УДК 519.24
doi:10.21685/2072-3059-2021-3-7

Новый статистический критерий большой мощности, полученный дифференцированием случайных данных малой выборки

А. И. Иванов[1], А. Ю. Малыгин[2], С. А. Полковникова[3]

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

[1]ivan@pniei.penza.ru, [2]mal890@yandex.ru, [3]1996svetlanaserikova@gmail.com

Аннотация. *Актуальность и цели*. Рассматривается проблема статистического анализа малых выборок путем синтеза новых статистических критериев. *Материалы и методы.* Предложено перед расчетами выполнить операцию дифференцирования случайных данных малой выборки. *Результаты*. Вероятность появления ошибок первого и второго рода классического хи-квадрат критерия при малой выборке в 16 опытов составляет 0,33, что недопустимо для практики. Новый статистический критерий при тех же условиях снижает вероятность ошибок до 0,075, что уже вполне допустимо для ряда приложений нейросетевой биометрии. *Выводы.* Обычно считается, что операция дифференцирования данных случайной выборки должна приводить к существенной утрате устойчивости вычислений. В данной работе показана ситуация, являющаяся исключением из общего правила. Синтезированный статистический критерий имеет существенно меньшую вероятность ошибок по сравнению с классическим хи-квадрат критерием при решении задачи нейросетевого разделения нормальных данных и данных с равномерным распределением.

Ключевые слова: статистический анализ малых выборок, проверка гипотезы нормальности, хи-квадрат критерий, дифференцирование случайных данных малых выборок, искусственные нейроны

Для цитирования: Иванов А. И., Малыгин А. Ю., Полковникова С. А. Новый статистический критерий большой мощности, полученный дифференцированием случайных данных малой выборки // Известия высших учебных заведений. Поволжский регион. Технические науки. 2021. № 3. С. 67–74. doi:10.21685/2072-3059-2021-3-7

New statistical high power test, obtained by differentiating random data of a small sample

A.I. Ivanov[1], A.Yu. Malygin[2], S.A. Polkovnikova[3]

[1]Penza Scientific Research Electrotechnical Institute, Penza, Russia
[2,3]Penza State University, Penza, Russia

[1]ivan@pniei.penza.ru, [2]mal890@yandex.ru, [3]1996svetlanaserikova@gmail.com

Abstract. *Background*. The research considers the problem of statistical analysis of small samples by synthesizing new statistical criteria. *Materials and methods.* It is proposed to perform the operation of differentiation of random data of a small sample before calculations. *Results*. The probability of errors of the first and second kind of classical hi-squared criterion with a small sample of 16 experiments is 0.33, which is unacceptable for practice. The new statistical criterion under the same conditions reduces the probability of errors to 0.075, which is already quite acceptable for a number of neural network biometrics applications. *Conclusions*. It is generally believed that the operation of differentiating random

sample data should lead to a significant loss of stability of calculations. This article shows a situation that is an exception to the general rule. The synthesized statistical criterion has a significantly lower probability of errors compared to the classical chi-squared criterion when solving the problem of neural network separation of normal data and data with a uniform distribution.

Keywords: statistical analysis of small samples, testing of the normality hypothesis, xi-square criteria, differentiation of random data of small samples, artificial neurons

For citation: Ivanov A.I., Malygin A.Yu., Polkovnikova S.A. New statistical high power test, obtained by differentiating random data of a small sample. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki = University proceedings. Volga region. Engineering sciences.* 2021;(3):67–74. (In Russ.). doi:10.21685/2072-3059-2021-3-7

Использование хи-квадрат критерия Пирсона для проверки гипотезы нормального распределения малых выборок

Современная математическая статистика в том виде, в каком мы ее знаем, опирается на хи-квадрат критерий Пирсона, построенный им в 1900 г. Эта математическая конструкция оказалась популярной из-за того, что она проста и долгое время ей не было равных по эффективности [1, 2]. К сожалению, на малых выборках хи-квадрат критерий плохо работает. Убедиться в этом можно, воспользовавшись программным обеспечением на языке MathCAD, воспроизводящем хи-квадрат критерий для нормальных и равномерных данных малой выборки в 16 опытов (программы приведены на рис. 1, результаты имитационного моделирования приведены на рис. 2).

$\chi 2 :=$
$x \leftarrow sort(rnorm(16,0,1))$
$m \leftarrow mean(x)$
$\sigma \leftarrow stdev(x)$
for $i \in 0..6$
$int_i \leftarrow \frac{(x_{15}-x_0)\cdot i}{6} + x_0$
$P_i \leftarrow pnorm\left[\left[\frac{(x_{15}-x_0)\cdot i}{6} + x_0\right], m, \sigma\right]$
$n \leftarrow hist(int, x)$
$\chi 2 \leftarrow \sum_{i=0}^{5} \frac{\left[n_i - 16\cdot(P_{i+1}-P_i)\right]^2}{16\cdot(P_{i+1}-P_i)}$
$\chi 2$

$\chi 2r :=$
$x \leftarrow sort(runif(16,-3,3))$
$m \leftarrow mean(x)$
$\sigma \leftarrow stdev(x)$
for $i \in 0..6$
$int_i \leftarrow \frac{(x_{15}-x_0)\cdot i}{6} + x_0$
$P_i \leftarrow pnorm\left[\left[\frac{(x_{15}-x_0)\cdot i}{6} + x_0\right], m, \sigma\right]$
$n \leftarrow hist(int, x)$
$\chi 2 \leftarrow \sum_{i=0}^{5} \frac{\left[n_i - 16\cdot(P_{i+1}-P_i)\right]^2}{16\cdot(P_{i+1}-P_i)}$
$\chi 2$

Рис. 1. Программная реализация хи-квадрат критерия для нормальных данных (левая часть) и равномерных данных (правая часть)

Очевидным является то, что для малых выборок вероятности ошибок первого и второго рода велики $P_1 = P_2 = P_{EE} = 0{,}330$. В связи с этим по стандартным рекомендациям [2] для приемлемых значений доверительных вероятностей критерий хи-квадрат должен применяться для выборок в 200 и бо-

лее опытов. Это условие невыполнимо для нейросетевой биометрии. Обычно автоматическое обучение нейросетей выполняют по алгоритму ГОСТ Р 52633.5, используя от 16 до 20 примеров биометрического образа «Свой». Перед проведением обучения необходимо убедиться в отсутствии в обучающей выборе грубых ошибок по критерию нормальности собранных данных по каждому из сотен биометрических параметров.

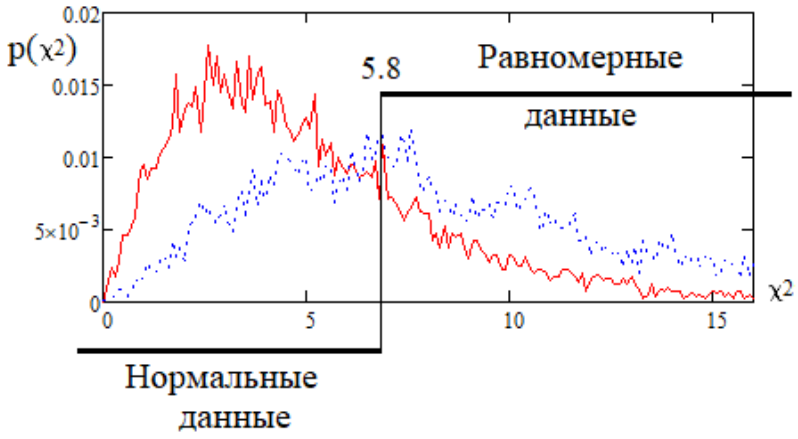

Рис. 2. Пример плохой линейной разделимости искусственным нейроном выходных состояний хи-квадрат критерия Пирсона для малых выборок в 16 опытов

Очевидным является то, что одного статистического критерия хи-квадрат для решения этой задачи недостаточно.

Нейросетевое обобщение множества известных статистических критериев

Одним из путей решения поставленной задачи является многокритериальный статистический анализ данных малых выборок. В частности, за прошлый век математической общественностью был создан 21 статистический критерий [1] для проверки гипотезы нормальности. В этом контексте представляется перспективным подход, по которому каждый их известных статистических критериев представляется в виде эквивалентного ему искусственного нейрона [3–5]. В частности, на рис. 2 представлена ситуация, когда входные данные малой выборки, обогащенные по Пирсону, подвержены квантованию с порогом $\chi^2 = 5{,}8$. При этом попадание данных в интервал от 0,0 до 5,8 будет соответствовать отклику искусственного нейрона «0», а наблюдение выходных значений искусственного нейрона $\chi^2 > 5{,}8$ будет соответствовать выходному состоянию «1».

Очевидно, что подобные искусственные нейроны могут быть получены и для других известных статистических критериев. То есть можно получить нейросеть, состоящую из 21 искусственного нейрона, соответствующих 21-му классическому статистическому критерию.

Очевидно, что при кодировке всех искусственных нейронов для нормальных данных «0» и равномерных данных «1» получаются выходные коды с высокой избыточностью. Эта избыточность может быть свернута каким-либо кодом, способным к обнаружению и исправлению ошибок [6]. Самым

простым свертыванием кодов является подсчет в них числа состояний «0» и «1». Решение принимается по большинству обнаруженных состояний.

Очевидным является то, что свертывающие коды (коды с обнаружением и исправлением ошибок) будут работать тем лучше, чем меньше будут корреляционные связи между разрядами кодов. Также желательно снижать вероятность ошибок в каждом разряде кода. К сожалению, 21 статистический критерий, созданный в прошлом веке, имеет сильные взаимные корреляционные связи и относительно низкую мощность. В связи с этим возникает задача синтеза новых статистических критериев, например критериев среднего геометрического и среднего гармонического [7–9]. На данный момент синтезировано порядка 7 новых статистических критериев, т.е. вместе с классическими критериями прошлого века мы можем иметь нейросеть с 28 выходами (рис. 3).

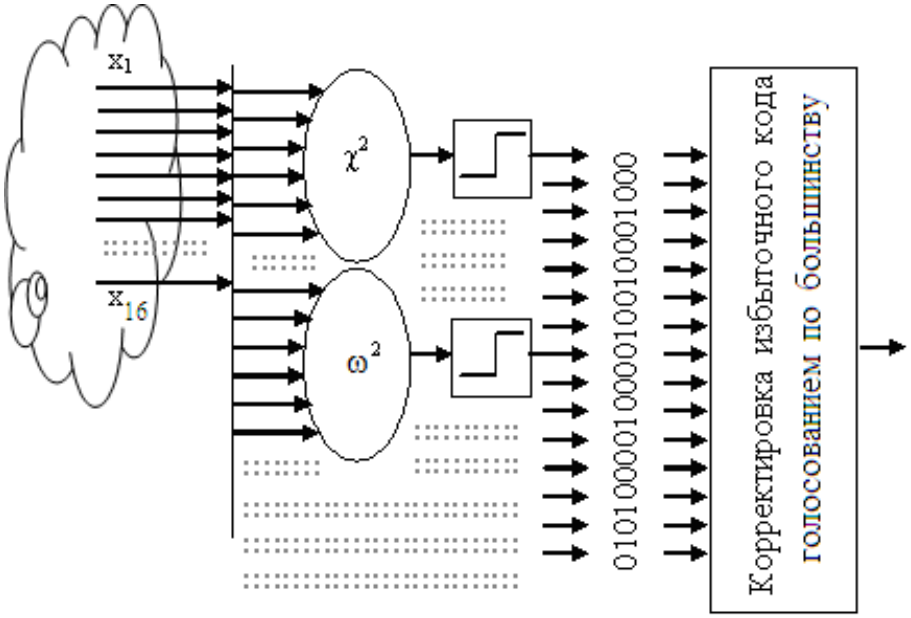

Рис. 3. Нейросетевое обобщение множества известных
на сегодня статистических критериев

Синтез нового статистического критерия высокой мощности через использование дифференцирования входных упорядоченных случайных данных малой выборки и последующего их взвешенного усреднения

Одним из перспективных способов синтеза новых статистических критериев является дифференцирование входных случайных данных. Формально дифференцирование может быть выполнено через упорядочивание данных малой выборки $x \leftarrow \mathrm{sort}(x)$ с последующим вычислением разности между соседними отсчетами. При этом должно быть выполнено нормирование данных делением на расстояние между крайними отсчетами. Фактически предлагается математическое ожидание разностей, упорядоченных по возрастанию данных. Пример программной реализации такого статистического функционала приведен на рис. 4.

$$D := \left| \begin{array}{l} x \leftarrow \text{sort}(\text{rnorm}(16,0,1)) \\ m \leftarrow \text{mean}(x) \\ \sigma \leftarrow \text{stdev}(x) \\ D \leftarrow \sum_{i=0}^{14} \frac{(x_{i+1} - x_i) \cdot \text{dnorm}(x_i, m, \sigma)}{x_{15} - x_0} \\ D \end{array} \right. \qquad Dr := \left| \begin{array}{l} x \leftarrow \text{sort}(\text{runif}(16,-3,3)) \\ m \leftarrow \text{mean}(x) \\ \sigma \leftarrow \text{stdev}(x) \\ D \leftarrow \sum_{i=0}^{14} \frac{(x_{i+1} - x_i) \cdot \text{dnorm}(x_i, m, \sigma)}{x_{15} - x_0} \\ D \end{array} \right.$$

Рис. 4. Программная реализация критерия среднего значения разностей упорядоченной выборки для нормальных данных (левая часть) и равномерных данных (правая часть)

Результаты численного моделирования нового статистического критерия приведены на рис. 5.

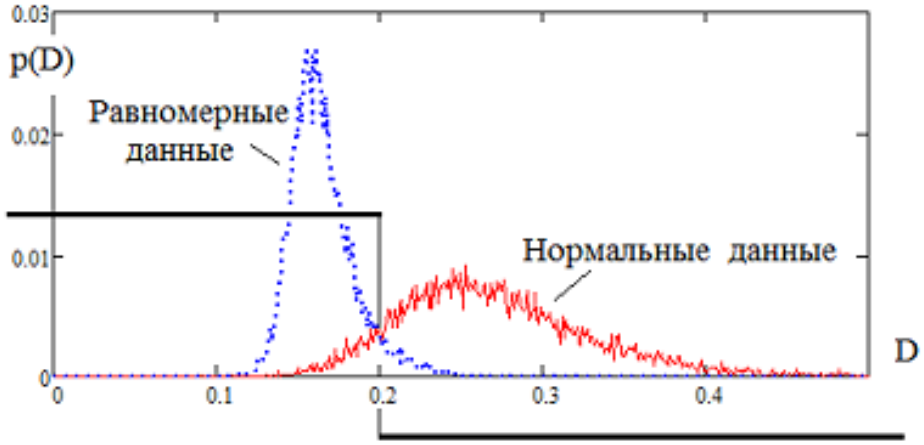

Рис. 5. Программная реализация дифференциального критерия *D* для нормальных данных (левая часть) и равномерных данных (правая часть)

Сравнивая данные, приведенные на рис. 2 и 5, видим значительный рост мощности нового критерия по сравнению с хи-квадрат критерием. Новый критерий при выходном квантовании в точке $D = 0{,}2$ дает равные вероятности ошибок первого и второго рода на уровне $P_1 = P_2 = P_{EE} = 0{,}075$, что эквивалентно росту мощности нового критерия в 4,4 раза по отношению к классическому хи-квадрат критерию.

Заключение

За пять лет исследований с 2016 г. [9] по настоящее время исследователями Пензенского государственного университета было создано 8 новых статистических критериев, включая критерий, описанный в этой статье. Предварительные прогнозы показывают, что для обеспечения доверительной вероятности 0,99 нейронная сеть должна состоять примерно из 60 искусственных нейронов. Примерно половина искусственных нейронов уже создана, предположительно через 10–15 лет удастся синтезировать вторую половину недостающих статистических критериев.

Подобный прогноз строился на предположении свертывания избыточных кодов по большинству состояний «0» или «1».

В случае использования более сложных кодовых конструкций [6] либо применения многослойного нейросетевого корректора (свертывателя данных) возможно существенное снижение числа необходимых статистических критериев.

Список литературы

1. Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников. М. : Физматлит, 2006. 816 с.
2. Р 50.1.037–2002. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим : в 2 ч. Часть I. Критерии типа χ2. Госстандарт России. Москва, 2001. 140 с.
3. 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.
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.
5. 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 : сб. науч. ст. по материалам I Всерос. науч.-техн. конф. Пенза, 2019. С. 156–164.
6. 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.
7. Лукин В. С. Сравнение мощности обычной и логарифмической форм статистических критериев среднего гармонического при использовании для проверки гипотезы нормального распределения данных малой выборки // Известия высших учебных заведений. Поволжский регион. Технические науки. 2020. № 4. С. 19–26.
8. Иванов А. И., Перфилов К. А, Лукин В. С. Нейросетевое обобщение семейства статистических критериев среднего геометрического и среднего гармонического для прецизионного анализа малых выборок биометрических данных // Информационно-управляющие телекоммуникационные системы, средства поражения и их техническое обеспечение : сб. науч. ст. Всерос. науч.-техн. конф. / под общ. ред. В. С. Безяева. Пенза : НПП «Рубин», 2019. С. 50–63.
9. Иванов А. И., Перфилов К. А., Малыгина Е. А. Оценка качества малых выборок биометрических данных с использованием дифференциального варианта статистического критерия среднего геометрического // Вестник Сибирского государственного аэрокосмического университета. 2016. № 4 (17). С. 864–871.

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. R 50.1.037–2002. *Rekomendatsii po standartizatsii. Prikladnaya statistika. Pravila proverki soglasiya opytnogo raspredeleniya s teoreticheskim: v 2 ch. Chast' I. Kriterii tipa χ2. Gosstandart Rossii = Recommendations for standardization. Applied statistics. Rules for checking the agreement of the experimental distribution with the theoretical: in 2 chapters. Chapter 1. χ2 criteria. State Standart of the Russian Federation*. Moscow, 2001:140. (In Russ.)

3. Ivanov A.I. *Iskusstvennye matematicheskie molekuly: povyshenie tochnosti statisticheskikh otsenok na malykh vyborkakh (programmy na yazyke MathCAD): preprint = Artificial mathematical molecules: improving the accuracy of statistical estimates on small samples (MathCAD programs): preprint*. Penza: Izd-vo PGU, 2020:36. (In Russ.)
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.)
5. Ivanov A.I., Bannykh A.G., Kupriyanov E.N. [et al.]. A collection of artificial neurons equivalent to statistical criteria for their combined use in testing the hypothesis of normality of small samples of biometric data. *Bezopasnost' informatsionnykh tekhnologiy: sb. nauch. st. po materialam I Vseros. nauch.-tekhn. konf. = Information technology security: proceedings of the 1st All-Russian scientific and engineering conference*. Penza, 2019:156–164. (In Russ.)
6. 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.)
7. Lukin V.S. Comparison of the power of ordinary and logarithmic forms of statistical tests of the harmonic mean when used to test the hypothesis of normal distribution of small sample data. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki = University proceedings. Volga region. Engineering sciences*. 2020;(4):19–26. (In Russ.)
8. Ivanov A.I., Perfilov K.A, Lukin V.S. Neural network generalization of a family of statistical criteria for geometric mean and harmonic mean for precision analysis of small samples of biometric data. *Informatsionno-upravlyayushchie telekommunikatsionnye sistemy, sredstva porazheniya i ikh tekhnicheskoe obespechenie: sb. nauch. st. Vseross. nauch.-tekhn. konf. = Information and control telecommunication systems, weapons and their technical support: proceedings of the All-Russian scientific and engineering conference.* Penza: NPP «Rubin», 2019:50–63. (In Russ.)
9. Ivanov A.I., Perfilov K.A., Malygina E.A. Evaluation of the quality of small samples of biometric data using the differential version of the statistical criterion of the geometric mean. *Vestnik Sibirskogo gosudarstvennogo aerokosmicheskogo universiteta = Bulletin of Siberian State Aerospace University*. 2016;(4):864–871. (In Russ.)

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 Scientific Research Electrotechnical Institute (9 Sovetskaya street, Penza, Russia)

Александр Юрьевич Малыгин
доктор технических наук, профессор, начальник межотраслевой лаборатории тестирования биометрических устройств и технологий, Пензенский государственный университет (Россия, г. Пенза, ул. Красная, 40)

E-mail: mal890@yandex.ru

Aleksandr Yu. Malygin
Doctor of engineering sciences, professor, head of the Intersectoral testing laboratory of biometric devices and technologies, Penza State University (40 Krasnaya street, Penza, Russia)

Светлана Андреевна Полковникова
аспирант, Пензенский государственный университет (Россия, г. Пенза, ул. Красная, 40)

Svetlana A. Polkovnikova
Postgraduate student, Penza State University (40 Krasnaya street, Penza, Russia)

E-mail: 1996svetlanaserikova@gmail.com

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

Поступила в редакцию / Received 04.09.2021

Поступила после рецензирования и доработки / Revised 01.10.2021

Принята к публикации / Accepted 20.10.2021