

УДК 004.056:519.24

doi: 10.21685/2072-3059-2023-4-3

**Распознавание малых выборок с заданным
распределением данных при использовании
искусственных нейронов, предсказывающих
доверительные вероятности собственных решений**

В. И. Волчихин¹, А. И. Иванов², А. В. Безяев³, И. А. Филипов⁴

^{1,3,4}Пензенский государственный университет, Пенза, Россия

²Пензенский научно-исследовательский электротехнический институт, Пенза, Россия

¹cnit@pnzgu.ru, ²ivan@pniei.penza.ru, ³tsib@pnzgu.ru, ⁴re.wolf@mail.ru

Аннотация. *Актуальность и цели.* Повышение достоверности статистической обработки данных на малых выборках является актуальной задачей. *Материалы и методы.* Предложено использовать три искусственных нейрона, являющихся аналогами хи-квадрат критерия, критерия четвертого статистического момента и критерия Гири. Также использована процедура дополнительного обучения выходных нелинейных функций искусственных нейронов для прогнозирования доверительных вероятностей относительно принимаемых нейронами решений. *Результаты.* Показан существенный рост числа обнаруживаемых и исправляемых ошибок при свертывании избыточных кодов нейросетевого классификатора. *Выводы.* Подтверждено, что использование параллельно нескольких статистических критериев дает более достоверный результат в сравнении с одним критерием, и для их объединения могут быть использованы сложные конструкции кодов, способных обнаруживать и исправлять ошибки. Численным экспериментом подтверждено, что двухслойная нейросеть позволяет снизить уровень обнаруженных, но не подлежащих исправлению ошибок до вероятности 0,141. Линейная экстраполяция результатов численного эксперимента позволяет ожидать доверительной вероятности 0,9 уже при использовании 5 искусственных нейронов первого слоя. Тем самым наблюдается существенное снижение затрат на защиту приложений за счет использования в доверенной вычислительной среде SIM-карт, RFID-карт, микроSD-карт, USB-Био Токенов, ПЛИС, DSP-контроллеров.

Ключевые слова: хи-квадрат критерий, критерий четвертого статистического момента, критерий Гири, малые выборки, проверка гипотезы нормальности

Для цитирования: Волчихин В. И., Иванов А. И., Безяев А. В., Филипов И. А. Распознавание малых выборок с заданным распределением данных при использовании искусственных нейронов, предсказывающих доверительные вероятности собственных решений // Известия высших учебных заведений. Поволжский регион. Технические науки. 2023. № 4. С. 31–39. doi: 10.21685/2072-3059-2023-4-3

**Recognition of small samples with a given data distribution
using artificial neurons that predict the confidence
probabilities of their own decisions**

V.I. Volchikhin¹, A.I. Ivanov², A.V. Bezyaev³, I.A. Filipov⁴

^{1,3,4}Penza State University, Penza, Russia

²Penza Scientific Research Electrotechnical Institute, Penza, Russia

¹cnit@pnzgu.ru, ²ivan@pniei.penza.ru, ³tsib@pnzgu.ru, ⁴re.wolf@mail.ru

Abstract. *Background.* Improving the reliability of statistical data processing on small samples. Materials and methods - it is proposed to use three artificial neurons, which are analogues of the chi-square test, the fourth statistical moment test and the Geary test. Additionally, the procedure for additional training of output nonlinear functions of artificial neurons was used to predict the confidence probabilities regarding decisions made by neurons. *Results.* A significant increase in the number of detected and corrected errors during the convolution of redundant codes of the neural network classifier is shown. *Conclusions.* It has been confirmed that the use of several statistical criteria in parallel gives a more reliable result in comparison with one criterion, and complex code designs capable of detecting and correcting errors can be used to combine them. A numerical experiment confirmed that a two-layer neural network can reduce the level of detected, but not correctable, errors to a probability of 0.141. Linear extrapolation of the results of a numerical experiment allows us to expect a confidence probability of 0.9 already when using 5 artificial neurons of the first layer. Thus, there is a significant reduction in the cost of protecting applications due to the use of SIM cards, RFID cards, microSD cards, USB BioTokens, FPGAs, DSP controllers in a trusted computing environment.

Keywords: chi-square test, fourth statistical moment test, Geary's test, small samples, testing of the normality hypothesis

For citation: Volchikhin V.I., Ivanov A.I., Bezyaev A.V., Filipov I.A. Recognition of small samples with a given data distribution using artificial neurons that predict the confidence probabilities of their own decisions. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki = University proceedings. Volga region. Engineering sciences.* 2023;(4):31–39. (In Russ.). doi: 10.21685/2072-3059-2023-4-3

Введение

Известные итерационные алгоритмы обучения сетей искусственных нейронов [1, 2] неустойчивы из-за неустойчивости вычисления производных приращения либо снижения показателя качества обучения. Как следствие, итерационные алгоритмы обучения трудно полностью автоматизировать. Проблему полной автоматизации обучения нейронов удастся решить, если отказаться от итерационных процедур оптимизации и воспользоваться детерминированными процедурами приближенного вычисления весовых коэффициентов однослойной сети перцептронов [3]¹.

При решении задач нейросетевой биометрико-криптографической аутентификации необходимо использовать доверенную вычислительную среду (SIM-карт, RFID-карт, microSD-карт, USB-БиоТокенов, ПЛИС, DSP-контроллеров). Как правило, доверенные контроллеры имеют ограниченные вычислительные ресурсы (ограниченный объем памяти, ограниченное число аппаратно реализованных криптографических функций, ограниченное энергопотребление). Предположительно, параллельно с аппаратной реализацией криптографических функций в доверенной вычислительной среде должны появиться аппаратно реализованные функции нейросетевых преобразований. В частности, перспективой является использование в контроллерах аппаратных реализаций искусственных нейронов, являющихся аналогами статистических критериев [4–6], построенных на проверке гипотез о нормальном и равномерном распределении малых выборок.

Например, вектор из нескольких сотен биометрических параметров может быть использован для формирования малых выборок по 16 опытов

¹ ГОСТ Р 52633.5–2011. Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия-код доступа.

с нормальным законом распределения значений и с равномерным распределением значений. Если воспользоваться классическим хи-квадрат критерием для проверки гипотезы нормального распределения данных, то мы получим распределение откликов, отображенное на рис. 1. В левой части рисунка дана программа, воспроизводящая критерий хи-квадрат при воздействии на него нормально распределенными данными.

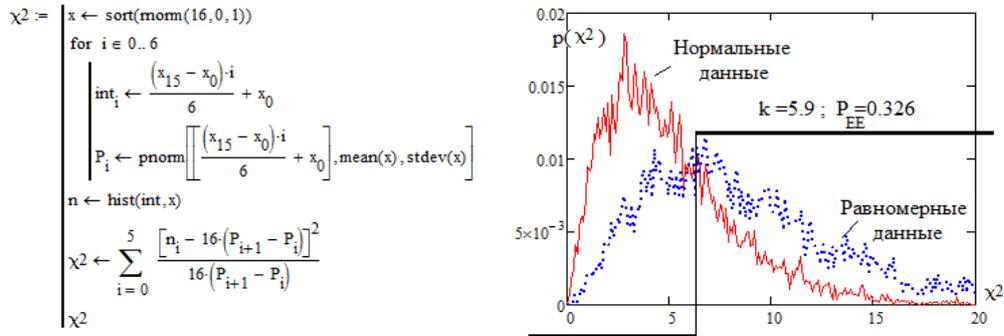


Рис. 1. Отклики хи-квадрат критерия, настроенного проверять гипотезу нормального распределения малой выборки в 16 опытов

В правой части рисунка отображены распределения откликов хи-квадрат критерия на нормальные и равномерные данные (для получения откликов на равномерные данные следует первую строку программы заменить на другую: $x \leftarrow \text{sort}(\text{runif}(16, -1, 1))$). Формально мы можем поставить в соответствие хи-квадрат критерию эквивалентный ему искусственный нейрон, поставив поле хи-квадрат обогатителя данных квантователь с порогом $k = 5,9$. В этом случае искусственный нейрон будет откликаться состоянием «0» при обнаружении нормальных данных и состоянием «1» при обнаружении равномерно распределенных данных. Значение порога квантователя подобрано так, чтобы вероятности ошибок первого и второго рода искусственного нейрона совпадали $P_1 \approx P_2 \approx P_{EE} \approx 0,326$. То есть одиночный искусственный хи-квадрат нейрон, настроенный на проверку гипотезы нормальности, позволяет принимать решения с низкой доверительной вероятностью 0,674.

Очевидно, что несколько статистических критериев, использованных параллельно, должны давать более достоверный результат в сравнении с одним критерием. Для их объединения могут быть использованы простейшие избыточные коды, способные обнаруживать и исправлять ошибки [6]. Очевидно также, что замена простейших кодов на более сложные конструкции должна приводить к росту достоверности принимаемых решений.

Дообучение искусственного хи-квадрат нейрона введением двух дополнительных функций прогноза уровня доверия

В случае снижения обнаруженного значения хи-квадрат отклика растет доверие к выходному состоянию «0». При росте значения хи-квадрат увеличивается доверие к выходному состоянию «1». Рассуждая формально, мы можем параллельно с бинарным квантователем использовать две непрерывные монотонные выходные функции хи-квадрат нейрона, выходное состояние которых оценивает уровень доверия к дискретному решению искус-

ственного нейрона. В итоге получается хи-квадрат нейрон с тремя выходами, отображенный на рис. 2.

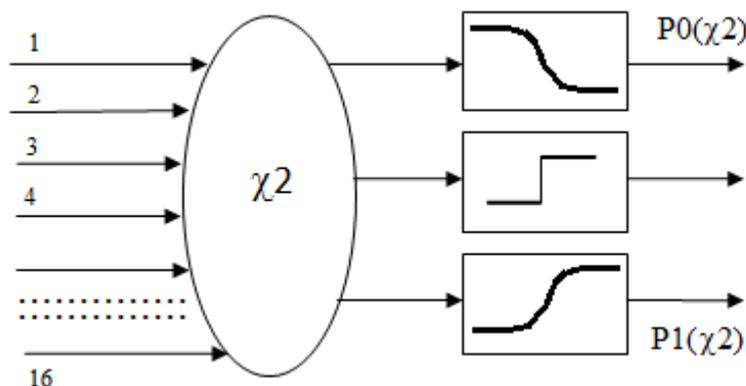


Рис. 2. Хи-квадрат нейрон с тремя выходами, два из которых непрерывные и дают оценку доверительной вероятности состояний третьего дискретного выхода

Очевидно, что дополнительные функции доверия $P_0(\chi^2)$ и $P_1(\chi^2)$ могут быть получены интегрированием плотностей распределения вероятностей, представленных на рис. 1. В частности, результаты интегрирования могут быть отображены таблицами доверительной вероятности, что является обычной формой для статистических справочников [7] и стандартов¹.

Ниже приведена таблица доверительных вероятностей для решения «0» хи-квадрат искусственного нейрона на малых выборках в 16 опытов (табл. 1).

Таблица 1

Таблица доверительных вероятностей для состояния «0» хи-квадрат нейрона

| | | | | | | | | | | | |
|----------|---|-------|-----|-----|-------|-------|-------|-------|-----|-------|-------------------|
| n | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| χ^2 | 0 | 1 | 1,5 | 2,3 | 2,9 | 4,4 | 6,1 | 7,3 | 9,4 | 11,3 | 22,2 |
| P_0 | 1 | 0,954 | 0,9 | 0,8 | 0,705 | 0,505 | 0,302 | 0,202 | 0,1 | 0,052 | $1 \cdot 10^{-3}$ |

Если хи-квадрат нейрон принимает иное решение «1», то таблица доверительных вероятностей к этому решению оказывается иной (табл. 2).

Таблица 2

Таблица доверительных вероятностей для состояния «1» хи-квадрат нейрона

| | | | | | | | | | | | |
|----------|-------------------|-------|-------|-------|-------|-------|-------|-----|-------|------|-------|
| n | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| χ^2 | 0 | 1 | 1,5 | 2,3 | 2,9 | 4,4 | 6,1 | 7,3 | 9,4 | 11,3 | 22,2 |
| P_1 | $1 \cdot 10^{-3}$ | 0,052 | 0,099 | 0,201 | 0,304 | 0,505 | 0,701 | 0,8 | 0,903 | 0,95 | 0,999 |

Так как плотности распределения вероятностей выходных состояний для нормальных данных и равномерных данных разные, таблицы доверительных вероятностей существенно различаются. При необходимости табли-

¹ Р 50.1.037–2002. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа χ^2 . М. : Госстандарт России, 2001. 140 с.

цы могут быть использованы для восстановления непрерывных функций вероятности через использование кусочно-линейной аппроксимации. На рис. 3 даны примеры восстановленных по таблицам непрерывных кусочно-линейных функций доверительной вероятности.

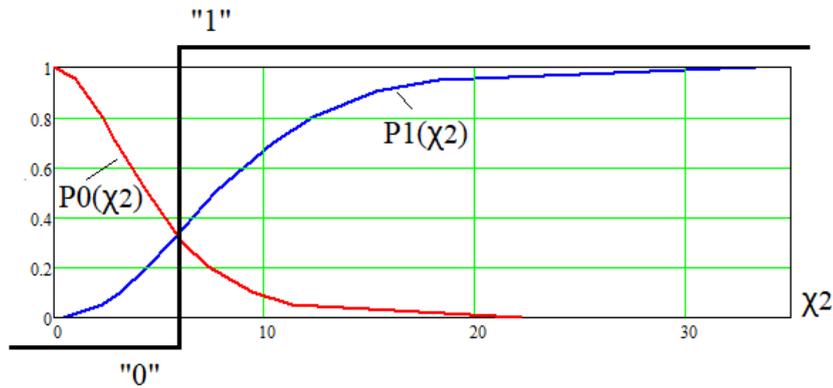


Рис. 3. Три выходных нелинейных функции модифицированного хи-квадрат нейрона для малых выборок в 16 опытов

Из рис. 3 видно, что функции доверия $P1(G)$ и $P0(G)$ не совпадают. В этом отношении многослойные сети искусственных нейронов Галушкина – Хинтона [8, 9] являются некоторым упрощением реально существующих статистических ситуаций. Многослойные сети искусственных нейронов не учитывают разницу функций доверительных вероятностей у двух разделяемых соседних классов.

Параллельное использование трех и более искусственных нейронов сетью с одним слоем

Очевидно, что при параллельном использовании трех разных статистических критериев мы можем обобщить их результаты [4–6] нейросетью. Выходной код нейросети будет обладать трехкратной избыточностью. Эта ситуация отражена на рис. 4.

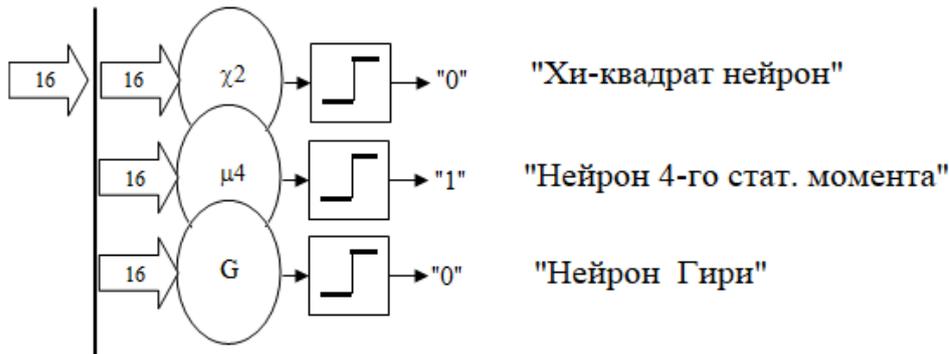


Рис. 4. Однослойная сеть из трех искусственных нейронов

Как итог – мы можем воспользоваться некоторым избыточным кодом, способным обнаруживать и исправлять ошибки. В простейшем случае это

может быть код, построенный «голосованием» по большинству состояний в его разрядах. Если мы наблюдаем код «000», то с высокой вероятностью можем считать входные данные малой выборки нормальными.

Если только в одном разряде код будет иметь состояние «1», то такие коды следует рассматривать как допустимые для выборок с нормально распределенными данными.

Синдромами обнаруженных ошибок следует считать коды с двумя состояниями «1» в их разрядах кода и код «111». Подобная кодовая конструкция дает среднее геометрическое значение вероятностей ошибок для нейронов 0,216. Корректировка синдромов допустимых ошибок позволяет трем, рассматриваемым нейронам снизить вероятности выходных ошибок до величины 0,165. Эта ситуация отображена на рис. 5 (верхняя линия, соединяющая две точки).

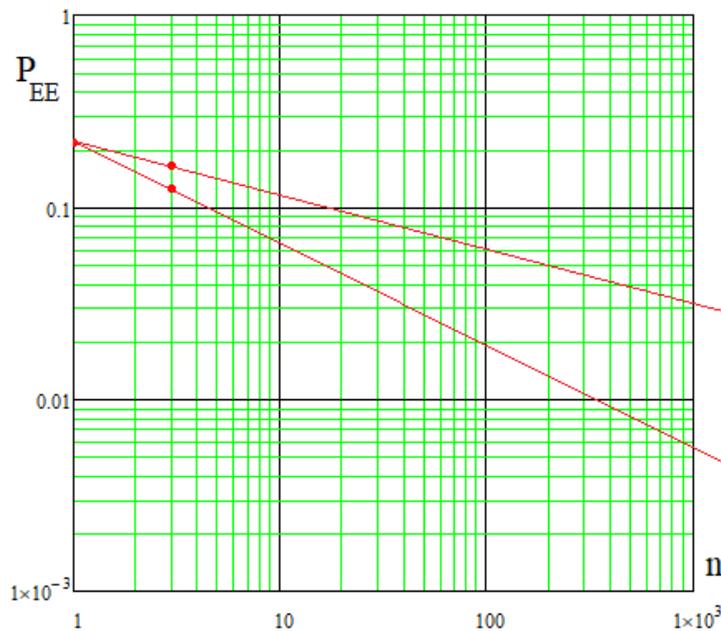


Рис. 5. Параллельное использование трех и более искусственных нейронов, эквивалентных трем классическим критериям

Из рис. 5 следует, что трех искусственных нейронов недостаточно для обеспечения доверия к принимаемым решениям 0,9. Однако, если число нейронов увеличить до 20, то доверительная вероятность на уровне 0,9 оказывается вполне достижима даже для малых выборок в 16 опытов.

Повышение эффективности нейросетевого обнаружения и исправления ошибок при учете достоверности состояний каждого из разрядов избыточного кода

Простейшие самокорректирующиеся коды «голосования по большинству» рассматривают код «110» как синдром обнаружения неисправимой ошибки. В нашем случае, когда искусственные нейроны дают не только выходной код, но и информацию о доверии к состоянию каждого из разрядов,

появляются новые возможности. Для учета дополнительной информации мы должны усложнить структуру нейросетевой обработки, перейдя к использованию двухслойных нейронных сетей (рис. 6).

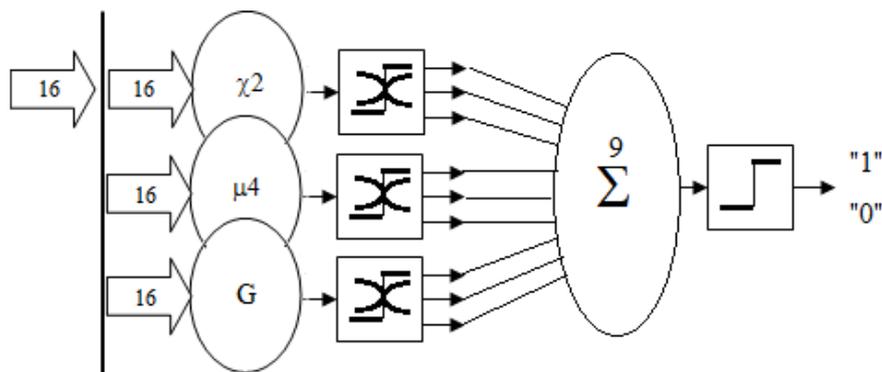


Рис. 6. Двухслойная сеть искусственных нейронов, учитывающая уровни доверия к состояниям персептронов первого слоя

В новой математической конструкции нейроны первого слоя имеют три выхода, а единственный нейрон второго слоя должен обобщать 9 выходных состояний нейронов первого слоя. Кроме синдрома ошибки «110» с выходов нейронов первого слоя мы получаем дополнительную информацию $\{P1(\chi^2), P1(\mu^4), P0(G)\}$. Ее учет позволяет либо подтвердить обнаружение неисправляемой ошибки, либо ее исправить. Появляется возможность подсчитать уровень доверия к первым двум разрядам $(P1(\chi^2) + P1(\mu^4))$ и сравнить его с доверием к третьему разряду $P0(G)$. В случае, если $(P1(\chi^2) + P1(\mu^4)) \geq P0(G)$, обнаружение синдрома неисправимой ошибки подтверждается. В ином случае, когда $(P1(\chi^2) + P1(\mu^4)) < P0(G)$, состояния первых двух разрядов рассматривается как ошибочное, подлежащее исправлению. В итоге мы получаем более сложную кодовую конструкцию, способную обнаруживать и исправлять большее число ошибок.

Численный эксперимент показал, что двухслойная нейросеть позволяет снизить уровень обнаруженных, но не подлежащих исправлению ошибок до вероятности 0,141. Эта ситуация отображена на рис. 5. Линейная экстраполяция результатов численного эксперимента (нижняя линия) позволяет ожидать доверительной вероятности 0,9 уже при использовании 5 искусственных нейронов первого слоя. Мы наблюдаем существенное снижение аппаратно-программных затрат на защиту приложений нейросетевого искусственного интеллекта путем использования малопотребляющих доверенных контроллеров (SIM-карт, RFID-карт, microSD-карт, USB-БиоТокенов, ПЛИС, DSP-контроллеров).

Список литературы

1. Хайкин С. Нейронные сети: полный курс. М. : Вильямс, 2006. 1104 с.
2. Рассел С., Норвиг П. Искусственный интеллект. Современный подход / пер. с англ. и под ред. К. А. Птицына. 2-е изд. М. : Вильямс, 2006. 1048 с.
3. Язов Ю. К., Волчихин В. И., Иванов А. И., Фунтиков В. А., Назаров И. Г. Нейросетевая защита персональных биометрических данных. М. : Радиотехника, 2012. 157 с.

4. Иванов А. И. Искусственные математические молекулы: повышение точности статистических оценок на малых выборках (программы на языке mathcad) : препринт. Пенза : Изд-во ПГУ, 2020. 36 с.
5. Иванов А. П., Иванов А. И., Малыгин А. Ю. [и др.]. Альбом из девяти классических статистических критериев для проверки гипотезы нормального или равномерного распределения данных малых выборок // Надежность и качество сложных систем. 2022. № 1. С. 20–29. doi: 10.21685/2307-4205-2022-1-3
6. Иванов А. И. Нейросетевой многокритериальный статистический анализ малых выборок : справочник. Пенза : Изд-во ПГУ, 2022. 160 с.
7. Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников. М. : ФИЗМАТЛИТ, 2006. 816 с.
8. Николенко С., Кудрин А., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей. СПб. : Питер, 2018. 480 с.
9. Аггарвал Ч. Нейронные сети и глубокое обучение. СПб. : Диалектика, 2020. 756 с.

References

1. Khaykin S. *Neyronnye seti: polnyy kurs = Neural networks: a complete course*. Moscow: Vil'yams, 2006:1104. (In Russ.)
2. Rassel S., Norvig P. *Iskusstvennyy intellekt. Sovremennyy podkhod / per. s angl. i pod red. K. A. Ptitsyna. 2-e izd. = Artificial intelligence. Modern approach / translated from English and edited by K.A. Ptitsyn. The 2nd edition*. Moscow: Vil'yams, 2006:1048. (In Russ.)
3. Yazov Yu.K., Volchikhin V.I., Ivanov A.I., Funtikov V.A., Nazarov I.G. *Neyrosetevaya zashchita personal'nykh biometricheskikh dannykh = Neural network protection of personal biometric data*. Moscow: Radiotekhnika, 2012:157. (In Russ.)
4. Ivanov A.I. *Iskusstvennye matematicheskie molekuly: povyshenie tochnosti statisticheskikh otsenok na malykh vyborkakh (programmy na yazyke mathcad): preprint = Artificial mathematical molecules: increasing the accuracy of statistical estimates on small samples (programs in mathcad): preprint*. Penza: Izd-vo PGU, 2020:36. (In Russ.)
5. Ivanov A.P., Ivanov A.I., Malygin A.Yu. et al. An album of nine classic statistical tests for testing the hypothesis of normal or uniform distribution of data from small samples. *Nadezhnost' i kachestvo slozhnykh system = Reliability and quality of complex systems*. 2022;(1):20–29. (In Russ.). doi: 10.21685/2307-4205-2022-1-3
6. Ivanov A.I. *Neyrosetevoy mnogokriterial'nyy statisticheskiy analiz malykh vyborok: spravochnik = Neural network multicriteria statistical analysis of small samples: reference book*. Penza: Izd-vo PGU, 2022:160. (In Russ.)
7. Kobzar' A.I. *Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnykh rabotnikov = Applied mathematical statistics. For engineers and scientists*. Moscow: FIZMATLIT, 2006:816. (In Russ.)
8. Nikolenko S., Kudrin A., Arkhangel'skaya E. *Glubokoe obuchenie. Pogruzhenie v mir neyronnykh setey = Deep learning. Dive into the world of neural networks*. Saint Petersburg: Piter, 2018:480. (In Russ.)
9. Aggarval Ch. *Neyronnye seti i glubokoe obuchenie = Neural networks and deep learning*. Saint Petersburg: Dialektika, 2020:756. (In Russ.)

Информация об авторах / Information about the authors

Владимир Иванович Волчихин

доктор технических наук, профессор,
президент Пензенского государственного
университета (Россия, г. Пенза,
ул. Красная, 40)

E-mail: cnit@pnzgu.ru

Vladimir I. Volchikhin

Doctor of engineering sciences, professor,
president of Penza State University
(40 Krasnaya street, Penza, Russia)

Александр Иванович Иванов

доктор технических наук, профессор,
научный консультант, Пензенский
научно-исследовательский
электротехнический институт (Россия,
г. Пенза, ул. Советская, 9)

E-mail: ivan@pniei.penza.ru

Aleksandr I. Ivanov

Doctor of engineering sciences, professor,
scientific adviser, Penza Scientific
Research Electrotechnical Institute
(9 Sovetskaya street, Penza, Russia)

Александр Викторович Безяев

кандидат технических наук, докторант,
Пензенский государственный
университет (Россия, г. Пенза,
ул. Красная, 40)

E-mail: tsib@pnzgu.ru

Aleksandr V. Bezyaev

Candidate of engineering sciences,
doctor's degree student, Penza State
University (40 Krasnaya street, Penza,
Russia)

Иван Александрович Филипов

преподаватель Военного учебного
центра, Пензенский государственный
университет (Россия, г. Пенза,
ул. Красная, 40)

E-mail: re.wolf@mail.ru

Ivan A. Filipov

Lecturer of the Military Training Center,
Penza State University (40 Krasnaya
street, Penza, Russia)

Авторы заявляют об отсутствии конфликта интересов / The authors declare no conflicts of interests.

Поступила в редакцию / Received 06.07.2023

Поступила после рецензирования и доработки / Revised 08.09.2023

Принята к публикации / Accepted 04.10.2023