УДК 519.24
doi:10.21685/2072-3059-2021-3-4

Расширение семейства статистических критериев Крамера – фон Мизеса за счет использования полиномов Лагерра при проверке гипотезы нормальности малых выборок

А. И. Иванов[1], А. П. Иванов[2], Е. Н. Куприянов[3]

[1]Пензенский научно-исследовательский электротехнический институт, Пенза, Россия
[2,3]Пензенский государственный университет, Пенза, Россия

[1]ivan@pniei.penza.ru, [2]ap_ivanov@pnzgu.ru, [3]evgnkupr@gmail.com

Аннотация. *Актуальность и цели.* Рассматривается проблема анализа малых выборок на примере синтеза новых статистических критериев, порождаемых классическим статистическим критерием Крамера – фон Мизеса. *Материалы и методы.* Предложено получать новые статистические критерии путем усиления результатов вычисления по классическому критерию умножением на четные ортогональные полиномы Лагерра. *Результаты и выводы.* Показано, что рассматриваемые новые статистические критерии дают решения, снижающие вероятности ошибок от трех до девяти раз для полиномов Лагерра 2, 4, 6-го порядков. С ростом порядка полинома Лагерра отмечается снижение вероятностей ошибок первого и второго рода новых статистических критериев. К семейству из двух ранее известных статистических критериев добавлено три новых статистических критерия, причем один из новых статистических критериев дает отклики, сильно коррелированные с откликами классического критерия Смиронова – Крамера – фон Мизеса.

Ключевые слова: анализ малых выборок, искусственные нейроны, статистические критерии проверки гипотезы нормальности, ортогональные полиномы Лагерра

Для цитирования: Иванов А. И., Иванов А. П., Куприянов Е. Н. Расширение семейства статистических критериев Крамера – фон Мизеса за счет использования полиномов Лагерра при поверке гипотезы нормальности малых выборок // Известия высших учебных заведений. Поволжский регион. Технические науки. 2021. № 3. С. 34–42. doi:10.21685/2072-3059-2021-3-4

Extension of statistical Cramer – von Mises tests using Laguerre polynomials while testing the hypothesis of small samples' normality

A.I. Ivanov[1], A.P. Ivanov[2], E.N. Kupriyanov[3]

[1]Penza Research Institute of Electrical Engineering, Penza, Russia
[2,3]Penza State University, Penza, Russia

[1]ivan@pniei.penza.ru, [2]ap_ivanov@pnzgu.ru, [3]evgnkupr@gmail.com

Abstract. *Background.* The research considers the problem of analyzing small samples using an example of the synthesis of new statistical tests generated by the classical statistical criterion of Cramer – von Mises. *Materials and methods.* It is proposed to obtain new statistical criteria by strengthening the calculation results according to the classical criterion by multiplying by even orthogonal Laguerre polynomials. *Results and conclusions.* It is shown that the considered new statistical criteria give solutions that reduce the error probabilities from three to nine times for Laguerre polynomials of 2, 4, 6 orders. With an increase

in the order of the Laguerre polynomial, a decrease in the probabilities of errors of the first and second kind of new statistical tests is noted. Three new statistical tests have been added to the family of two previously known statistical tests, with one of the new statistical tests giving responses strongly correlated with the responses of the classical Smironov – Cramer – von Mises test.

Keywords: analysis of small samples, artificial neurons, statistical criteria for testing the hypothesis of normality, orthogonal Laguerre polynomials

For citation: Ivanov A.I., Ivanov A.P., Kupriyanov E.N. Extension of statistical Cramer – von Mises tests using Laguerre polynomials while testing the hypothesis of small samples' normality. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki = University proceedings. Volga region. Engineering sciences.* 2021;(3):34–42. (In Russ.). doi:10.21685/2072-3059-2021-3-4

Введение

Одной из проблем статистической обработки реальных данных является то, что их мало. Так, средства биометрико-нейросетевой аутентификации приходится автоматически обучать на малой выборке из 20 примеров образа «Свой» (алгоритм ГОСТ Р 52633.5–2011). Та же самая ситуация возникает у врача, имеющего в своей личной практике всего 20 примеров историй некоторого заболевания. Аналогичные ситуации возникают у экономистов, биологов, бактериологов, материаловедов.

Классическая статистика позволяет делать достаточно достоверные выводы с доверительной вероятностью 0,97, пользуясь хи-квадрат критерием[1], при проверке гипотезы нормального распределения только на выборках в 200 и более опытов. То же самое относится и к иным статистическим критериям[2].

После хи-квадрат критерия наиболее распространенным на практике является критерий Крамера – фон Мизеса [1], созданный в 1928 г.:

$$\omega^2 = \sum_{i=0}^{n-1} \left\{ P\left(\{x_i\}, E(x), \sigma(x)\right) - \frac{i-0,5}{n} \right\}^2 , \qquad (1)$$

где $P\left(\{x_i\}, E(x), \sigma(x)\right)$ – вероятность появления состояния $\{x_i\}$ для нормального распределения данных, отсортированных по возрастанию $x_0 \le x_1 \le x_2 ..... \le x_n$ с математическим ожиданием $E(x)$ и стандартным отклонение малой выборки $\sigma(x)$.

Проблему малых выборок легко показать, выполнив соответствующий численный эксперимент. Результаты численного эксперимента приведены на рис. 1. Из рис. 1 видно, что малые выборки в 16 опытов с нормальным и равномерным законами распределения при обработке их по формуле (1) дают почти перекрывающиеся распределения значений. Формально мы можем обработку (1) рассматривать как некоторую процедуру накопления (обогащения) входных данных в квадратичном пространстве. Если на выходе обогатителя поставить квантователь, то все пространство входных состояний обога-

[1] Р 50.1.037–2002. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим : в 2 ч. Часть I. Критерии типа χ^2.

[2] Р 50.1.037–2002. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим т: в 2 ч. Часть II. Непараметрические критерии.

тителя будет разделено на два состояния. Состояние «0» будет соответствовать подтверждению гипотезы нормальности входных данных. Состояние «1» будет соответствовать отвержению гипотезы. При срабатывании квантователя в точке $\omega^2 = 0{,}119$ вероятности ошибок первого и второго рода совпадают и составляют $P_1 \approx P_2 \approx P_{EE} \approx 0{,}4$. То есть искусственный нейрон Крамера – фон Мизеса способен разделять малые выборки в 16 опытов с нормальным и равномерным законами распределения с доверительной вероятностью 0,6. Это недопустимо мало для практики.

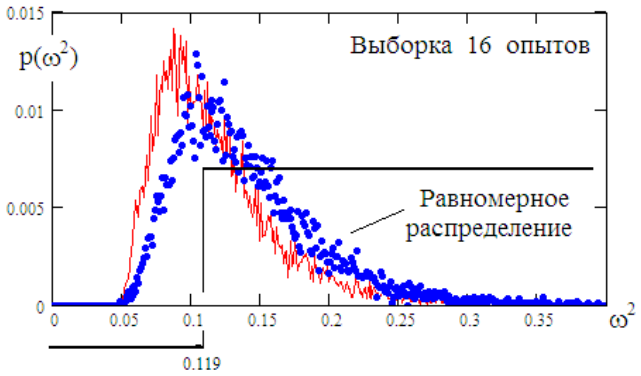

Рис. 1. Работа искусственного нейрона, эквивалентного классическому статистическому критерию Крамера – фон Мизеса, обеспечивающего $P_1 \approx P_2 \approx P_{EE} \approx 0{,}4$

В связи с катастрофически низкой доверительной вероятностью критерия Крамера – фон Мизеса на малых выборках в 1936 г. он был усовершенствован нашим соотечественником Н. В. Смирновым [1]:

$$S\omega^2 = \sum_{i=0}^{n-1} P\left[\left\{P\left(\{x_i\}, E(x), \sigma(x)\right) - \frac{i-0.5}{n}\right\}, E(x), \sigma(x)\right]. \qquad (2)$$

Результаты численного моделирования обогатителя данных в пространстве, деформированном ожидаемым нормальным законом (2), приведены на рис. 2.

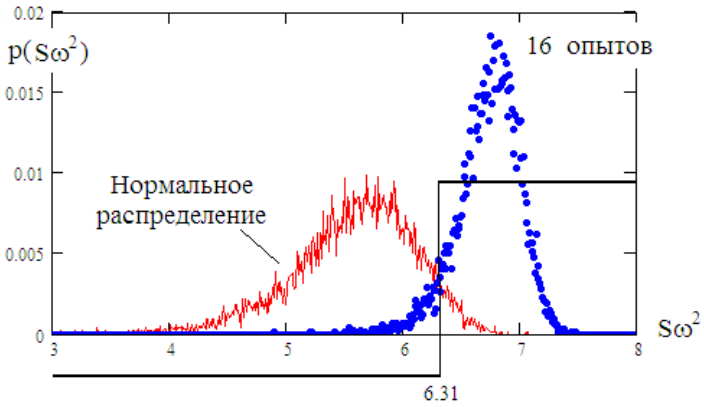

Рис. 2. Работа искусственного нейрона, эквивалентного классическому статистическому критерию Смирнова – Крамера – фон Мизеса, обеспечивающего $P_1 \approx P_2 \approx P_{EE} \approx 0{,}06$

Сравнивая рис. 1 и 2, мы видим существенное повышение линейной разделимости откликов накопителя данных (2). При настройке порога квантования $S\omega^2 = 6{,}31$ вероятности появления ошибок первого и второго рода составляют $P_1 \approx P_2 \approx P_{EE} \approx 0{,}06$. Это в 6,7 раза лучше, чем у исходного критерия Крамера – фон Мизеса.

Увеличение числа представителей семейства критериев Крамера – фон Мизеса

Н. В. Смирнов в 1936 г. [1] получил собственную более эффективную модификацию критерия Крамера – фон Мизеса, фактически выполнив некоторую модификацию пространства накопления (обогащения) входных данных. Последуем этим же путем, модифицировав пространство накопления данных полиномом Лагерра второго порядка для упорядоченных и центрированных данных:

$$L_2\omega = \sum_{i=0}^{n-1} \frac{\left\{x_i^{\,2} - 4x_i + 2\right\}}{2\left(\sigma(x)\right)^2 n} \times \left\{ P\left(\{x_i\}, E(x) = 0, \sigma(x)\right) - \frac{i - 0{,}5}{n} \right\}. \qquad (3)$$

Результаты численного моделирования новой модификации статистического критерия приведены на рис. 3.

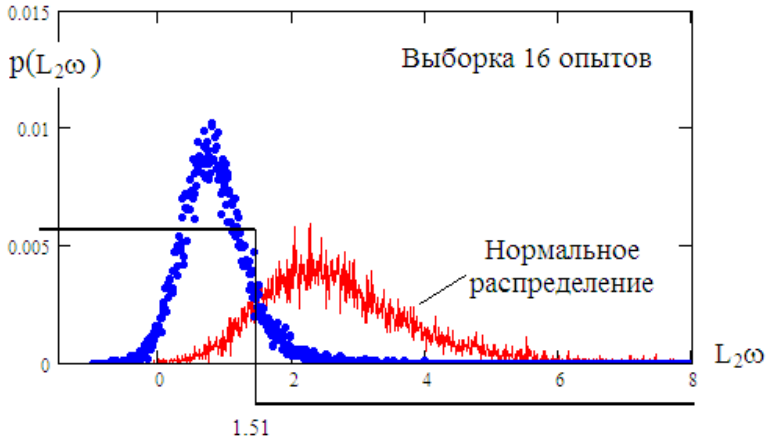

Рис. 3. Работа искусственного нейрона, эквивалентного модификации критерия Крамера – фон Мизеса полиномом Лагерра второго прядка, обеспечивающего $P_1 \approx P_2 \approx P_{EE} \approx 0{,}121$

Сравнивая рис. 1 и 3, мы видим, что по отношению к исходному критерию линейная разделимость данных существенно улучшилась. При пороге квантования $L_2\omega = 1{,}51$ вероятности ошибок первого и второго рода совпадают и составляют $P_1 \approx P_2 \approx P_{EE} \approx 0{,}121$. Это в 3,3 раза лучше, чем у исходной математической конструкции, но примерно в 2 раза хуже, чем у конструкции Смирнова – Крамера – фон Мизеса.

Продолжим модифицировать исходную математическую конструкцию, применив искажение пространства полиномом Лагерра четвертого порядка:

$$L_4\omega = \sum_{i=0}^{n-1} \frac{\left\{x_i^4 - 16x_i^3 + 72x_i^2 - 96x_i + 24\right\}}{24\left(\sigma(x)\right)^4 n} \times \left\{P\left(\{x_i\}, E(x) = 0, \sigma(x)\right) - \frac{i - 0{,}5}{n}\right\}. \quad (4)$$

Результаты моделирования новой математической конструкции приведены на рис. 4.

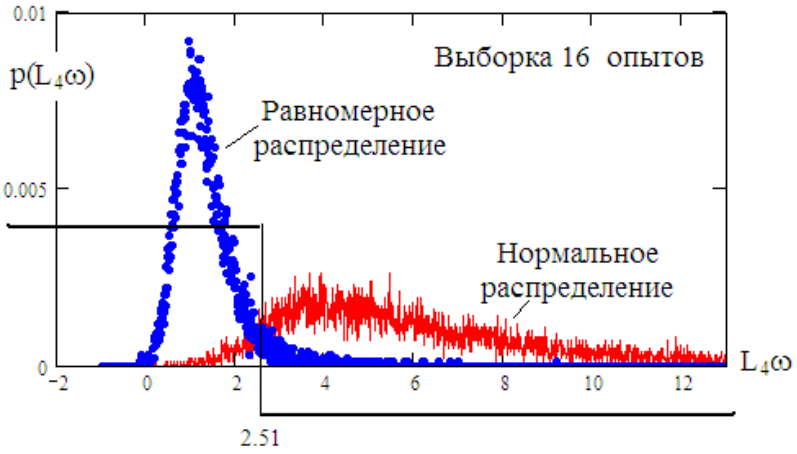

Рис. 4. Работа искусственного нейрона, эквивалентного модификации критерия Крамера – фон Мизеса полиномом Лагерра четвертого прядка, обеспечивающего $P_1 \approx P_2 \approx P_{EE} \approx 0{,}06$

Из рис. 4 видно, что при пороге квантования $L_4\omega = 2{,}51$ вероятности ошибок первого и второго рода совпадают и составляют $P_1 \approx P_2 \approx P_{EE} \approx 0{,}06$. Эта математическая конструкция по своим вероятностным показателям повторяет критерий Смирнова – Крамера – фон Мизеса. Однако в силу иного искажения пространства накопления данных результаты этих двух решающих правил разные. Корреляция между ними составляет $\mathrm{corr}(S\omega^2, L_4\omega) = 0{,}792$. Это означает, что эти два статистических критерия можно использовать совместно.

Деформация пространства накопления входных данных полиномом Лагерра шестого порядка дает еще один вариант статистического критерия:

$$L_6\omega = \sum_{i=0}^{n-1} \frac{L_6(x_i)}{\left(\sigma(x)\right)^6 n} \times \left\{P\left(\{x_i\}, E(x) = 0, \sigma(x)\right) - \frac{i - 0{,}5}{n}\right\}, \quad (5)$$

где $L_6(x_i)$ – полином Лагерра шестого порядка:

$$L_6(x_i) = \left\{x_i^6 - 36x_i^5 + 450x_i^4 - 2400x_i^3 + 5400x_i^2 - 4320x_i + 720\right\} \cdot \frac{1}{720}. \quad (6)$$

Работа эквивалентного этому критерию нейрона иллюстрируется данными, отображенными на рис. 5.

Из рис. 5 видно, что нейрон Крамера – фон Мизеса, модифицированный полиномом Лагерра шестого порядка, обладает рекордной мощностью при разделении нормальных и равномерных данных. При пороге $L_6\omega = 2{,}31$

вероятности появления ошибок первого и второго рода снижаются до величины $P_1 \approx P_2 \approx P_{EE} \approx 0{,}045$.

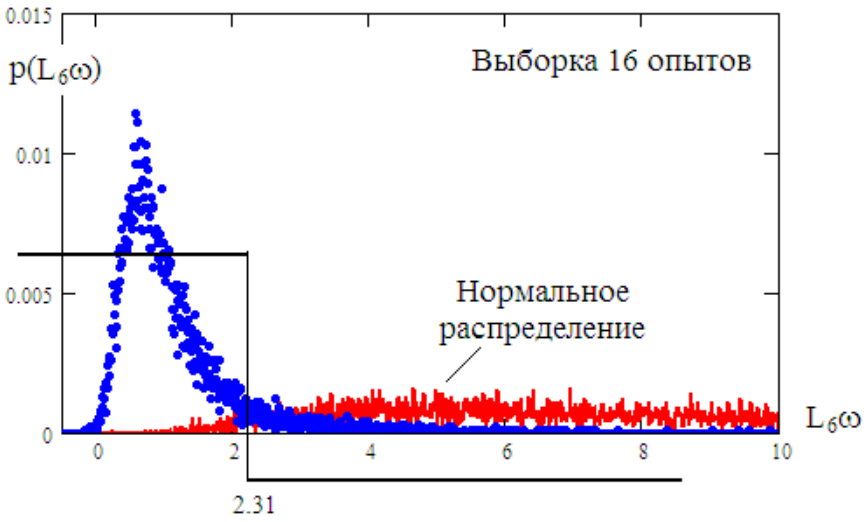

Рис. 5. Работа искусственного нейрона, эквивалентного модификации критерия Крамера – фон Мизеса полиномом Лагерра шестого прядка, обеспечивающего $P_1 \approx P_2 \approx P_{EE} \approx 0{,}045$

В итоге получается, что с ростом порядка полинома Лагерра мы наблюдаем монотонное снижение вероятностей появления ошибок первого и второго рода. Существует два пути дальнейшего повышения качества статистической обработки данных малых выборок. Во-первых, можно попытаться применять полиномы Лагерра более высоких порядков. При этом может происходить дальнейшее снижение вероятностей ошибок классификации. Во-вторых, можно использовать совместно все рассмотренные в статье статистические критерии.

Ожидаемые результаты при одновременном использовании пяти статистических критериев семейства Крамера – фон Мизеса

Одним из путей повышения достоверности статистических оценок малых выборок является использование сетей, состоящих из множества искусственных нейронов, если каждый из нейронов выполняет накопление (обогащение) входных данных в своем многомерном пространстве. Формально каждый из известных на текущий момент статистических критериев может рассматриваться как некоторый искусственный нейрон [2–4]. То есть 21 наиболее часто используемому статистическому критерию проверки гипотезы нормальности [1] можно поставить в соответствие сеть из 21 искусственного нейрона. Если эти все нейроны сети построены так же, как описанные в данной статье нейроны Крамера – фон Мизеса, то в наилучшем случае все нейроны дадут состояние «0». Выходной код нейросети будет иметь 21 нулевое состояние. В самом плохом случае, когда все нейроны отвергнут гипотезу нормальности, код будет состоять только из 21 единицы. Коды «000..00» и «111..11» будут встречаться достаточно редко. В основном мы

будем наблюдать коды, в которых будут встречаться как состояния «0», так и состояния «1».

Самым простым способом избавиться от неоднозначности (21 кратной кодовой избыточности) является подсчет состояний «0» и состояний «1» [5]. Если число состояний «0» больше числа состояний «1», то рассматриваемую малую выборку следует считать нормальной. В противном случае сеть из 21 нейрона должна обнаружить факт равномерного распределения данных малой выборки.

Очевидно, что тот же принцип устранения кодовой избыточности можно применить и для рассматриваемых в данной статье пяти нейронов Крамера – фон Мизеса. Качество работы такой сети можно оценить, проведя процедуру ее симметризации [6]. Для этой цели следует вычислить среднее геометрическое равновероятных ошибок каждого из нейронов, оно составит 0,095. Также следует вычислить среднее значение модулей коэффициентов вне диагонали корреляционной матрицы, выходов нейросети, составляющее 0,465.

Данные о векторе равновероятных ошибок и о матрице коэффициентов корреляции отражены на рис. 6.

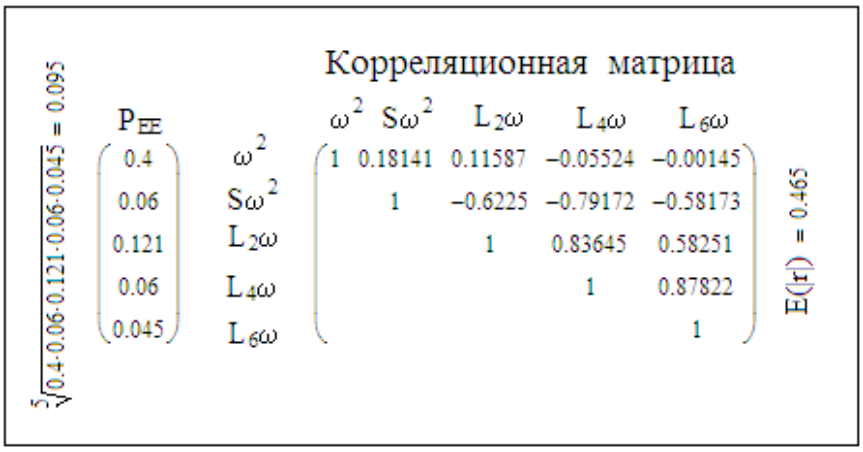

Рис. 6. Данные численного моделирования о векторе вероятностей появления ошибок P_{EE} и данные корреляционной матрицы для пяти рассмотренных нейронов

Проведенный численный эксперимент показал, что для пяти рассмотренных в данной статье нейронов с использованием простейшего кода обнаружения и исправления ошибок по большинству состояний позволяет поднять доверительную вероятность с 0,965 (когда используется один самый мощный нейрон $L_6\omega$) до доверительной вероятности 0,981. И та и другая доверительные вероятности являются приемлемыми для практики.

Заключение

Крайне важным обстоятельством является так же то, что остается возможность увеличения числа нейронов семейства Крамера – фон Мизеса за счет применения полиномов Лагерра более высоких порядков. Кроме того, при вычислениях могут быть использованы более сложные коды обнаруже-

ния и исправления ошибок для устранения изначально заложенной кодовой избыточности. Все это позволяет надеяться на то, что в ближайшем будущем применение достаточного числа искусственных нейронов позволит существенно увеличить достоверность статистических решений, принимаемых на малых выборках.

Еще одним направлением продолжения исследований является возможность расширения других классов статистических критериев. Так, статистические критерии семейства среднего геометрического и семейства среднего гармонического [2] имеют от двух до трех вариантов реализации. Наблюдается аналогия с семейством статистических критериев семейства Крамера – фон Мизеса. Так как полиномы Лагерра позволили расширить семейство критериев Крамера – фон Мизеса, предположительно они же могут позволить расширить многообразие и иных статистических критериев.

Список литературы

1. Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников. М. : Физматлит, 2006. 816 с.
2. Иванов А. И., Банных А. Г., Куприянов Е. Н. [и др.]. Коллекция искусственных нейронов, эквивалентных статистическим критериям, для их совместного применения при проверке гипотезы нормальности малых выборок биометрических данных // Безопасность информационных технологий : труды I Всерос. науч.-техн. конф. Пенза : Изд-во ПГУ, 2019. С. 163–172.
3. Волчихин В. И., Иванов А И., Безяев А. В., Куприянов Е. Н. Нейросетевой анализ нормальности малых выборок биометрических данных с использованием хи-квадрат критерия и критериев Андерсона – Дарлинга // Инженерные технологии и системы. 2019. Т. 29, № 2. С. 205–217. doi:10.15507/2658-4123.029.201902.205-217
4. Иванов А. И., Банных А. Г., Безяев А. В. Искусственные молекулы, собранные из искусственных нейронов, воспроизводящих работу классических статистических критериев // Вестник Пермского университета. Серия: Математика. Механика. Информатика. 2020. № 1 (48). С. 26–32. doi:10.17072/1993-0550-2020-1-26-32.
5. Безяев А. В. Биометрико-нейросетевая аутентификация: обнаружение и исправление ошибок в длинных кодах без накладных расходов на избыточность : препринт. Пенза : Изд-во ПГУ, 2020. 40 с.
6. Иванов А. И., Банных А. Г., Серикова Ю. И. Учет влияния корреляционных связей через их усреднение по модулю при нейросетевом обобщении статистических критериев для малых выборок // Надежность. 2020. № 2. С. 28–34. doi: 10.21683/1729-2646-2020-20-2-28-34

References

1. Kobzar' A.I. *Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnykh rabotnikov = Applied mathematical statistics. For engineers and scientists*. Moscow: Fizmatlit, 2006:816. (In Russ.)
2. Ivanov A.I., Bannykh A.G., Kupriyanov E.N. [et al.]. Collection of artificial neurons, equivalent to statistical criteria, for their joint use in testing the hypothesis of small samples' normality of biometric data. *Bezopasnost' informatsionnykh tekhnologiy: trudy I Vseros. nauch.-tekhn. konf. = Information technology security: proceedings of the 1st All=Russian scientific and engineering conference*. Penza: Izd-vo PGU, 2019:163–172. (In Russ.)
3. Volchikhin V.I., Ivanov AI., Bezyaev A.V., Kupriyanov E.N. Neural network analysis of normality of small samples of biometric data using the chi-square test and Anderson-Darling tests. *Inzhenernye tekhnologii i sistemy = Engineering technologies and systems*. 2019;29(2):205–217. (In Russ.). doi:10.15507/2658-4123.029.201902.205-217

4. Ivanov A.I., Bannykh A.G., Bezyaev A.V. Artificial molecules assembled from artificial neurons that reproduce the work of classical statistical criteria. *Vestnik Permskogo universiteta. Seriya: Matematika. Mekhanika. Informatika = Bulletin of Perm University. Series: Mathematics. Mechanics. Informatics*. 2020;(1):26–32. (In Russ.). doi:10.17072/1993-0550-2020-1-26-32
5. Bezyaev A.V. *Biometriko-neyrosetevaya autentifikatsiya: obnaruzhenie i ispravlenie oshibok v dlinnykh kodakh bez nakladnykh raskhodov na izbytochnost': preprint = Biometrical neural network authentication: detecting and correcting errors in long codes without the overhead of redundancy: preprint*. Penza: Izd-vo PGU, 2020:40. (In Russ.)
6. Ivanov A.I., Bannykh A.G., Serikova Yu.I. Taking into account the influence of correlations through their averaging in modulus in a neural network generalization of statistical criteria for small samples. *Nadezhnost' = Reliability*. 2020;(2):28–34. (In Russ.). doi: 10.21683/1729-2646-2020-20-2-28-34

Информация об авторах / Information about the authors

Александр Иванович Иванов
доктор технических наук, доцент, 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, associate professor, scientific adviser, Penza Research Institute of Electrical Engineering (9 Sovetskaya street, Penza, Russia)

Алексей Петрович Иванов
кандидат технических наук, доцент, заведующий кафедрой технических средств информационной безопасности, Пензенский государственный университет (Россия, г. Пенза, ул. Красная, 40)

E-mail: ap_ivanov@pnzgu.ru

Aleksey P. Ivanov
Candidate of engineering sciences, associate professor, head of the sub-department of technical means of information security, Penza State University (40 Krasnaya street, Penza, Russia)

Евгений Николаевич Куприянов
аспирант, Пензенский государственный университет (Россия, г. Пенза, ул. Красная, 40)

E-mail: evgnkupr@gmail.com

Evgeniy N. Kupriyanov
Postgraudate student, 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 09.08.2021

Поступила после рецензирования и доработки / Revised 18.09.2021

Принята к публикации / Accepted 07.10.2021