УДК 004.93
https://doi.org/10.37661/1816-0301-2021-18-3-83-96

Оригинальная статья
Original Paper

Нормализация данных в машинном обучении

В. В. Старовойтов✉, Ю. И. Голуб

Объединенный институт проблем информатики
Национальной академии наук Беларуси,
ул. Сурганова, 6, Минск, 220012, Беларусь
✉*E-mail: valerys@newman.bas-net.by*

Аннотация. В задачах машинного обучения исходные данные часто заданы в разных единицах измерения и типах шкал. Такие данные следует преобразовывать в единое представление путем их нормализации или стандартизации. В работе показана разница между этими операциями. Систематизированы основные типы шкал, операции над данными, представленными в этих шкалах, и основные варианты нормализации функций. Предложена новая шкала частей и приведены примеры использования нормализации данных для их более корректного анализа.

На сегодняшний день универсального метода нормализации данных, превосходящего другие методы, не существует, но нормализация исходных данных позволяет повысить точность их классификации. Кластеризацию данных методами, использующими функции расстояния, лучше выполнять после преобразования всех признаков в единую шкалу.

Результаты классификации и кластеризации разными методами можно сравнивать различными оценочными функциями, которые зачастую имеют разные диапазоны значений. Для выбора наиболее точной функции можно выполнить нормализацию нескольких из них и сравнить оценки в единой шкале.

Правила разделения признаков древовидных классификаторов инвариантны к шкалам количественных признаков. Они используют только операцию сравнения. Возможно, благодаря этому свойству классификатор типа «случайный лес» в результате многочисленных экспериментов признан одним из лучших при анализе данных разной природы.

Ключевые слова: классификация объектов, кластеризация, нормализация данных, нормализация функций, сигмоида, гиперболический тангенс, случайный лес

Благодарности. Работа частично выполнена в рамках проектов БРФФИ Ф20РА–014 и Ф21ПАКГ–001.

Для цитирования. Старовойтов, В. В. Нормализация данных в машинном обучении / В. В. Старовойтов, Ю. И. Голуб // Информатика. – 2021. − Т. 18, № 3. – С. 83–96. https://doi.org/10.37661/1816-0301-2021-18-3-83-96

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 02.07.2021
Подписана в печать | Accepted 11.08.2021
Опубликована | Published 29.09.2021

Data normalization in machine learning

Valery V. Starovoitov✉, Yuliya I. Golub

The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus,
st. Surganova, 6, Minsk, 220012, Belarus
✉*E-mail: valerys@newman.bas-net.by*

Abstract. In machine learning, the input data is often given in different dimensions. As a result of the scientific papers review, it is shown that the initial data described in different types of scales and units of measurement should be converted into a single representation by normalization or standardization. The difference between these operations is shown. The paper systematizes the basic operations presented in these scales, as well as the main variants of the function normalization. A new scale of parts is suggested and examples of the data normalization for correct analysis are given. Analysis of publications has shown that there is no universal method of data normalization, but normalization of the initial data makes it possible to increase the accuracy of their classification. It is better to perform data clustering by methods using distance functions after converting all features into a single scale. The results of classification and clustering by different methods can be compared with different scoring functions, which often have different ranges of values. To select the most accurate function, it is reasonable to normalize several functions and to compare their estimates on a single scale. The rules for separating features of tree-like classifiers are invariant to scales of quantitative features. Only comparison operation is used. Perhaps due to this property, the random forest classifier, as a result of numerous experiments, is recognized as one of the best classifiers in the analysis of data of different nature.

Keywords: object classification, clustering, data normalization, function normalization, sigmoid, hyperbolic tangent, random forest

Acknowledgements. This work was partially performed within the framework of the BRFFR projects F20RA–014 and F21PAKG–001.

For citation. Starovoitov V. V., Golub Y. I. Data normalization in machine learning. *Informatics*, 2021, vol. 18, no. 3, pp. 83−96 (In Russ.). https://doi.org/10.37661/1816-0301-2021-18-3-83-96

Conflict of interest. The authors declare of no conflict of interest.

Введение. Нормализация данных в машинном обучении – это метод предварительной обработки, при котором данные преобразуются, чтобы обеспечить равный вклад каждого показателя [1, 2]. Успех алгоритмов машинного обучения зависит от точности описания данных для получения обобщенной прогнозной модели проблемы классификации [3]. Важность нормализации данных для улучшения описания и повышения точности алгоритмов машинного обучения была отмечена многими исследователями [4]. Для применения алгоритмов, использующих некую метрику или сравнение данных разных типов, предварительно требуется представление данных в одной шкале измерения.

В литературе часто путают понятия нормализации и стандартизации данных (см., например, [1–4]). Цель настоящей публикации – уточнить эти понятия, показать разницу между ними и продемонстрировать, как их следует применять в области машинного обучения.

Шкалы описания данных. В машинном обучении данные представляют собой признаки, описывающие некоторые объекты, понятия или события. Признаки могут быть записаны в разных шкалах: категориальных (неметрических) и количественных (численных). Такая классификация шкал была предложена Стивенсом [5]. Она представлена первыми четырьмя типами шкал в табл. 1 и подвергалась множественной критике [6]. Общепризнанной классификации шкал данных не существует, предлагаются и другие варианты этой классификации [7]. Основная идея любой классификации шкал заключается в группировании однотипно описываемых данных и определении допустимых для каждой группы операций над данными одного типа.

Таблица 1. Дополненная классификация шкал Стивенса

Table 1. Augmented classification of Stevens scales

Шкала *Scale*	Свойство *Property*	Единицы измерения *Units measurements*	Матем. операции *Mathem. operations*	Доп. операции *Add. operations*	Центральная тенденция *Central trend*
Номинальная или наименований	Принадлежность классу	Названия	=, ≠	Группировка	Мода
Порядковая или рангов	Сравнение, уровень	Порядковые величины	–//–, >, <	Сортировка	Медиана
Интервальная	Разница в интервалах	Относительные величины	–//–, +, –	Сравнение	Среднее арифметическое
Отношений	Абсолютная величина	Неотрицательные величины	–//–, *, /	Отношение	Среднее геометрическое и гармоническое
Частей	Часть от целого	Числа в диапазоне [0; 1]	–//–	Сравнение	Нет

Все шкалы делятся на категориальные и количественные, категориальные шкалы – на номинальные и порядковые (табл. 1). Чаще всего данные, представленные в категориальных шкалах, носят субъективный характер. Данные в шкалах наименований, например названия улиц, номера телефонов, могут быть представлены символами. Данные такого типа можно только сравнивать: равны они или нет. Данные в порядковых шкалах могут содержать оценки в виде чисел, позволяющих задать порядок и сравнить описания разных объектов. Например, ранжировать людей по росту не в сантиметрах, а выше или ниже, первый, второй или десятый. Школьные оценки описываются так же. Их можно сравнивать (больше или меньше, лучше или хуже), используя операции больше, меньше, равно. Вместе с тем описание данных в этих шкалах субъективно и применение к ним арифметических операций является некорректным. Например, нельзя утверждать, что ученик, получивший оценку 10, знает в два раза больше ученика, получившего оценку 5.

Данные в количественных шкалах представлены числами. Числа принимают значения из определенных шкал, которые можно разделить на три группы: интервальные, отношений и частей. Третья группа предлагается авторами настоящей статьи. Интервальные шкалы разбиты на равные интервалы, но не имеют строго определенного начала (нуля). В частности, время можно измерять в часах или годах, но начало исчисления условно. Такие данные можно складывать и вычитать, но их нельзя множить и делить, так как меняется цена интервала измерения. Шкалы отношений также разделены на равные интервалы, но они имеют строго определенный ноль, с которого начинается отсчет. Данные, представленные в таких шкалах, не могут быть отрицательными. К ним можно применять дополнительно операции умножения и деления, статистические операции. Например, двадцатилетний человек в два раза моложе сорокалетнего (данные представлены в шкале отношений), но родившийся в 2000 г. не будет в два раза младше родившегося в 1980 г. (данные представлены в шкале интервалов).

Классификацию Стивенса можно дополнить пятой группой – шкалами частей от целого (например, процентов в долях, КПД, частей угла, вероятности и т. п.). Такие шкалы начинаются с нуля и заканчиваются единицей, в них описываются безразмерные нормализованные данные.

Признаки одного и того же объекта могут описываться в разных шкалах. Например, ученик Иванов (по шкале наименований), ему 10 лет (по шкале отношений), он родился в 2011 г. (по шкале интервалов), он отличник (по шкале порядка), решил все задачи контрольной работы (единица по шкале частей), у него черные волосы (по шкале наименований).

В задачах классификации и кластеризации исходные данные должны быть описаны в числовом представлении и, желательно, преобразованы в единую шкалу измерений. Рассмотрим способы преобразования данных.

Нормализация и стандартизация данных. Данные, описанные в категориальной шкале, нормализовать невозможно.

Определение 1. Преобразование численных данных в диапазон с крайними значениями [0; +1] будем называть нормализацией данных.

Следует отличать понятия нормализации и нормировки данных. Нормировка – это преобразование данных, измеренных в разных единицах (например, когда длины объектов описаны в сантиметрах и дюймах), в единую шкалу.

Чаще всего нормализованные данные имеют значения в диапазоне [0; +1], реже – в диапазоне [−1; +1], но данные, представленные в этих диапазонах, легко трансформируются из одного диапазона в другой. Если количественные данные описаны в одной шкале с разными интервалами, после нормализации их можно сравнивать и оценивать математически. Фактически нормализация данных – это их преобразование в шкалу частей.

Линейная нормализация набора произвольных данных x выполняется по формуле

$$y = (x - \min(x)) / (\max(x) - \min(x)), \quad (1)$$

где x – исходное множество данных, y – преобразованное множество данных, min и max – операции вычисления минимального и максимального значений.

Десятичное масштабирование является нелинейным методом нормализации данных в диапазон [0; +1]:

$$y = (x - \min(x)) / 10^j, \quad j = \log_{10} |\max(x) - \min(x)|. \quad (2)$$

Еще одним способом нормализации данных можно считать метод вычисления интервала значений, который используется при построении блочной диаграммы распределения данных, называемой ящиком с усами и предложенной в работе [8]. Если дано множество одномерных величин $X_n = \{x_1, x_2, \ldots, x_n\}$, «ящик с усами» строится следующим образом. Вычисляются значения квартилей массива X_n $Q1$, $Q2$, $Q3$, $Q4$; межквартильный размах $IQR = Q3 - Q1$; границы диапазона основных значений ящика $L = Q1 - 1{,}5IQR$ и $R = Q3 + 1{,}5IQR$ (рис. 1). Диапазон основных значений данных будет равен $[L; R]$. Все точки вне этого диапазона классифицируются как выбросы. Для данных, имеющих нормальное распределение, всего около 0,7 % значений будут лежать за пределами вычисленного межквартильного интервала [9], а его легко преобразовать в диапазон значений [0; +1] по формуле (1), где $L = \min(x)$, $R = \max(x)$.

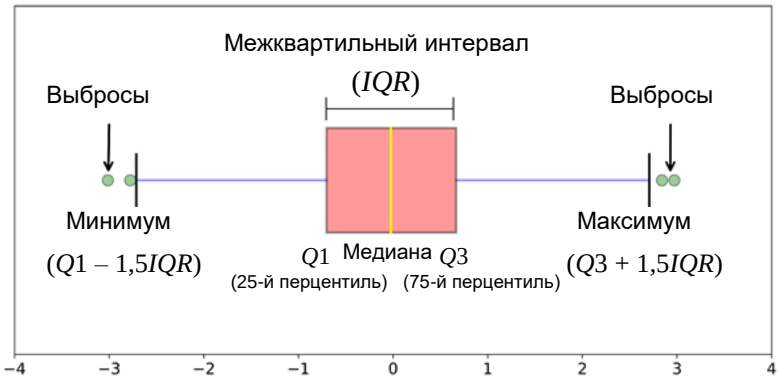


Рис. 1. Идея стандартного вычисления межквартильного интервала

Fig. 1. The idea of standard calculation of interquartile range

Для данных, не соответствующих нормальному распределению, в работе [10] скорректировали границы межквартильного интервала в сторону искажения данных, используя нижний и верхний полуквартильный диапазоны $A = Q2 - Q1$ и $B = Q3 - Q2$, которые определяют грани-

цы интервала как [$Q1 – 3A$; $Q3 + 3B$]. В статье [11] значения квартилей $Q1$ и $Q3$ заменены медианным значением второго квартиля $Q2$, а константа 1,5 заменена формулой, зависящей от размера диапазона. Такой подход увеличивает число вычислительных операций, поскольку применяется дополнительная сортировка, но, как правило, он незначительно уточняет границы искомого интервала.

В работе [12] для данных, соответствующих непрерывному унимодальному распределению, предложено вычислять отклонения от $Q1$ и $Q3$ в виде определенных функций. Однако для их построения требуется найти значение медианы попарных отклонений MC (medcouple), предложенной в статье [13] и вычисляемой от медианного значения исходного множества данных. Этот подход также требует дополнительных вычислений, поскольку сложность вычисления MC равна O ($n \log n$) для n величин.

Обобщая описанные подходы, можно сделать вывод, что если множество данных X_n имеет ограниченный диапазон значений [x_0; x_k], где $x_0 < x_k$, то его можно нормализовать, преобразовав в другое множество $X_n \rightarrow Y_n$ с фиксированным диапазоном значений [0; +1] или [–1; +1] либо с заданными свойствами. Например, преобразовав медианное значение множества, получим $y_{\text{med}}(x_{\text{med}}) = 0{,}5$ при $y(Q1) = y_1$, $y(Q3) = y_3$, $y(L) = 0$, $y(B) = 1$. Очевидно, что можно назначить и другие значения y для указанных значений x, например $y(x_{\min}) = 0$, $y(x_{\max}) = 1$.

Если данные имеют нормальное или примерно нормальное распределение, можно выполнить процедуру их стандартизации.

Определение 2. Преобразование данных в набор, имеющий определенные статистические характеристики, но неопределенные минимальные и максимальные значения, будем называть стандартизацией данных.

Наиболее распространенные методы стандартизации данных [2] приведены в табл. 2. Например, Z-преобразование меняет значения набора данных так, что он будет иметь нулевое среднее и единичную дисперсию.

Таблица 2. Методы стандартизации набора данных x

Table 2. Methods for dataset x standardization

Название *Name*	Формула *Formula*
Z-преобразование	$y = (x - \mu) / \sigma$, μ – среднее, σ^2 – дисперсия
Преобразование Парето	$y = (x - \mu) / \sqrt{\sigma^2}$
Масштабирование стабильности переменной (Variable Stability Scaling)	$y = \mu (x - \mu) / \sigma^2$
Степенное преобразование	$y = p - \mu^{p}$, $p = \sqrt{x - \min(x)}$
Med-MAD-преобразование	$y = (x - \text{med}(x)) / \text{MAD}$, $\text{MAD} = \text{med}(\lvert x - \text{med}\rvert)$

Стандартизованные данные можно нормализовать, убрав выбросы и преобразовав значения в диапазон [0; +1]. Если данные не имеют нормального распределения, указанные выше методы просто приведут их к другому масштабу, но не преобразуют в шкалу с диапазоном [0; +1]. Это методы стандартизации, а не нормализации данных.

Нормализация функций. Не всегда известно полное множество данных, однако может быть известна функция, описывающая их природу или позволяющая оценить границы диапазона значений.

Определение 3. Под нормализацией функции будем понимать преобразование, нормализующее область ее значений в диапазон [0; +1] или [–1; +1].

Основные варианты нормализации функций приведены в табл. 3. Основное отличие нормализации данных от нормализации функций состоит в том, что набор данных всегда конечен и можно вычислить его минимальное и максимальное, среднее и медианное значения, среднеквадратическое отклонение и другие статистические характеристики, которые используются

для нормализации и стандартизации данных. При нормализации функций известно только уравнение функции. По нему можно определить предельные значения, а также значения функции в некоторых точках.

Таблица 3. Основные варианты нормализации функций [14]

Table 3. The main options of function normalization [14]

Название *Name*	Формула *Formula*	Стандартные параметры *Standard parameters*	Диапазон значений *Range of values*
Алгебраическая функция	$f(x) = x / (\|x\| + a)$, $a > 0$	$a = 1$	[−1; +1]
Обобщение алгебраической функции	$f(x) = x^n / (\|x\|^n + a)$, $a > 0$	$n = 2$, $a = 1$	[0; +1]
Обобщенный вариант сигмоиды	$f(x) = (1 + \exp^{-bx})^{-a}$, $a,b > 0$	$a = b = 1$	[0; +1]
Арктангенс	$f(x) = \arctan(x/a)$, $a > 0$	$a = 1$	[–1; +1]
Гиперболический тангенс	$f(x) = \tanh(x) = (e^{x/a} - e^{-x/a}) / (e^{xa/} + e^{-xa/})$, $a > 0$	$a = 1$	[–1; +1]
На базе гиперболического тангенса	$f(x) = 0{,}5 \cdot x[\tanh(0{,}01(x/a - \mu) / \sigma) + 1]$, $a > 0$	$a = 1$	[0; +1]
Функция Гудермана	$f(x) = \mathrm{gd}(x) = 4 \cdot \arctan(\tanh(x/a) / \pi$, $a > 0$	$a = 2$	[−1; +1]

Если область значений функции ограничена, к функции можно применить те же преобразования, которые применяют для нормализации данных. Например, если некоторые данные описываются функцией $y = \sin(x) + \cos(x)$, то она нормализуется в диапазон [0; +1] следующим образом:

$$f(y) = (y + \sqrt{2}) / (2 \cdot \sqrt{2}).$$

Под нормализацией функции $a \le y(x) < +\infty$ понимается такое ее преобразование $f(y)$, что $0 \le f(y) \le 1$ и $f(a) = 0$, $f(\infty) = 1$. Если функция монотонна, то из $x_1 < x_2$ следует $f(y_1) \le f(y_2)$. Здесь x является переменной.

Функция с бесконечным диапазоном значений $-\infty < y(x) < +\infty$ обычно нормализуется в диапазон значений [−1; +1], где $f(-\infty) = -1$ и $f(\infty) = 1$. Значения функции f можно преобразовать в диапазон [0; +1] посредством дополнительного преобразования $g(f) = (f + 1) / 2$. Поэтому будем считать, что нормализованная функция вычисляет некоторую величину в диапазоне [0; +1], где крайние значения соответствуют минимальному и максимальному значениям функции f.

Если количественный признак описывает объекты двух классов со средними значениями классов μ1, μ2 и дисперсией σ^2, а данные в каждом классе имеют нормальное распределение с одинаковой дисперсией, то нормализованное значение признака дает логистическая функция (сигмоида) [15].

В работе [16] описаны свойства преобразования Box-Cox (3), предложенного для «исправления» данных, не имеющих нормального распределения, и показано, что применение этого преобразования улучшает результаты классификации. Экспериментальные исследования показали, что если данные плохо описываются нормальным распределением, то требуется многократное применение этого преобразования и оптимизация оценки максимального правдоподобия относительно нормального распределения для каждого варианта. На рис. 2 слева изображена гистограмма множества данных, представленных в диапазоне $[3{,}944 \cdot 10^{-05}; 3{,}372]$ и имеющих распределение Вейбулла, а справа – результат их преобразования в новое представление в диапазоне [–4,342; 1,376], имеющее нормальное распределение с уровнем значимости $p = 0{,}90$. Аппроксимация обоих наборов данных функцией нормального распределения показана кривыми красного цвета. Результат стандартизации данных получен посредством применения преобразования Box-Cox к набору данных x с параметром $\lambda = 0{,}2$. Далее при необходимости можно нормализовать эти данные методами (1) или (2):

$$y = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} \text{ для } \lambda \neq 0, \\ \log(x) \text{ для } \lambda = 0, \end{cases} \quad (3)$$

где λ – константа.

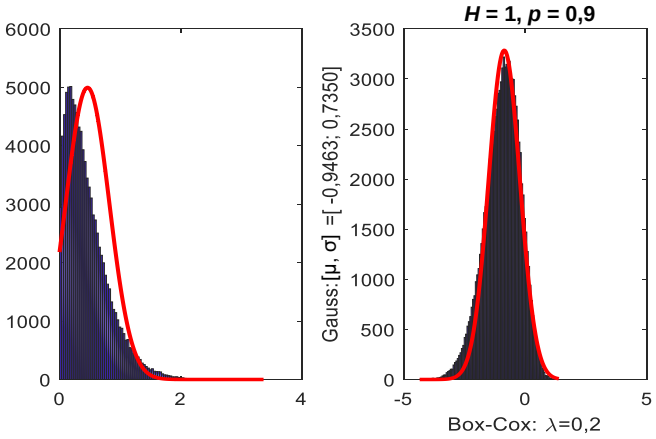


Рис. 2. Пример стандартизации данных, имеющих распределение Вейбулла с параметрами (0,5; 1,3), в нормально распределенное представление с параметрами (–0,946; 0,735)

Fig. 2. An example of standardization of data with Weibull distribution with parameters (0,5; 1,3) into a normally distributed representation with parameters (–0,946; 0,735)

Недостатком метода линейной нормализации (1) является наличие аномальных значений данных (выбросов), которые «увеличивают» диапазон. Это приводит к тому, что нормализованные значения концентрируются в узком диапазоне вблизи нуля. Чтобы избежать этого, следует определять диапазон с помощью не максимальных и минимальных значений, а среднего значения и дисперсии.

Дополнительно данным можно придать определенные свойства, например убрать выбросы, отцентрировать данные в середину диапазона.

Применение описанных функций. Рассмотрим примеры описания температуры воды в разных шкалах (табл. 4).

Таблица 4. Температура воды в единицах измерения разных шкал

Table 4. Water temperature in different scales measurement

Состояние воды *Water condition*	Кельвин *Kelvin*	Градус Цельсия *Celsius*	Градус Фаренгейта *Fahrenheit*	Градус Ньютона *Newton*	Градус Реомюра *Reaumur*
Замерзание	273,15	0	32	0	0
Кипение	373,15	100	212	33	80

Шкала Кельвина относится к шкалам отношений, поскольку в ней есть абсолютный ноль, а остальные шкалы – к интервальным. Все шкалы в табл. 4 связаны линейными зависимостями. Теоретический верхний предел температуры равен примерно 10^{32} К, т. е. можно считать, что диапазоны температурных шкал не ограничены справа. Самая низкая температура воздуха, зафиксированная на Земле, равна –91,2 °С (181,95 К), а самая высокая +56,7 °С. Учитывая эти экстремальные значения, температуру воздуха, описанную в любой шкале, можно перевести в нормализованный вид в шкале частей со значениями от 0 до +1. Если нормализовать температуры, указанные в табл. 1, в шкале Кельвина по формуле $y = x / (x + 1)$, то для 0 °С получим число 0,996 352, а для 100 °С – число 0,997 327. Новые значения различаются только тысячными долями. Если температура воздуха нормализована линейно в указанном интервале [min; max], то значение 0 °С будет преобразовано в число 0,616 63.

В зависимости от решаемой задачи нормализовать такой параметр, как значение температуры, можно разными способами. Пользуясь функциями принадлежности нечеткому множеству (например, функцией Гаусса), на базе показаний температуры можно также сформировать новый параметр, например степень комфортности температуры на улице или в помещении с диапазоном значений [0; +1] и пиком, равным единице и соответствующим 22 °C.

Формирование новых признаков на базе исходных данных. Один из самых известных алгебраических методов конструирования признаков называется методом главных компонент (principal component analysis, PCA). Главные компоненты – это новые признаки, сконструированные в виде линейных комбинаций исходных признаков [15].

Авторы работы [17] предложили применять PCA к численным признакам и нормализовать значение каждого признака, разделив на большее собственное число и умножив на собственное число, соответствующее этому признаку. На четырех наборах данных разных типов было продемонстрировано повышение точности классификации с помощью искусственных нейронных сетей (ИНС).

Функции принадлежности нечеткому множеству определяют степень принадлежности количественных величин определенному нечеткому множеству [18]. Диапазон их значений равен [0; +1]. Четыре основных типа функций определяются формой кривой функции принадлежности, типичными их представителями являются треугольная и трапециевидная функции, а также сигмоида и функция Гаусса (рис. 3). Следует отметить следующие особенности этих функций: не всегда максимальное значение функции соответствует крайним значениям диапазона области определения функции, одно и то же значение функции может соответствовать разным величинам диапазона области определения.

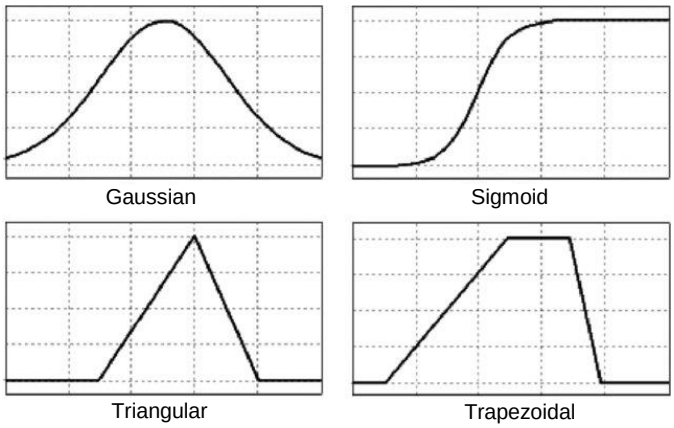


Рис. 3. Графики четырех основных типов функций нечеткой принадлежности
Fig. 3. Plots of four main types of fuzzy membership functions

В рамках настоящей статьи функции нечеткой принадлежности можно рассматривать как функции нормализации, которые преобразуют количественное описание данных в качественную шкалу значений, т. е. формируют другое представление этих данных. Например, рост человека, измеренный в сантиметрах или дюймах, можно качественно описать как низкий, средний и высокий. Рост двух и более человек можно сравнить как в количественной шкале, так и качественной.

Применение при кластеризации и классификации. Методы кластерного анализа и многие классификаторы рассчитывают расстояние между двумя точками в признаковом пространстве, используя евклидову метрику. Если один из признаков имеет более широкий диапазон значений, расстояние будет определяться этим доминирующим признаком. Диапазон всех признаков при классификации должен быть нормализован так, чтобы каждый вносил приблизительно пропорциональный вклад [19]. Еще одна причина, по которой применяется масштабирование значений признаков, заключается в том, что градиентный спуск сходится примерно в 14 раз

быстрее на нормализованных данных [20]. Авторы статьи [20] тестировали преобразование Парето на базе данных ImageNet. Для этой цели также применяют *Z*-преобразование [3]. Алгоритм логистической калибровки признака (*z*-оценки) описан в работе [15].

Сигмоида часто используется в ИНС в качестве функции активации, которая позволяет усиливать слабые сигналы и не усиливать сильные. Производная сигмоиды может быть легко выражена через саму функцию:

$$S'(x) = S(x) \times (1 - S(x)),$$

что позволяет сократить вычислительную сложность методов обучения ИНС. Вместо сигмоиды можно использовать гиперболический тангенс. Его производная также вычисляется через исходную функцию:

$$\tanh'(x) = 1 - \tanh^2(x).$$

Отметим, что максимальное значение производной сигмоиды равно 0,25, а максимальное значение производной тангенса – единице. Поэтому обучение ИНС с использованием гиперболического тангенса вместо сигмоиды происходит быстрее.

Если количественный признак описывает объекты двух классов со средними значениями классов μ1, μ2 и дисперсией σ^2, а в каждом классе данные имеют нормальное распределение с одинаковой дисперсией, то лучшее нормализованное значение признака дает сигмоида [15].

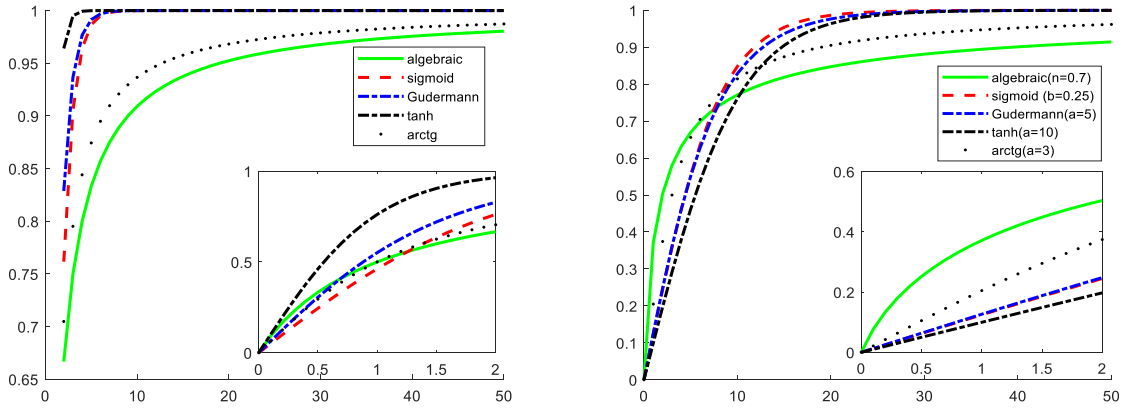
В статье [2] рассматривалось применение 14 методов нормализации и стандартизации данных и их влияние на производительность классификации с учетом полного набора функций, выбора функций и их взвешивания. Эксперименты были выполнены на 21 общедоступной базе реальных и синтезированных данных [2]. Было отмечено, что ни один метод не превосходит другие во всех экспериментах. Поэтому авторы отметили наборы лучших методов. Лучшими вариантами нормализации (точнее, стандартизации) были признаны *Z*-преобразование и преобразование Парето, а также гиперболический тангенс.

Преимущество древовидных классификаторов. Древовидные классификаторы используют деление признакового пространства по правилу сравнения значений одного признака с некоторой пороговой величиной, т. е. основой является операция сравнения. Если изменяется диапазон значений признаков, пороговые значения в правилах древовидного классификатора можно легко изменить, но сами правила не меняются. Древовидные модели нечувствительны к шкале количественного признака. Например, для обученного дерева не важно, измерена температура по шкале Цельсия или Фаренгейта. Несущественен и переход от линейной шкалы к логарифмической, т. е. порог разделения просто будет равен log (*v*), а не *v* [15]. Древовидные модели нечувствительны к монотонным преобразованиям шкалы признака – преобразованиям, не изменяющимся относительного порядка его значений. Таким образом, древовидные модели не используют шкалу количественных признаков, а трактуют признаки как порядковые. То же самое справедливо для моделей на основе правил. Возможно, это является причиной того, что классификатор типа «случайный лес» в результате многочисленных экспериментов оказался лучшим среди 179 вариантов классификаторов 17 классов, протестированных на 121 множестве данных в статье [21]. В работе [22] показано, что классификатор «случайный лес» продемонстрировал более точную классификацию рака груди на два класса по 31 признаку по сравнению с байесовским классификатором.

Сравнение оценочных функций с разными диапазонами значений. На рис. 4 видно, что графики сигмоиды, гиперболического тангенса и функции Гудермана очень близки, графики арктангенса и алгебраической функции имеют несколько отличную форму, а последняя функция – наиболее плавную форму. При большом диапазоне нормализуемых значений лучше использовать арктангенс и алгебраическую функцию, задавая определенные значения функции нормализации $y = f(x)$ посредством выбора ее параметров. Например, можно получить $y = 0{,}5$ для исходного значения $x = 8$, применяя любую функцию нормализации. При этом диапазон

исходных значений [0; 8] будет преобразован в диапазон [0; 0,5], а оставшийся диапазон [8; +∞) – в диапазон [0,5; +1]. Разные функции нормализации по-разному преобразуют эти два поддиапазона.

В работе [23] сравнивались результаты бинарной классификации данных разной природы. Результаты оценивались по матрице ошибок размером 2×2. Целью исследования было сравнение точности оценок двух функций: коэффициента корреляции Мэтьюса (MCC) с диапазоном значений [–1; +1] и популярного в медицинских исследованиях диагностического отношения шансов (DOR) с диапазоном значений [0; +∞). Обе функции были нормализованы в диапазон значений [0; +1]: первая – линейно, вторая – с помощью алгебраической функции $y = x/(x + 1)$. Было показано, что функция DOR инварианта к дисбалансу классов, но функция MCC является более информативной. При числе анализируемых объектов более 53 коэффициент корреляции Пирсона между значениями нормализованных функций в среднем был более 0,95 для четырех миллионов искусственно сгенерированных матриц ошибок.



a) *b)*

Рис. 4. Графики функций нормализации со стандартными параметрами (*a*) и измененными (*b*)

Fig. 4. Plots of normalization functions with the standard parameters (a) and modified (b)

Аналогично можно сравнить точность оценок, получаемых другими функциями с разными диапазонами значений, например коэффициентом корреляции Пирсона (диапазон [0; +1]) и среднеквадратическим отклонением (диапазон [0; +∞)), нормализуя их значения.

Следует отметить: если данные двух типов преобразованы в одну шкалу, то из выражения $y_1 = f_1(x) > y_2 = f_2(x)$ не следует, что первое значение или признак лучше второго. Примером смогут служить коэффициенты корреляции Пирсона, Спирмена и Кендалла. Их значения обычно находятся в отношениях Пирсон > Спирмен > Кендалл. Следует оценивать динамику изменений этих функций на равных диапазонах из области их определения.

Проблема адекватности применения математических операций к данным, представленным в разных шкалах. Известно, что к данным рангового типа нельзя применять арифметические операции [24], однако это происходит повсеместно. Например, после защиты диссертации каждый член совета по защите выставляет оценку диссертации, затем вычисляется средний балл. Изначально выставляются оценки «отлично» (5), «хорошо» (4), «удовлетворительно» (3) и «неудовлетворительно» (2). В данном случае цифры – это символы, означающие только порядковые эквиваленты оценок. Пусть две трети совета поставила оценку «хорошо» и одна треть – «отлично». Сложить оценки «хорошо» и «отлично» нельзя, но их заменяют числами и тогда средний балл считается равным 4,3. Что это означает в данной шкале оценок? Оценка диссертации на 30 % выше оценки «хорошо»? Но такой оценки нет в применяемой шкале, по ней можно вычислить только медианное значение оценки, т. е. «хорошо».

Если те же оценки представить в виде чисел и перейти в шкалу частей, то можно вычислить, какую долю от максимально возможной суммы получила диссертация. Оценка будет равна 0,86.

Аналог такого подхода имеет место при вычислении оценок централизованного тестирования, если их разделить на максимальное число 100.

Нормализация данных в цифровой обработке изображений. В области классификации и кластеризации данных цифровые изображения часто используются как исходные данные. При линейном растяжении контраста или гамма-коррекции изображений они сначала подвергаются процедуре нормализации, т. е. преобразования в диапазон значений [0; 1] по формуле (1). Насколько корректно такое преобразование? Напомним, что значения яркостей цифрового изображения являются безразмерными величинами и обозначают номер диапазона, в который попадает значение сигнала в результате его квантования при выполнении аналого-цифрового преобразования. Таким образом, значение яркости одного и того же сигнала может быть нулем или другим числом в зависимости от выбранного числа диапазонов при квантовании (два или более). Фактически яркость описывается в порядковой шкале с фиксированным числом значений, обычно кратным степени двойки. К таким данным нельзя применять арифметические операции, поэтому при обработке изображений значения яркости сначала преобразуют в другую шкалу (численную), а затем обратно в порядковую шкалу.

А. Чеддад в статье [25] предложил интересный метод нормализации цифровых изображений методом Box-Cox, утверждая, что он повышает качество преобразованных изображений по сравнению с другими подходами. Авторы настоящей статьи проверили, действительно ли возможна нормализации исходных данных, представленных в виде изображений. На рис. 5 приведены примеры преобразования изображения методом, предложенным Чеддадом, с разными значениями параметра λ. На рис. 6 показаны гистограммы двух вариантов преобразования изображений методом Box-Cox. При $\lambda < 1$ значения яркости нелинейно увеличиваются, при $\lambda > 1$ – нелинейно уменьшаются. Оказалось, что, как правило, яркости цифровых изображений имеют распределения, существенно отличные от нормального, и методом Box-Cox их невозможно привести к нормальному распределению. В данном примере результатом является степенное преобразование значений яркости из исходного диапазона [14; 248] в новые диапазоны [3,3322; 5,5683], [5,2013; 12,0929], [10,6401; 45,4126], [34,6116; 414,7514] с последующим преобразованием по формуле (1) в диапазон [0; 255].

Рис. 5. Слева – оригинальное изображение, далее – изображения, преобразованные по формуле (3) с параметром λ, равным 0; 0,25; 0,6; 1,1
Fig. 5. First left is the original image, then the images are transformed according to the formula (3) with the parameter λ equal to 0; 0,25; 0,6; 1,1

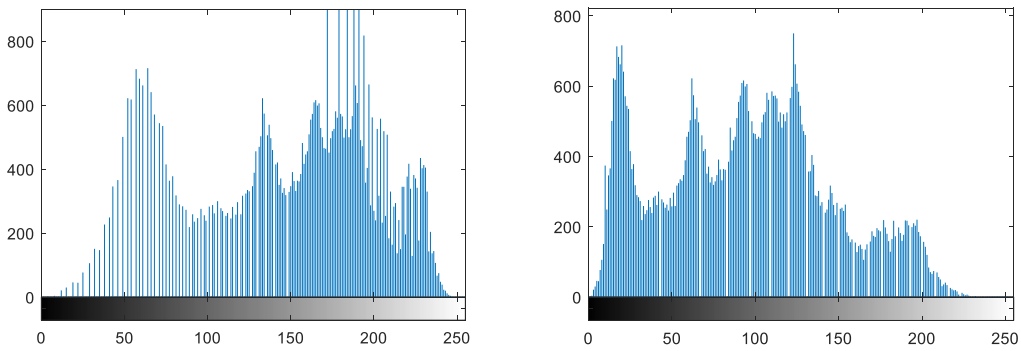


Рис. 6. Гистограммы изображений после преобразований с параметром λ, равным 0 и 1,1
Fig. 6. Histograms of the images after transformations with parameters λ as 0 and 1,1

Заключение. Задачи классификации и кластеризации относятся к области машинного обучения. Данные в них часто представлены в разных шкалах. Больше информации и возможностей для анализа несут данные, представленные в количественных шкалах. Данные, представленные во всех шкалах, кроме номинальной, можно сравнивать, используя операции больше, меньше либо равно.

В статье описаны основные классы шкал представления данных разных типов. Приведены основные подходы к нормализации и стандартизации данных, а также нормализации функций с конечным и бесконечным диапазонами значений.

Многие экспериментальные исследования показывают, что универсального метода нормализации данных, превосходящего другие методы, не существует, но нормализация исходных данных позволяет повысить точность их классификации [4, 19, 20, 26, 27]. Следовательно, кластеризацию данных (особенно методами, использующими функции расстояния) лучше выполнять после преобразования всех признаков в единую шкалу.

Результаты классификации и кластеризации разными методами можно сравнивать различными оценочными функциями. Они имеют разные диапазоны значений. Для выбора наиболее подходящей функции можно выполнить нормализацию всех сравниваемых функций, тогда сравниваемые оценки будут в одной шкале – шкале частей. Байесовский классификатор использует понятие вероятности наступления некоторого события, а вероятность также можно считать описанием данных в шкале частей.

Древовидные модели классификации инвариантны к шкалам количественных признаков. При изменении шкал просто пересчитываются пороговые значения, но не меняются правила в древовидных моделях классификации. Возможно, благодаря этому свойству классификатор типа «случайный лес» является одним из лучших при анализе данных разной природы.

В следующей статье будет показано, как использовать нормализацию признаков и оценки классификации при анализе данных различной природы.

Вклад авторов. В. В. Старовойтов определил план статьи и задачи, которые необходимо было решить при проведении исследований, принял участие в интерпретации результатов; Ю. И. Голуб описала представления данных разных типов при их анализе, выполнила экспериментальные исследования различных вариантов нормализации и стандартизации данных, предложила рекомендации по нормализации и стандартизации входных данных для решения задач классификации. Шкала частей для представления данных была разработана совместно.

Список использованных источников

1. Aksoy, S. Feature normalization and likelihood-based similarity measures for image retrieval / S. Aksoy, R. M. Haralick // Pattern Recognition Letters. – 2001. – Vol. 22, no. 5. – P. 563–582.

2. Singh, B. Investigating the impact of data normalization on classification performance / B. Singh // Applied Soft Computing J. – 2020. – Vol. 97. – P. 105524.

3. Nayak, S. C. Impact of data normalization on stock index forecasting / S. C. Nayak, B. B. Misra, H. S. Behera // Intern. J. of Computer Information Systems and Industrial Management Applications. – 2014. – Vol. 6. – P. 257–269.

4. Naeini, A. A. Assessment of normalization techniques on the accuracy of hyperspectral data clustering / A. A. Naeini, M. Babadi, S. Homayouni // Intern. Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences. – 2017. – Vol. 42. – P. 27–30.

5. Stevens, S. S. On the theory of scales of measurement / S. S. Stevens // Science. New Series. – 1946. – Vol. 103, no. 2684. – P. 677–680.

6. Орлов, А. И. Теория измерений как часть методов анализа данных / А. И. Орлов // Социология: методология, методы, математическое моделирование. – 2012. – № 35. – С. 155–174.

7. Velleman, P. F. Nominal, ordinal, interval, and ratio typologies are misleading / P. F. Velleman, L. Wilkinson // The American Statistician. – 1993. – Vol. 47, no. 1. – P. 65–72.

8. Tukey, J. W. Exploratory Data Analysis / J. W. Tukey. – Massachusetts : Addison-Wesley, 1977. – P. 39–49.

9. Bruffaerts, C. A generalized boxplot for skewed and heavy-tailed distributions / C. Bruffaerts, V. Verardi, C. Vermandele // Statistics & Probability Letters. – 2014. – Vol. 95. – P. 110–117.

10. Kimber, A. C. Exploratory data analysis for possibly censored data from skewed distributions / A. C. Kimber // Applied Statistics. – 1990. – Vol. 39. – P. 21–30.

11. Carling, K. Resistant outlier rules and the non-Gaussian case / K. Carling // Computational Statistics & Data Analysis. – 2000. – Vol. 33, no. 3. – P. 249–258.

12. Hubert, M. An adjusted boxplot for skewed distributions / M. Hubert, E. Vandervieren // Computational Statistics & Data Analysis. – 2008. – Vol. 52, no. 12. – P. 5186–5201.

13. Brys, G. A robust measure of skewness / G. Brys, M. Hubert, A. Struyf // J. of Computational and Graphical Statistics. – 2004. – Vol. 13. – P. 996–1017.

14. Kyurkchiev, N. Sigmoid Functions: Some Approximation and Modelling Aspects / N. Kyurkchiev, S. Markov. – Saarbrucken : LAP Lambert Academic Publishing, 2015. – 120 p.

15. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах. – М. : ДМК Пресс, 2015. – 402 с.

16. Bicego, M. Properties of the Box-Cox transformation for pattern classification / M. Bicego, S. Baldo // Neurocomputing. – 2016. – Vol. 218. – P. 390–400.

17. Zhang, Q. Weighted data normalization based on eigenvalues for artificial neural network classification / Q. Zhang, S. Sun // Proc. of Intern. Conf. Neural Information Processing. – 2009. – Vol. 5863. – P. 349–356. https://doi.org/10.1007/978-3-642-10677-4_39

18. Zadeh, L. A. Fuzzy sets / L. A. Zadeh // Information and Control. – 1965. – Vol. 8, no. 3. – P. 338–353.

19. Więckowski, J. How the normalization of the decision matrix influences the results in the VIKOR method? / J. Więckowski, W. Sałabun // Procedia Computer Science. – 2020. – Vol. 176. – P. 2222–2231.

20. Ioffe, S. Batch normalization: accelerating deep network training by reducing internal covariate shift / S. Ioffe, C. Szegedy // 32nd Intern. Conf. on Machine Learning, Lille, France, 7–9 July 2015. – Lille, 2015. – Vol. 37. – P. 448–456.

21. Do we need hundreds of classifiers to solve real world classification problems? / M. Fernández-Delgado [et. al.] // The J. of Machine Learning Research. – 2014. – Vol. 15, no. 1. – P. 3133–3181.

22. Lemons, K. Comparison between Naive Bayes and random forest to predict breast cancer / K. A. Lemons // Intern. J. of Undergraduate Research & Creative Activities. – 2020. – Vol. 12, art. 12. – P. 1–5. http://doi.org/10.7710/2168-0620.0287

23. Chicco, D. The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment / D. Chicco, V. Starovoitov, G. Jurman // IEEE Access. – 2021. – Vol. 9. – P. 47112–47124. https://doi.org/10.1109/ACCESS.2021.3068614

24. Новиков, Д. А. Статистические методы в педагогических исследованиях (типовые случаи) / Д. А. Новиков. – М. : МЗ-Пресс, 2004. – 67 с.

25. Cheddad, A. On box-cox transformation for image normality and pattern classification // IEEE Access. – 2020. – Vol. 8. – P. 154975–154983. https://doi.org/10.1109/ACCESS.2020.3018874

26. Han, J. The influence of the sigmoid function parameters on the speed of backpropagation learning / J. Han, C. Moraga // Intern. Workshop on Artificial Neural Networks, Malaga-Torremolinos, Spain, 7–9 June 1995. – Malaga-Torremolinos, 1995. – P. 195–201.

27. Jain, A. Score normalization in multimodal biometric systems / A. Jain, K. Nandakumar, A. Ross // Pattern Recognition. – 2005. – Vol. 38, no. 12. – P. 2270–2285.

References

1. Aksoy S., Haralick R. M. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters*, 2001, vol. 22, no. 5, pp. 563–582.

2. Singh B. Investigating the impact of data normalization on classification performance. *Applied Soft Computing Journal*, 2020, vol. 97, p. 105524.

3. Nayak S. C., Misra B. B., Behera H. S. Impact of data normalization on stock index forecasting. *International Journal of Computer Information Systems and Industrial Management Applications*, 2014, vol. 6, pp. 257–269.

4. Naeini A. A., Babadi M., Homayouni S. Assessment of normalization techniques on the accuracy of hyperspectral data clustering. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2017, vol. 42. pp. 27–30.

5. Stevens S. S. On the theory of scales of measurement. *Science. New Series*, 1946, vol. 103, no. 2684, pp. 677–680.

6. Orlov A. I. *Measurement theory as part of data analysis methods*. Sotsiologiya: metodologiya, metody, matematicheskoe modelirovanie [*Sociology: Methodology, Methods, Mathematical Modeling*], 2012, no. 35, pp. 155–174 (In Russ.).

7. Velleman P. F., Wilkinson L. Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 1993, vol. 47, no. 1, pp. 65–72.

8. Tukey J. W. *Exploratory Data Analysis*. Massachusetts, Addison-Wesley, 1977, pp. 39–49.

9. Bruffaerts C., Verardi V., Vermandele C. A generalized boxplot for skewed and heavy-tailed distributions. *Statistics & Probability Letters*, 2014, vol. 95, pp. 110–117.

10. Kimber A. C. Exploratory data analysis for possibly censored data from skewed distributions. *Applied Statistics*, 1990, vol. 39, pp. 21–30.

11. Carling K. Resistant outlier rules and the non-Gaussian case. *Computational Statistics & Data Analysis*, 2000, vol. 33, no. 3, pp. 249–258.

12. Hubert M., Vandervieren E. An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, 2008, vol. 52, no. 12, pp. 5186–5201.

13. Brys G., Hubert M., Struyf A. A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 2004, vol. 13, pp. 996–1017.

14. Kyurkchiev N., Markov S. *Sigmoid Functions: Some Approximation and Modelling Aspects*. Saarbrucken, LAP Lambert Academic Publishing, 2015, 120 p.

15. Flach P. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. 1st ed. Cambridge University Press, 2012, 409 p.

16. Bicego M., Baldo S. Properties of the Box-Cox transformation for pattern classification. *Neurocomputing*, 2016, vol. 218, pp. 390–400.

17. Zhang Q., Sun S. Weighted data normalization based on eigenvalues for artificial neural network classification. *Proceedings of International Conference Neural Information Processing*, 2009, vol. 5863, pp. 349–356. https://doi.org/10.1007/978-3-642-10677-4_39

18. Zadeh L. A. Fuzzy sets. *Information and Control*, 1965, vol. 8, no. 3, pp. 338–353.

19. Więckowski J., Sałabun W. How the normalization of the decision matrix influences the results in the VIKOR method? *Procedia Computer Science*, 2020, vol. 176, pp. 2222–2231.

20. Ioffe S., Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015*. Lille, 2015, vol. 37, pp. 448–456.

21. Fernández-Delgado M., Cernadas E., Barro S., Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 2014, vol. 15, no. 1, pp. 3133–3181.

22. Lemons K. A. Comparison between Naive Bayes and random forest to predict breast cancer. *International Journal of Undergraduate Research & Creative Activities*, 2020, vol. 12, art. 12, pp. 1–5. http://doi.org/ 10.7710/2168-0620.0287

23. Chicco D., Starovoitov V., Jurman G. The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment. *IEEE Access*, 2021, vol. 9, pp. 47112–47124. http://doi.org/10.1109/ACCESS.2021.3068614

24. Novikov D. A. Statisticheskie metody v pedagogicheskikh issledovaniyakh (tipovye sluchai). *Statistical Methods in Pedagogical Research (Typical Cases)*. Moscow, MZ-Press, 2004, 67 p. (In Russ.).

25. Cheddad A. On box-cox transformation for image normality and pattern classification. *IEEE Access*, 2020, vol. 8, pp. 154975–154983. http://doi.org/10.1109/ACCESS.2020.3018874

26. Han J., Moraga C. The influence of the sigmoid function parameters on the speed of backpropagation learning. *International Workshop on Artificial Neural Networks, Malaga-Torremolinos, Spain, 7–9 June 1995*. Malaga-Torremolinos, 1995, pp. 195–201.

27. Jain A., Nandakumar K., Ross A. Score normalization in multimodal biometric systems. *Pattern Recognition*, 2005, vol. 38, no. 12, pp. 2270–2285.

Информация об авторах

Старовойтов Валерий Васильевич, доктор технических наук, профессор, главный научный сотрудник, Объединенный институт проблем информатики Национальной академии наук Беларуси.
E-mail: valerys@newman.bas-net.by

Голуб Юлия Игоревна, кандидат технических наук, доцент, старший научный сотрудник, Объединенный институт проблем информатики Национальной академии наук Беларуси.
E-mail: 6423506@gmail.com

Information about the authors

Valery V. Starovoitov, Dr. Sci. (Eng.), Professor, Chief Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus.
E-mail: valerys@newman.bas-net.by

Yuliya I. Golub, Cand. Sci. (Eng.), Associate Professor, Senior Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus.
E-mail: 6423506@gmail.com