F04A54D940DAE62B00A1D293F4FAD7A2226FCCE2361DDA (дата обращения: 21.01.2019).

2. *Васильев Т. Р., Кокуев А. Г.* Контроль расхода нефтепродуктов на основе искусственной нейронной сети // Вестн. Астрахан. гос. техн. ун-та. Сер.: Управление, вычислительная техника и информатика. 2018. № 2. С. 43–52.

3. *Johari A., Navein C.* Hierarchical density-based clustering of malware behaviour // Journal of telecommunication electronic and computer engineering. 2017. № 2-10. P. 151–158.

4. *Казимиров Д. Ю., Исаченко А. С.* Формирование последовательности запуска в производство изделий одновременной кластеризацией по технологическим признакам и классам деталей // Вестн. Иркут. гос. техн. ун-та. 2016. № 7 (114). С. 24–36.

5. *Московкин В. М., Казимиру Э.* Матричная кластеризация как кластеризация матриц одинаковой размерности // Науч. вед. Белгор. гос. ун-та. Сер.: Экономика. Информатика. 2017. № 23 (272). С. 123–127.

6. *Russel S., Norvig P.* Artificial Intelligence: A Modern Approach. Prentice Hall, 2015. 1164 p.

7. *Демина Р. Ю., Ажмухамедов И. М.* Методика формирования обучающего множества при использовании статических антивирусных методов эвристического анализа // Инженер. вестн. Дона. 2015. № 3. URL: http://ivdon.ru/uploads/article/pdf/IVD_204_demina_azhmuhamedov.pdf_0b8ea4a2fc.pdf (дата обращения: 21.01.2019).

8. *Демина Р. Ю.* Особенности программной реализации алгоритмов методики формирования обучающего множества для бинарных классификаторов, используемых в антивирусном эвристическом статическом анализе // Вестн. Астрахан. гос. техн. ун-та. Сер.: Управление, вычислительная техника и информатика. 2017. № 2. С. 62–68.

9. *Потапов А. С.* Распознавание образов и машинное восприятие. СПб.: Политехника, 2007. 552 с.

10. *Флах П.* Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс, 2015. 400 с.

Статья поступила в редакцию 19.09.2019

ИНФОРМАЦИЯ ОБ АВТОРАХ

Демина Раиса Юрьевна – Россия, 414056, Астрахань; Астраханский государственный технический университет; аспирант кафедры информационной безопасности; raisapereverzeva@gmail.com.

Ажмухамедов Искандар Маратович – Россия, 414056, Астрахань; Астраханский государственный университет; д-р техн. наук; зав. кафедрой информационной безопасности; aim_agtu@mail.com.

INCREASING QUALITY OF CLASSIFYING OBJECTS USING NEW METRICS OF CLUSTERING

R. Yu. Demina[1], I. M. Azhmukhamedov[2]

[1] Astrakhan State Technical University, Astrakhan, Russian Federation

[2] Astrakhan State University, Astrakhan, Russian Federation

Abstract. The article touches upon one of the main problems of machine learning - clustering objects. It has been widely used in various subject areas: marketing, sociology, psychology, etc. Clusterization algorithms, as a rule, are based on a metric that reflects the distance between objects. However, in some cases it is not practical to use the distance between objects. In certain situations, it is possible to say that one object is similar to the other, the latter being not similar to the former. The original picture and its copy may serve as an example. For such cases, a measure of object similarity is proposed in the work, which shows how many features of one object are contained in another one. A similarity matrix is built on this measure, the analysis of which allows revealing clusters of mutually similar objects. When testing the proposed clustering method, the Rand index (the proportion of correctly connected or unrelated objects) made 0.93. There has been proposed an algorithm that allows to form a set of objects absolutely different from each other. A set of objects formed in this way can later become a learning set for classifiers and increase their fidelity in recognition.

Key words: clustering, metric, comparison, degree of likeliness, training set, object's features, Rand index.

For citation: Demina R. Yu., Azhmukhamedov I. M. Increasing quality of classifying objects using new metrics of clustering. *Vestnik of Astrakhan State Technical University. Series: Management, Computer Science and Informatics*. 2019;4:106-114. (In Russ.) DOI: 10.24143/2072-9502-2019-4-106-114.

REFERENCES

1. Kumar A., Kuppusamy K. S., Aghila G. A learning model to detect maliciousness of portable executable using integrated feature set. *Journal of King Saud University Computer and Information Sciences*. King Saud University, 2017. Available at: https://reader.elsevier.com/reader/sd/pii/S1319157817300149?token=A3C9C0EE7EC36B541CE5E33E71C3C5383BA752D2A973FBD698F04A54D940DAE62B00A1D293F4FAD7A2226FCCE2361DDA (accessed: 21.01.2019).

2. Vasil'ev T. R., Kokuev A. G. Kontrol' raskhoda nefteproduktov na osnove iskusstvennoi neironnoi seti [Control of consumption of petroleum products based on artificial neural network]. *Vestnik Astrakhanskogo gosudarstvennogo tekhnicheskogo universiteta. Seriia: Upravlenie, vychislitel'naia tekhnika i informatika*, 2018, no. 2, pp. 43-52.

3. Johari A., Navein C. Hierarchical density-based clustering of malware behaviour. *Journal of telecommunication electronic and computer engineering*, 2017, no. 2-10, pp. 151-158.

4. Kazimirov D. Iu., Isachenko A. S. Formirovanie posledovatel'nosti zapuska v proizvodstvo izdelii odnovremennoi klasterizatsiei po tekhnologicheskim priznakam i klassam detalei [Formation of sequential launching products by simultaneous clustering according to technological characteristics and classes of parts]. *Vestnik Irkutskogo gosudarstvennogo tekhnicheskogo universiteta*, 2016, no. 7 (114), pp. 24-36.

5. Moskovkin V. M., Kazimiru E. Matrichnaia klasterizatsiia kak klasterizatsiia matrits odinakovoi razmernosti [Matrix clustering as clustering matrices of similar nulity]. *Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta. Seriia: Ekonomika. Informatika*, 2017, no. 23 (272), pp. 123-127.

6. Russel S., Norvig P. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2015. 1164 p.

7. Demina R. Iu., Azhmukhamedov I. M. Metodika formirovaniia obuchaiushchego mnozhestva pri ispol'zovanii staticheskikh antivirusnykh metodov evristicheskogo analiza [Methods of forming training set when using static antivirus methods of heuristic analysis]. *Inzhenernyi vestnik Dona*, 2015, no. 3. Available at: http://ivdon.ru/ uploads/article/pdf/IVD_204_demina_ azhmuhamedov.pdf_0b8ea4a2fc.pdf (accessed: 21.01.2019).

8. Demina R. Iu. Osobennosti programmnoi realizatsii algoritmov metodiki formirovaniia obuchaiushchego mnozhestva dlia binarnykh klassifikatorov, ispol'zuemykh v antivirusnom evristicheskom staticheskom analize [Characteristics of programming algorithms of developing training set for binary classifiers used in antivirus heuristic static analysis]. *Vestnik Astrakhanskogo gosudarstvennogo tekhnicheskogo universiteta. Seriia: Upravlenie, vychislitel'naia tekhnika i informatika*, 2017, no. 2, pp. 62-68.

9. Potapov A. S. *Raspoznavanie obrazov i mashinnoe vospriiatie* [Pattern recognition and machine perception]. Saint-Petersburg, Politekhnika Publ., 2007. 552 p.

10. Flakh P. Mashinnoe obuchenie. *Nauka i iskusstvo postroeniia algoritmov, kotorye izvlekaiut znaniia iz dannykh* [Machine learning. Science and art of building algorithms extracting knowledge from data]. Moscow, DMK Press, 2015. 400 p.

The article submitted to the editors 19.09.2019

INFORMATION ABOUT THE AUTHORS

Demina Raisa Yurievna – Russia, 414056, Astrakhan; Astrakhan State Technical University; Postgraduate Student of the Department of Information Security; raisapereverzeva@gmail.com.

Azhmukhamedov Iskandar Maratovich – Russia, 414056, Astrakhan; Astrakhan State University; Doctor of Technical Sciences, Assistant Professor; Professor of the Department of Information Security; aim_agtu@mail.ru.