УДК 004.622
DOI: 10.26102/2310-6018/2022.38.3.013

Ансамблирование методов обнаружения выбросов при подготовке обучающей выборки данных

В.С. Дорофеев✉, Т.М. Волосатова

Московский государственный технический университет имени Н.Э. Баумана, Москва, Российская Федерация
do.wladimir@gmail.com✉

Резюме. Большинство методов машинного обучения показывают наибольшую эффективность при работе с данными, удовлетворяющими нормальному распределению. С другой стороны, обучающая выборка часто содержит «выбросы» различной природы, способные значительно снизить точность методов машинного обучения. Таким образом, в любой задаче машинного обучения возникает проблема обнаружения выбросов. В статье приведена классификация основных типов выбросов. Рассмотрены различные методы обнаружения одномерных выбросов: метод, использующий критерий Граббса; метод Z-оценки; метод надежной Z-оценки (RZ-оценки); метод межквартильного размаха (IQR); метод процентильного уплотнения (Winsorization). Выполнено сравнение методов обнаружения одномерных выбросов. Для автоматизированного обнаружения выбросов предложен ансамблевый метод, объединяющий различные методы обнаружения одномерных выбросов. Ансамблирование позволяет настроить автоматизированную процедуру обнаружения выбросов по правилу требуемой строгости. Предложенный метод применен для анализа и обнаружения выбросов в данных по продажам товаров в период акции в крупной розничной сети. Показана возможность применения ансамблирования методов обнаружения выбросов для стратификации обучающей выборки. При этом абсолютная и относительная ошибка прогнозирования итоговой модели была снижена на 5 % по сравнению с исходной.

Ключевые слова: выбросы, машинное обучение, обучающая выборка, ансамблирование, метод Z-оценки, метод межквартильного размаха.

Для цитирования: Дорофеев В.С., Волосатова Т.М. Ансамблирование методов обнаружения выбросов при подготовке обучающей выборки данных. *Моделирование, оптимизация и информационные технологии.* 2022;10(3). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=1210
DOI: 10.26102/2310-6018/2022.38.3.013

Ensemble methods for detecting outliers in the preparation of a training data set

V.S. Dorofeev✉, T.M. Volosatova

Bauman Moscow State Technical University, Moscow, Russian Federation
do.wladimir@gmail.com✉

Abstract. Most machine learning methods are most effective when working with data that satisfies a normal distribution. On the other hand, the training set often contains “outliers” of various nature, which can significantly reduce the accuracy of machine learning methods. Thus, in any machine learning task, there is a problem of detecting outliers. The article provides a classification of the main types of emissions. Various methods for detecting one-dimensional outliers are considered: the method using the Grubbs criterion; Z-score method; robust Z-score (RZ-score) method; interquartile range (IQR) method; Winsorization method. The methods for detecting one-dimensional outliers are compared. For the automated detection of outliers, an ensemble method has been proposed that combines various methods

for detecting one-dimensional outliers. The ensemble method helps to configure an automated outlier detection procedure according to the rule of the required severity. The suggested method is applied to analyze and detect outliers in data on sales of goods during the promotion in a large retail network. The applicability of using outlier detection method ensemble to stratification of the training sample is shown. At the same time, the absolute and relative forecasting error of the final model decreased by 5% compared to the initial one.

Keywords: outliers, machine learning, training sample, ensemble method, Z-score, interquartile range method.

For citation: Dorofeev V.S., Volosatova T.M. Ensemble methods for detecting outliers in the preparation of a training data set. *Modeling, Optimization and Information Technology*. 2022;10(3). Available from: https://moitvivt.ru/ru/journal/pdf?id=1210 DOI: 10.26102/2310-6018/2022.38.3.013 (In Russ.).

Введение

На сегодняшний день информация в цифровом виде представляет собой колоссальный массив данных. По оценкам экспертов, объем общемировых данных вырастет с 45 зеттабайт в 2019 году до 175 в 2025 1. Цифровые данные, в отличие от других видов экономических ресурсов, растут экспоненциально, и источники данных продолжают расширяться. При этом типичной задачей является поиск скрытых закономерностей в большом объеме данных. Так как человек не способен обрабатывать значительный объем поступающих данных за приемлемое время, для выполнения этой задачи широко используются методы машинного обучения.

Данные, используемые при обучении модели машинного обучения, называют обучающей выборкой. Формирование обучающей выборки имеет принципиально важное значение для успешного решения задач машинного обучения. Нередко основная сложность задачи машинного обучения 2 сводится к правильной подготовке обучающей выборки. При этом в массиве данных обучающей выборки может присутствовать ошибочная или сильно выбивающаяся из общего ряда значений информация об исследуемых объектах или процессах. Численно выделяющееся наблюдение, которое расходится с общей закономерностью в выборке, называют «выброс».

В некоторых классах задач поиск выбросов может быть основной целью, например, в задаче обнаружении мошенничества 3 или идентификации признаков вовлеченности кредитных организаций и их клиентов в сомнительные операции 4.

Однако в общем случае поиск выбросов является побочной целью, а выбросы являются результатом ошибок в данных. Выбросы могут мешать настройке модели 2 и привести к значительному снижению точности прогнозирования или классификации.

Таким образом, возникает задача обнаружения и фильтрации выбросов – обнаружения в обучающей выборке ограниченного количества атипичных значений.

Существующие подходы к решению задачи обнаружения выбросов можно разделить на три типа аналогично типам задач машинного обучения.

– Определение выбросов без априорной информации о данных. По сути, этот подход аналогичен кластеризации.

– Моделирование как «нормальных» данных, так и выбросов. Такой подход аналогичен классификации и требует данных, предварительно размеченных на классы.

– Моделирование либо только «нормальных» данных, либо только выбросов. Данный подход также требует размеченных данных, при этом модель может продолжать учиться по мере получения новых данных.

Анализируя результаты большого количества тщательно выполненных физических экспериментов 5, получен вывод, что наличие 5–10 % выбросов в реальных

статистических данных является скорее правилом, чем исключением. Так, анализ результатов внутренних сличений, выполненных во вторичном эталоне ВЭТ 1-5 ВНИИФТРИ (г. Иркутск), показал 6, что результаты, полученные на суточных интервалах за 90 суток, содержат в среднем 6 % выбросов. Анализ причин, порождающих выбросы, при этом не проводился.

Можно предположить, что данные, собранные в других областях деятельности, могут содержать значительно больший процент выбросов.

Анализируя причины, приводящие к отклонениям от параметрических моделей в условиях реальных экспериментов, можно составить следующую классификацию основных типов выбросов по их источнику:

– Ошибки ввода. Это ошибки, допущенные во время сбора, записи или ввода данных, вызванные человеческим фактором. Они соответствуют замещающей модели 7 загрязнения данных.

– Ошибки измерения. Ошибки, вызванные неисправностью измерительного прибора, либо ошибки округления. Соответствуют аддитивной модели загрязнения данных.

– Естественный выброс. Выбросы, не являющиеся ошибкой. Однако в данном случае принятая параметрическая модель может быть неадекватной рассматриваемой задаче.

Также выделим два следующих типа выбросов по размерности:

– Одномерные выбросы. Для обнаружения данного типа выбросов достаточно исследовать распределение одного признака.

– Многомерные выбросы. Для обнаружения данного типа выбросов необходимо исследовать распределение n-признаков в n-мерном пространстве.

В данной работе рассмотрена задача обнаружения одномерных выбросов в неразмеченном наборе данных. Планируется также и дальнейшее развитие работы – многомерный анализ.

В задачах машинного обучения наиболее применимы данные, удовлетворяющие нормальному распределению. Абсолютное большинство методов машинного обучения показывает наилучшую эффективность при использовании данных, удовлетворяющих нормальному распределению. Такие модели, как Байесовский алгоритм 8, Латентное размещение Дирихле, а также методы, основанные на различных модификациях линейной и логистической регрессии 9, явно исходят из предположения, что распределение обучающей выборки является одномерным или многомерным нормальным распределением. Кроме того, сигмовидные функции, лежащие в основе архитектуры многих нейросетевых моделей, показывают лучшие результаты при работе с нормально распределенными данными 10.

Таким образом, следует тщательно изучить данные, проверить базовые распределения для каждого непрерывного признака и очистить обучающую выборку от выбросов, прежде чем подходить к работе с моделью машинного обучения.

Методы обнаружения выбросов

Рассмотрим следующие методы обнаружения выбросов:

1) метод на основе визуализации;
2) метод, использующий критерий Граббса;
3) метод Z-оценки;
4) метод надежной Z-оценки (RZ-оценки);
5) метод межквартильного размаха (IQR);
6) метод процентильного уплотнения (Winsorization).

Визуализация данных. Визуализация данных является одним из наиболее эффективных и очевидных подходов к обнаружению выбросов 11. В данном случае решение, является ли значение выбросом, принимает человек. С присутствием человека в качестве лица, принимающего решение, связаны и другие недостатки данного подхода:

– данный подход невозможно автоматизировать;

– сложности при работе с большим объемом данных;

– отсутствие математического обоснования для принятия решения.

Критерий Граббса — статистический тест, используемый для определения выбросов в одномерном наборе данных, подчиняющихся нормальному закону распределения. Был предложен в 1950 году Франком Граббсом 12. Критерий Граббса определен для следующих гипотез:

1. H0 − в наборе данных нет выбросов;

2. Ha − в наборе данных присутствует, как минимум, один выброс.

Пусть имеется набор данных $X_1,..,X_{|X|}$. Тогда критерий Граббса для значения X_i рассчитывается как

$$G = \frac{\max\limits_{i=1,..,|X|} |X_i - \overline{X}|}{s},$$

где $\overline{X}$ и s означают выборочное среднее и среднеквадратичное отклонение соответственно. При этом критическое значение критерия Граббса рассчитывается как

$$G_{крит} = \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{(t_{\frac{a}{2N},N-2})^2}{N-2+(t_{\frac{a}{2N},N-2})^2}},$$

где $t_{\frac{a}{2N},N-2}$ означает максимальное критическое значение распределения Стьюдента с N − 2 степенями свободы и уровнем значимости α/(2N).

Значение критерия Граббса показывает максимальное абсолютное отклонение от выборочного среднего в единицах среднеквадратичного отклонения. Если вычисленное значение G больше критического $G_{\text{крит}}$ 13, можно отклонить нулевую гипотезу H0 и сделать вывод Ha, что значение X_i является выбросом.

Метод Z-оценки. Z-оценка базируется на следующем эмпирическом правиле из предположения о нормальности распределения анализируемого набора данных 14:

– 68,26 % данных будут находиться в пределах 1 стандартного отклонения от среднего ($\mu \pm 1\sigma$);

– 95,44 % в пределах 2σ от среднего ($\mu \pm 2\sigma$);

– 99,7 % в пределах 3σ от среднего ($\mu \pm 3\sigma$);

– 95 % в пределах ($\mu \pm 1{,}96\sigma$);

– 99 % в пределах ($\mu \pm 2{,}75\sigma$).

Z-оценка рассчитывается по формуле

$$Z = \frac{X_i - \mu}{\sigma},$$

где X_i– анализируемое значение; μ – среднее значение для набора данных; σ – стандартное отклонение для набора данных [15].

Если величина Z-оценки значения X_i в точке больше трех (поскольку она покрывает 99,7 % площади под кривой при нормальном распределении), это означает,

что точка сильно отличается от других значений. Данное значение воспринимается как выброс.

Z-оценка имеет ряд ограничений. Во-первых, среднее и среднеквадратическое отклонение также изменяется под воздействием аномальных значений, что может вызвать маскировку выбросов. Во-вторых, максимально возможная величина Z-оценки зависит от размера выборки.

Метод надежной Z-оценки (RZ-оценки) также называют методом среднего абсолютного отклонения. Он является модифицированным методом RZ-оценки с некоторыми изменениями параметров 16. Поскольку на среднее значение и стандартное отклонение, используемые в методе Z-оценки, сильно влияют выбросы, в качестве альтернативы этих параметров можно использовать медиану и абсолютное отклонение от медианы.

С помощью метода RZ-оценки можно с большей достоверностью обнаружить выбросы даже при наличии их значительного числа в данных, используемых для вычисления медианного и медианного абсолютного отклонения:

$$RZ = \frac{0{,}6745 \cdot (X_i - Med)}{MAD},$$

где X_i – анализируемое значение; Med − медиана для набора данных; MAD – абсолютное отклонение от медианы.

Метод также базируется на предположении, что набор данных следует стандартному нормальному распределению. В этом случае абсолютное отклонение от медианы будет сходиться к медиане полунормального распределения, которая является 75-м процентилем нормального распределения, при этом $N(0{,}75) = 0{,}6745$.

Если RZ-оценка значения в точке больше трех, аналогично методу Z-оценки, данное значение воспринимается как выброс.

Метод межквартильного размаха (метод IQR). В методе IQR (метод межквартильного размаха, англ. InterQuartile Range) для обнаружения выбросов используется 17 межквартильный размах (IQR), основанный на понятии квартилей:

– $Q1$ представляет 1-й квартиль / 25-й процентиль данных;

– $Q2$ представляет 2-й квартиль / медиана / 50-й процентиль данных;

– $Q3$ представляет 3-й квартиль / 75-й процентиль данных.

Величина IQR представляет собой расстояние между 1-м и 3-м 18 квартилями (25-м и 75-м процентилями):

$$IQR = N(0{,}75) - N(0{,}25).$$

Таким образом, в данном методе IQR используется в качестве меры изменения набора данных. Величина $(Q1 - 1{,}5 \cdot IQR)$ представляет наименьшее значение в наборе данных, а $(Q3 + 1{,}5 \cdot IQR)$ – наибольшее значение. Любое значение, выходящее за пределы диапазона от $(-1{,}5 \cdot IQR)$ до $(1{,}5 \cdot IQR)$, рассматривается как выбросы.

Метод процентильного уплотнения также называют методом Winsorization. Данный метод аналогичен методу IQR: Величина $N(0{,}01)$ представляет наименьшее значение в наборе данных, а $N(0{,}99)$ – наибольшее значение 19. Если анализируемое значение превышает значение 99-го процентиля и ниже 1-го процентиля, данные значения рассматриваются как выбросы.

Ассемблирование методов обнаружения выбросов

В большинстве прикладных задач, в том числе задаче обнаружения выбросов, неочевидно, какой метод подходит наилучшим образом. В таких случаях хорошо зарекомендовал себя подход с использованием ансамбля методов 19, когда результаты сразу нескольких из них участвуют в формировании конечного результата. Ансамбль методов (алгоритмов) – метод, который использует несколько алгоритмов с целью получения лучшей эффективности прогнозирования, чем можно было бы получить от каждого алгоритма по отдельности.

Рассмотрим формальную постановку задачи классификации.

Пусть X – множество описаний объектов, Y– множество классов. Существует неизвестная целевая зависимость – отображение $f: X \to Y$.

Требуется построить алгоритм $a: X \to Y$, аппроксимирующий целевую зависимость на всем множестве X с требуемой точностью.

Пусть имеется M классификаторов $b_1, b_2, .., b_M$, где $b_i: X \to Y, b_i \in B, i \in (1..M)$, позволяющих разделить множество объектов X на K классов.

Тогда возможно построить новый классификатор на основе данных с помощью следующих методов:

– простое голосование, когда

$$b(x) = \frac{1}{M}\sum_{i=1}^{M} b_i(x);$$

– взвешенное голосование, в котором

$$b(x) = \frac{1}{M}\sum_{i=1}^{M} \omega_i b_i(x), \quad \sum_{i=1}^{M} \omega_i = 1, \omega_i > 0$$

– смесь экспертов, при котором

$$b(x) = \frac{1}{M}\sum_{i=1}^{M} \omega_i(x) b_i(x), \quad \sum_{i=1}^{M} \omega_i(x) = 1, \forall x \in X.$$

Очевидно, что простое голосование является частным случаем взвешенного голосования, а взвешенное голосование является частным случаем смеси экспертов 21.

Вероятность правильного решения задачи при помощи ансамбля алгоритмов можно определить с использованием теоремы Кондорсе 22 о присяжных.

Теорема Кондорсе. Пусть M – количество классификаторов, m – *степень строгости* (минимально необходимое число классификаторов, которое должно детектировать значение как выброс для идентификации значения в качестве выброса). В случае минимального большинства, простым голосованием степень строгости m определяет формула

$$m = \left[\frac{M}{2}\right] + 1.$$

Вероятность правильного решения всего ансамбля классификаторов равна

$$R = \sum_{i=m}^{M} C_M^i p^i (1-p)^{M-i},$$

где p – вероятность правильного решения одного классификатора.

При p > 0,5 наблюдается эффект Кондорсе – стремление R к 1 при увеличении числа классификаторов 23. Для p < 0,5 мы наблюдаем обратный эффект – уменьшение R до 0.

Исходим из двух следующих предположений.

1) Характер распределения значений в наборе данных должен стремиться к нормальному распределению как наиболее оптимальному для использования при обучении большинства моделей машинного обучения.

2) Вероятность корректного обнаружения выбросов каждым из рассматриваемых методов больше 0,5.

Тогда возможно построить ансамбль на основе методов обнаружения выбросов. При этом, если число методов нечетно, то возможно построить новый классификатор на основе данных простого голосования.

Также при построении ансамбля возможно выбирать степень строгости m при определении выбросов: чем большее количество методов m должно детектировать значение как выброс, тем меньше значений будет идентифицировано как выброс.

Для автоматизации процедуры обнаружения выбросов может быть заранее определена степень строгости m при определении выбросов в типичных для решаемой задачи наборах данных.

Вычислительные эксперименты

Данные о продажах в период акции (скидки) в российской крупной розничной сети. Рассмотрены продажи более 2500 товаров в 700 магазинах компании по всей территории России, на которые распространялось действие скидок. Набор данных содержит выбросы, предположительно вызванные оптовыми закупками в розничных магазинах, либо ошибки учета товарного запаса и продаж. Предложенные критерии были применены для анализа и обнаружения выбросов в рассматриваемых данных. Результаты исследования представлены в Таблице 1.

Таблица 1 – Количество выбросов, обнаруженных с использованием различных критериев
Table 1 – Number of outliers detected using different criteria

Критерий	Скорость продажи до акции	Скорость продажи на 7 день акции	Коэффициент увеличения продаж
Граббс	3375	3717	15
Winsorization	5294	5479	3697
Z-оценка	10241	9406	22
RZ-оценка	94771	63203	67440
IQR	75330	63207	32436
Всего значений	460139	460139	460139

Наиболее чувствительными критериями обнаружения выбросов показали себя метод IQR 17 и метод надежной Z-оценки 16. В основе данных методов лежит использование медианы рассматриваемого набора данных.

Критерий Граббса 12 и метод Z-оценки 15 оказались менее чувствительными. В основе данных методов лежит среднее значение и стандартное отклонение, сильно подверженные влиянию величины и количества выбросов в наборе. В случае набора данных, при получении которого имело место деление на близкое к 0 числу (коэффициент увеличения продаж), данные методы детектировали как выбросы минимальное число значений, имеющих наибольшую величину в наборе.

Проценты значений, детектированных как выброс при различных значениях строгости m, представлены в Таблице 2.

Таблица 2 – Проценты значений, детектированных как выброс
Table 2 - Percentage of values detected as an outlier

Степень строгости m	Скорость продажи до акции	Скорость продажи на 7 день акции	Коэффициент увеличения продаж
1	0,170	0,113	0,121
2	0,135	0,113	0,058
3	0,018	0,017	0,007
4	0,009	0,010	0,00003
5	0,006	0,007	0,00002
Проценты значений, не идентифицированных как выброс	0,829	0,886	0,878

При этом остается возможность выбирать степень строгости m при определении выбросов в зависимости от применяемого в дальнейшем метода машинного обучения: чем более метод чувствителен к выбросам, тем меньшее количество методов m должно детектировать значение как выброс для исключения его из ОВ.

Задача прогнозирования целевого признака. В качестве прогностической модели был использован метод машинного обучения CatBoost как один из наиболее перспективных 24 при решении задачи прогнозирования целевого признака 25. Модель была обучена на обучающей выборке, отфильтрованной от выбросов с различной степенью строгости.

Для оценки качества модели прогнозирования в качестве метрики использована среднеквадратичная ошибка

$$RMSE = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)}{n}},$$

где y_i – реальное значение целевого признака; $\hat{y}_i$ – спрогнозированное значение целевого признака; n – размер выборки. Данный метод 24 дает информацию о размере ошибки, в то же время накладывая штраф на ее величину.

Полученные результаты приведены в Таблице 3.

Таблица 3 – Оценка качества модели CatBoost, обученной на отфильтрованной обучающей выборке
Table 3 – Quality evaluation of CatBoost model trained on the filtered training set

Степень строгости m	Процент использованных значений обучающей выборки	RMSE, обучающая выборка	RMSE, тестовая выборка	Абсолютная ошибка, шт	Относительная ошибка, шт
Без фильтрации	1	0,959	1,022	173910	0,504
5	0,994	0,804	0,925	158343	0,489
4	0,9905	0,784	0,837	154722	0,503
3	0,9815	0,752	0,805	148347	0,509
2	0,8645	0,528	0,566	100565	0,592
1	0,8295	0,428	0,448	77367	0,649

Из представленных результатов видно, что среднеквадратичная ошибка для модели уменьшается с повышением степени строгости. Главным образом это связано с тем, что из выборки исключаются выбросы с большой амплитудой, вносящие наибольший вклад в погрешность при прогнозировании выбросов. При некотором пороговом значении строгости наблюдается резкий рост относительной ошибки, что связано с тем, что из обучающей выборки удаляется слишком большой процент значений. Для парирования данного эффекта существует несколько подходов:

– заменить значения выброса другой величиной, не являющейся выбросом, например медианным значением для выборки;

– рассматривать значения, классифицированные как выбросы, отдельно от основного объема выборки.

В рассматриваемом вычислительном эксперименте был выбран последний из подходов. Примем степень строгости $m = 3$, что равнозначно большинству при простом голосовании. При значении строгости $m = 3$ не наблюдается роста относительной ошибки прогнозирования модели, обученной на отфильтрованной обучающей выборке. Вынесем значения, детектированные как выбросы, в отдельную обучающую выборку и обучим на ней вторую модель. Таким образом, данные не будут потеряны. Для определенности обозначим ситуацию, в которой рассматриваются только детектированные как выбросы данные отрицательной степенью строгости $m = -3$. Результаты обучения модели приведены в Таблице 4.

Таблица 4 – Оценка качества модели CatBoost, обученной на отфильтрованной обучающей выборке
Table 4 – Quality evaluation of CatBoost model trained on the filtered training set

Степень строгости m	Процент использованных значений обучающей выборки	*RMSE*, обучающая выборка	*RMSE*, тестовая выборка	Абсолютная ошибка, шт	Относительная ошибка, шт
3	0,9815	0,752	0,805	148347	0,509
-3	0,0185	3,793	5,973	16450	0,309
Итоговое значение	1			164797	0,478

Из представленных результатов следует, что итоговая абсолютная ошибка прогнозирования в количестве товаров значительно снизилась, как и относительная ошибка – более чем на 5 % по сравнению с прогнозом по нестратифицированной обучающей выборке.

Заключение

Рассмотрены различные методы обнаружения одномерных выбросов: критерий Граббса, метод Z-оценки, метод надежной Z-оценки, метод межквартального размаха и метод процентильного уплотнения. Эти методы применены для анализа и обнаружения выбросов в данных по продажам товаров в период акции в крупной розничной сети. В ходе вычислительного эксперимента выявлено, что на исследуемом наборе данных методы определяют как выбросы различное число значений. Наиболее чувствительными критериями обнаружения выбросов показали себя методы межквартального размаха и метод надежной Z-оценки. Критерий Граббса и метод Z-оценки оказались менее чувствительными.

Для автоматизированного обнаружения выбросов предложен ансамблевый метод, объединяющий различные методы обнаружения одномерных выбросов. При определении выбросов простым большинством было детектировано как выброс 1-2 % от всех значений в рассматриваемых наборах данных. В то же время было детектировано как выброс хотя бы одним из рассматриваемых методов до 15-20 % всех значений.

Таким образом, ансамблирование позволяет настроить автоматизированную процедуру обнаружения выбросов по правилу требуемой строгости в зависимости от специфики задачи.

Показана возможность применения ансамблирования методов обнаружения выбросов для стратификации обучающей выборки. При этом абсолютная и относительная ошибка прогнозирования итоговой модели может быть снижена на 5 % по сравнению с исходной.

Однако на основе решения одной конкретной задачи нельзя делать далеко идущие выводы. Предполагается дальнейшее расширение исследования в последующих работах. Планируется также и дальнейшее развитие работы – многомерный анализ, методы которого будут рассмотрены в будущей работе.

СПИСОК ИСТОЧНИКОВ

1. Reinsel D., Gantz J., Rydning J. *The Digital of the World – From Edge to Core.* IDC White Paper; 2018. Доступно по: https://www.seagate.com/ru/ru/our–story/data–age–2025/ (дата обращения: 17.05.2021).
2. Парасич А.В., Парасич В.А., Парасич И.В. Формирование обучающей выборки в задачах машинного обучения. Обзор. *Информационно-управляющие системы.* 2021;4(113):61–68.
3. Якимова В.А. Возможности и перспективы использования цифровых технологий в аудиторской деятельности. *Вестник Санкт-Петербургского университета. Экономика.* 2020;2:287–318.
4. Бекетнова Ю.М. Сравнительный анализ методов машинного обучения при идентификации признаков вовлеченности кредитных организаций и их клиентов в сомнительные операции. *Финансы: теория и практика.* 2021;5:186–199.
5. Shulenin V.P. Robust Alternatives to the Standard Deviation in Processing of Physics Experimental Data. *Russian Physics Journal*. 2016:59(6):824–832.
6. Серышева И.А. Фильтрация выбросов в задачах статической и динамической обработки данных в эталонах времени и частоты. *Вестник Иркутского государственного технического университета.* 2018;22(10):67–77.
7. Горяинов В.Б., Горяинова Е.Р. Влияние аномальных наблюдений на оценку наименьших квадратов параметра авторегрессионного уравнения со случайным коэффициентом. *Вестник МГТУ им. Н.Э. Баумана. Сер. Естественные науки.* 2016;2:16–24. DOI: 10.18698/1812-3368-2016-2-16-24.
8. Piryonesi S. Madeh, El-Diraby, Tamer E. Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. *Journal of Transportation Engineering, Part B: Pavements*. 2020:146–148.
9. David A. Freedman. *Statistical Models: Theory and Practice.* Cambridge University Press; 2009. 442 p.
10. Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Séverine Dubuisson, Isabelle Bloch. TRADI: Tracking deep neural network weight distributions. *16th European Conference on Computer Vision.* 2020:1–27.
11. Лежебоков А.А., Кулиев Э.В. Технологии визуализации для прикладных задач интеллектуального анализа данных. *Известия КБНЦ РАН*. 2019;4(90):14–23.

12. Житный М.В., Девяткина Т.Ю., Хубларова Т.С., Прохватова И.С. Методика экспериментального моделирования ударного воздействия имитаторов частиц космического мусора на солнечные элементы космического аппарата. *Известия ТулГУ. Технические науки.* 2020;5:32–40.
13. Ширяева Л.К., Репина Е.Г. О некоторых свойствах симметричной копулы Граббса. *Вестн. Сам. гос. техн. ун-та. Сер. Физ.-мат. Науки*. 2018;22(4):714–734. DOI: 10.14498/vsgtu1640.
14. McLeod S.A. Z-score: definition, calculation and interpretation. Simply Psychology; 2019. Доступно по: https://www.simplypsychology.org/z-score.html (дата обращения 17.05.2021).
15. Sapoetra D.B., Basuki R. Effect of service quality, religiosity, relationship closeness, and customer trust on customer satisfaction and loyalty at Bank Jatim Syariah. *RJOAS*. 2019;3:200–219.
16. Nurunnabi A., West G., Belton D. Robust Outlier Detection and Saliency Features Estimation in Point Cloud Data. *2013 International Conference on Computer and Robot Vision*. 2013:98–105.
17. Выходцев Н.А. Использование искусственного интеллекта для оценки стоимости недвижимого имущества. *Доклады ТУСУР*. 2021;1:68–72.
18. Chernov G. How to learn to defeat noisy robot in rock-paper-scissors game: an exploratory study. *Экономический журнал ВШЭ*. 2020;4:503–538.
19. Евсеева С.А. Исследование эффективности процедур коллективного вывода при решении задачи классификации. *Актуальные проблемы авиации и космонавтики*. 2019;2:41–43.
20. Lee B.K., Lessler J., Stuart E.A. Weight Trimming and Propensity Score Weighting. *PLoS ONE*. 2011;6(3). DOI: 10.1371/journal.pone.0018174.
21. Микрюков А.А., Бабаш А.В., Сизов В.А. Классификация событий в системах обеспечения информационной безопасности на основе нейросетевых технологий. *Открытое образование.* 2019;1:57–63.
22. Протасов В.И., Потапова З.Е. Методика кардинального снижения вероятности принятия ошибочных решений в системах коллективного интеллекта. *Современные информационные технологии и ИТ-образование*. 2019;3:588–601.
23. Baharad E., Goldberger J., Koppel M., Nitzan S. Beyond Con-dorcet: optimal aggregation rules using voting records. *Theory and Decision.* 2012;72(1):113–130.
24. Дорофеев В.С., Волосатова Т.М. Алгоритм подготовки обучающей выборки в задаче прогнозирования спроса. *Математические методы в технологиях и технике*. 2021;2:64–68.
25. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*. 2018:6637–6647.

REFERENCES

1. Reinsel D., Gantz J., Rydning J. The Digital of the World – From Edge to Core. *IDC White Paper*. 2018. Available by: https://www.seagate.com/ru/ru/our–story/data–age–2025/ (accessed on 17.05.2021).
2. Parasich A.V., Parasich V.A., Parasich I.V. Formirovanie obuchajushhej vyborki v zadachah mashinnogo obuchenija. Obzor. *Informacionno-upravljajushhie sistemy.* 2021;4(113):61–68. (In Russ.).

3. Jakimova V.A. Vozmozhnosti i perspektivy ispol'zovanija cifrovyh tehnologij v auditorskoj dejatel'nosti. *Vestnik Sankt-Peterburgskogo universiteta. Jekonomika*. 2020;2:287–318. (In Russ.).
4. Beketnova Ju.M. Sravnitel'nyj analiz metodov mashinnogo obuchenija pri identifikacii priznakov vovlechennosti kreditnyh organizacij i ih klientov v somnitel'nye operacii. *Finansy: teorija i praktika.* 2021;5:186–199. (In Russ.).
5. Shulenin V.P. Robust Alternatives to the Standard Deviation in Processing of Physics Experimental Data. *Russian Physics Journal*. 2016;59(6):824–832.
6. Serysheva I.A. Fil'tracija vybrosov v zadachah staticheskoj i dinamicheskoj obrabotki dannyh v jetalonah vremeni i chastoty. *Vestnik Irkutskogo gosudarstven-nogo tehnicheskogo universiteta.* 2018;22(10):67–77. (In Russ.).
7. Gorjainov V.B., Gorjainova E.R. Vlijanie anomal'nyh nabljudenij na ocenku naimen'shih kvadratov parametra avtoregressionnogo uravnenija so sluchajnym kojefficietom. *Vestnik MGTU im. N.Je. Baumana. Ser. Estestvennye nauki.* 2016;2:16–24. DOI: 10.18698/1812-3368-2016-2-16-24. (In Russ.).
8. Piryonesi S. Madeh, El-Diraby, Tamer E. Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. *Journal of Transportation Engineering, Part B: Pavements.* 2020:146–148.
9. David A. Freedman. Statistical Models: Theory and Practice. *Cambridge University Press;* 2009. 442 p.
10. Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Séverine Dubuisson, Isabelle Bloch. TRADI: Tracking deep neural network weight distributions. *16th European Conference on Computer Vision.* 2020:1–27.
11. Lezhebokov A.A., Kuliev Je.V. Tehnologii vizualizacii dlja prikladnyh zadach intellektual'nogo analiza dannyh. *Izvestija KBNC RAN.* 2019;4(90):14–23. (In Russ.).
12. Zhitnyj M.V., Devjatkina T.Ju., Hublarova T.S., Prohvatova I.S. Metodika jeksperimental'nogo modelirovanija udarnogo vozdejstvija imitatorov chastic kosmi-cheskogo musora na solnechnye jelementy kosmicheskogo apparata. *Izvestija TulGU. Tehnicheskie nauki.* 2020;5:32–40. (In Russ.).
13. Shirjaeva L.K., Repina E.G. O nekotoryh svojstvah simmetrichnoj kopuly Grabbsa. *Vestn. Sam. gos. tehn. un-ta. Ser. Fiz.-mat. Nauki.* 2018;22(4):714–734. DOI: 10.14498/vsgtu1640. (In Russ.).
14. McLeod S.A. Z-score: definition, calculation and interpretation. *Simply Psychology*; 2019. Available by: https://www.simplypsychology.org/z-score.html (accessed on 17.05.2021).
15. Sapoetra, D.B., Basuki, R. Effect of service quality, religiosity, relationship closeness, and customer trust on customer satisfaction and loyalty at Bank Jatim Syariah. *RJOAS.* 2019;3:200–219.
16. Nurunnabi A., West G., Belton D. Robust Outlier Detection and Saliency Features Estimation in Point Cloud Data. *2013 International Conference on Computer and Robot Vision.* 2013:98–105.
17. Vyhodcev N.A. Ispol'zovanie iskusstvennogo intellekta dlja ocenki stoimosti nedvizhimogo imushhestva. *Doklady TUSUR.* 2021;1:68–72. (In Russ.).
18. Chernov G. How to learn to defeat noisy robot in rock-paper-scissors game: an ex-ploratory study. *Jekonomicheskij zhurnal VShJe.* 2020;4:503–538.
19. Evseeva S.A. Issledovanie jeffektivnosti procedur kollektivnogo vyvoda pri reshenii zadachi klassifikacii. *Aktual'nye problemy aviacii i kosmonavtiki.* 2019;2:41–43. (In Russ.).
20. Lee B.K., Lessler J., Stuart E.A. Weight Trimming and Propensity Score Weighting. *PLoS ONE.* 2011;6(3). DOI: 10.1371/journal.pone.0018174.

21. Mikrjukov A.A., Babash A.V., Sizov V.A. Klassifikacija sobytij v sistemah obespechenija informacionnoj bezopasnosti na osnove nejrosetevyh tehnologij. *Otkrytoe obrazovanie.* 2019;1:57–63. (In Russ.).
22. Protasov V.I., Potapova Z.E. Metodika kardinal'nogo snizhenija verojatnosti prinjatija oshibochnyh reshenij v sistemah kollektivnogo intellekta. *Sovremennye informacionnye tehnologii i IT-obrazovanie.* 2019;3:588–601. (In Russ.).
23. Baharad E., Goldberger J., Koppel M., Nitzan S. Beyond Condorcet: optimal aggre-gation rules using voting records. *Theory and Decision.* 2012;72(1):113–130.
24. Dorofeev V.S., Volosatova T.M. Algoritm podgotovki obuchajushhej vyborki v zadache prognozirovanija sprosa. *Matematicheskie metody v tehnologijah i tehnike.* 2021;2:64–68. (In Russ.).
25. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: unbi-ased boosting with categorical features. *Advances in Neural Information Processing Systems.* 2018:6637–6647.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Дорофеев Владимир Сергеевич, аспирант Московского государственного технического университета имени Н.Э. Баумана, Москва, Российская Федерация.
e-mail: do.wladimir@gmail.com
SPIN-код: 4400-6513

Vladimir Sergeevich Dorofeev, Postgraduate Student, Bauman Moscow State Technical University, Moscow, Russian Federation.

Волосатова Тамара Михайловна, кандидат технических наук, доцент Московского государственного технического университета имени Н.Э. Баумана, Москва, Российская Федерация.
e-mail: tamaravol@gmail.com
SPIN-код: 6758-1172

Tamara Mikhailovna Volosatova, Candidate of Technical Sciences, Associate Professor at Bauman Moscow State Technical University, Moscow, Russian Federation.

Статья поступила в редакцию 11.07.2022; одобрена после рецензирования 25.07.2022; принята к публикации 16.09.2022.

The article was submitted 11.07.2022; approved after reviewing 25.07.2022; accepted for publication 16.09.2022.