УДК 519.24; 53; 57.017
doi: 10.21685/2072-3059-2024-3-5

Дифференциальный критерий Джини для нейросетевого анализа малых выборок биометрических данных

С. А. Гужова

Пензенский государственный университет, Пенза, Россия
1996svetlanaserikova@gmail.com

Аннотация. *Актуальность и цели.* Обучение искусственных нейронных сетей в соответствии с алгоритмами по распознаванию биометрических образов, описанными в ГОСТ Р 52633.5–2011, должно выполняться на малых выборках. Рассматриваются выборки, состоящие из 16 и 64 опытов. *Материалы и методы.* Показано, что для выборок в 16 опытов хи-квадрат критерий 1900 г. дает недопустимые вероятности ошибок. Классический критерий Джини 1941 г. дает вероятности ошибок на 29 % выше хи-квадрат критерия. *Результаты и выводы.* Использование предложенного нового дифференциального варианта критерия Джини позволяет на выборках в 16 опытов получить результаты примерно в 9 раз лучше, чем у хи-квадрат критерия на выборках того же объема. Рассмотрен вариант нейросетевого использования классических и вновь созданных на их базе критериев. При этом нейросетевое использование 9 новых статистических критериев позволяет снизить вероятности ошибок первого и второго рода до значения 0,031. Если в этой группе заменить наихудший критерий на дифференциальный критерий Джини, то вероятность появления ошибок удается снизить до значения 0,027.

Ключевые слова: быстрые алгоритмы обучения нейросетей, малые выборки, статистические критерии проверки гипотезы нормальности и равномерности, бинарные нейрокритерии, искусственные нейроны, повышение точности оценок за счет устранения кодовой избыточности

Для цитирования: Гужова С. А. Дифференциальный критерий Джини для нейросетевого анализа малых выборок биометрических данных // Известия высших учебных заведений. Поволжский регион. Технические науки. 2024. № 3. С. 47–54. doi: 10.21685/2072-3059-2024-3-5

The Gini differential test for neural network analysis of small biometric data samples

S.A. Guzhova

Penza State University, Penza, Russia
1996svetlanaserikova@gmail.com

Abstract. *Backgroud.* Training of artificial neural networks in accordance with the algorithms for recognizing biometric patterns described in GOST R 52633.5-2011 should be performed on small samples. The research considers samples consisting of 16 and 64 experiments. *Materials and methods.* Shown for samples of 16 experiments, the chi-square test of 1900 gives unacceptable error probabilities. The classic 1941 Gini test gives error probabilities 29% higher than the chi-square test. *Results and conclusions.* The use of the proposed new differential version of the Gini test allows to obtain results in samples of 16 ex-

periments approximately 9 times better than those of the chi-square test in samples of the same size. A variant of neural network use of classical and newly created criteria on their basis is considered. At the same time, the neural network use of 9 new statistical criteria makes it possible to reduce the probability of errors of the first and second types to a value of 0.031. If the worst-case test is replaced with the Gini differential test in this group, the probability of errors can be reduced to a value of 0.027.

Keywords: fast neural network training algorithms, small samples, statistical criteria for testing the normality and uniformity hypothesis, binary neurocriteria, artificial neurons, improving the accuracy of estimates by eliminating code redundancy

For citation: Guzhova S.A. The Gini differential test for neural network analysis of small biometric data samples. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki = University proceedings. Volga region. Engineering sciences.* 2024;(3):47–54. (In Russ.). doi: 10.21685/2072-3059-2024-3-5

Постановка задачи

К сожалению, большинство статистических вычислений построено на использовании выборок реальных данных большого объема. Например, для достоверной оценки данных по хи-квадрат критерию [1, 2] желательно иметь выборку в 64 и более опытов. Если выборка мала, например имеет только 16 опытов [3], то доверительная вероятность для хи-квадрат критерия оказывается недопустимо низкой. На рис. 1 представлены результаты имитационного моделирования хи-квадрат критерия для выборки в 16 опытов.

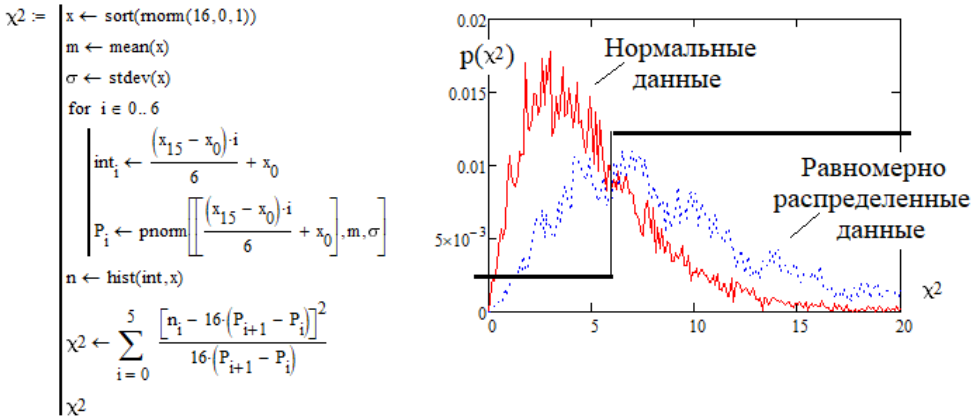

Рис. 1. Распределения плотности вероятности значений хи-квадрат критерия на малой выборке в 16 опытов

Из рис. 1 видно, что при пороге принятия решения $k = 5{,}3$ вероятности ошибок первого и второго рода близки и составляют величину $P_1 \approx P_2 \approx \approx P_{EE} \approx 0{,}332$. То есть использование одного классического хи-квадрат критерия на малой выборке в 16 опытов обеспечивает недопустимо низкую доверительную вероятность 0,668.

Программная реализация в левой части рис. 1 ориентирована на моделирование нормально распределенных данных. Для того чтобы эту программу переделать под равномерные данные, необходимо заменить в ней первую строку на обращение к другой функции: x←sort(runif(16,0,16)).

Положение кардинально меняется, если хи-квадрат критерий Пирсона использовать для больших выборок. На рис. 2 представлены отклики хи-квадрат критерия для большой выборки в 64 опыта.

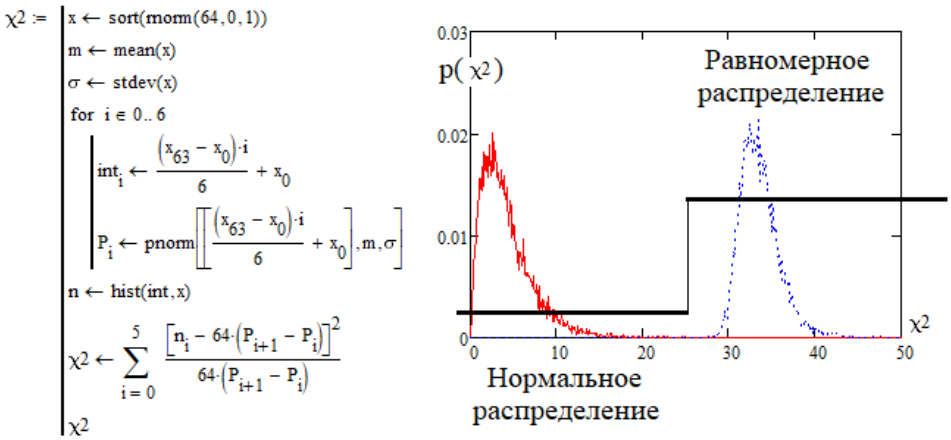

Рис. 2. Хорошее разделение данных с нормальным и равномерным распределением при выборке в 64 опыта и пороге сравнения $k \approx 27$

Из рис. 2 видно, что на большой выборке в 64 опыта нормальное распределение данных и равномерное распределение данных хорошо различимы.

Вероятности появления ошибок первого и второго рода малы и составляют менее 0,0001. При запуске программы моделирования 10 000 раз (левая часть рис. 2) ошибок не обнаружено.

К сожалению, пользователи негативно относятся к необходимости 64 раза писать рукописный пароль или произносить свой голосовой пароль.

Классический критерий Джини

В прошлом веке усилиями исследователей математической статистики было создано более 20 статистических критериев для проверки гипотезы нормального и равномерного распределения данных. В том числе в 1941 г. был создан критерий Джини [2].

Аналитическая запись этого критерия сводится к вычислению интеграла модуля разности наблюдаемой функции вероятности и ожидаемой функции вероятности:

$$D = \int_{-\infty}^{\infty} \left| P(x) - P^*(x) \right| dx, \qquad (1)$$

где $P(.)$ – ожидаемая функция вероятности; $P^*(.)$ – наблюдаемая функция вероятности, анализируемой выборки.

К сожалению, на малых выборках критерий Джини работает хуже хи-квадрат критерия. На рис. 3 представлены результаты моделирования этого критерия на малых выборках в 16 опытов.

Из данных рис. 3 следует рост вероятности ошибок первого рода критерия Джини до величины $P_1 \approx P_2 \approx P_{EE} \approx 0{,}428$. В сравнении с хи-квадрат

критерием рост вероятности ошибок составляет 29 %, что ограничивает применение классического критерия Джини при анализе малых выборок в группе с другими классическими статистическими критериями [4, 5].

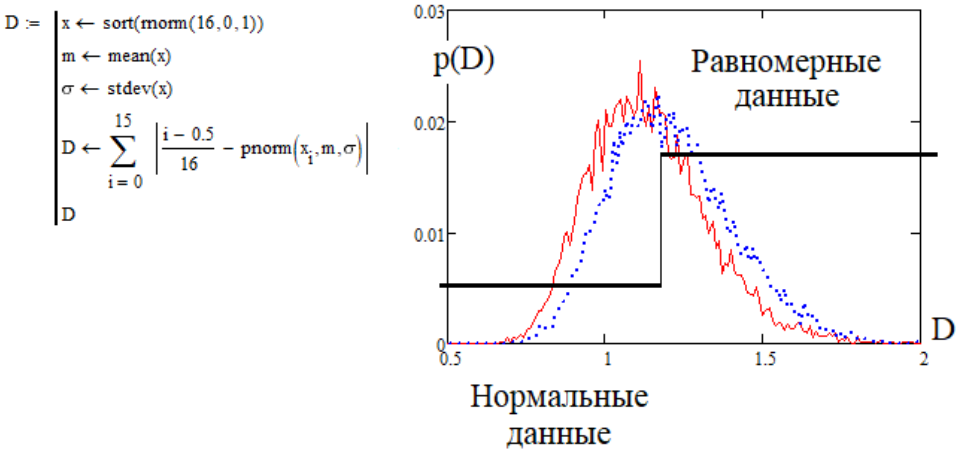

Рис. 3. Плохая разделимость малых выборок
классическим критерием Джини при значении порога $k \approx 1{,}16$

Дифференциальный вариант критерия Джини

Наряду с классической формой критерия Джини (1) может быть использован ее дифференциальный вариант. Этот вариант получается заменой в выражении (1) функций вероятности на их производные (на функции плотности вероятности):

$$dD = \int_{-\infty}^{\infty} \left| p(x) - p^{*}(x) \right| dx, \qquad (2)$$

где $p(.)$ – ожидаемая функция плотности вероятности; $p^{*}(.)$ – наблюдаемая функция плотности вероятности анализируемой выборки.

Проведенные исследования показали, что переход к статистическому анализу плотностей вероятности малых выборок значительно повышает доверительную вероятность принимаемых критерием статистических решений. На рис. 4 приведены результаты моделирования дифференциального варианта критерия Джини.

Из рис. 4 следует, что вероятности ошибок первого и второго рода для нового критерия снижаются до величины $P_1 \approx P_2 \approx P_{EE} \approx 0{,}037$ при пороге $k \approx 3{,}28$. Этот показатель примерно в 9 раз лучше, чем у хи-квадрат критерия на выборках того же объема. Показатель снижения вероятностей ошибок для дифференциального критерия Джини является одним из самых высоких среди группы новых статистических критериев, созданных в этом веке [5].

Нейросетевое объединение классических и новых статистических критериев

Следует обратить особое внимание на то, что каждый из рассмотренных выше статистических критериев имеет собственную шкалу. Формально

можно попытаться решить задачу приведения шкал разных критериев к одной шкале (например, к шкале наиболее часто применяемого хи-квадрат критерия Пирсона).

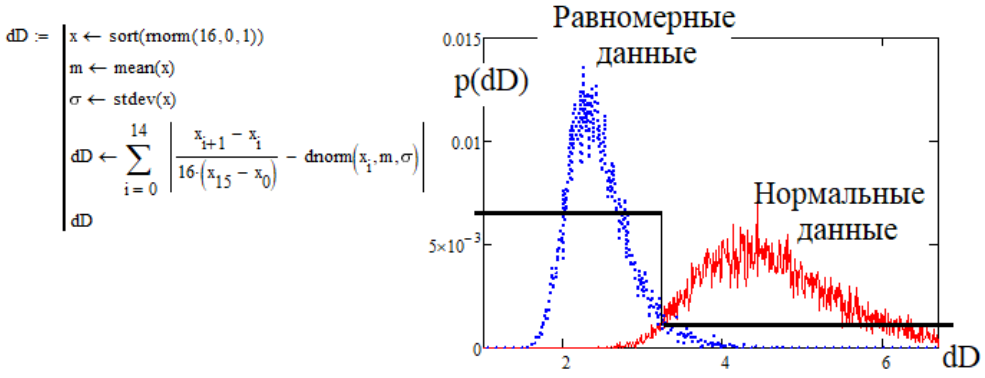

Рис. 4. Хорошая разделимость равномерно и нормально распределенных данных малой выборки дифференциальным вариантом критерия Джини

Перевод в единую обобщенную шкалу может быть выполнен таблично либо через использование полиномов. В частности, такая практика широко используется при измерении температуры термопарами.

Только в единственной обобщенной шкале возможно повысить точность оценок через усреднение откликов разных критериев.

Возможен иной, менее трудоемкий подход, построенный на нейросетевом объединении множества разных статистических критериев. В простейшем случае достаточно использования бинарных искусственных нейронов [6–8], каждый из которых построен на квантовании данных своего статистического критерия.

На рис. 5 представлена нейронная сеть, обобщающая отклики нескольких искусственных нейронов.

Заключение

Проведенные исследования показали, что переход к статистическому анализу плотностей вероятности малых выборок значительно повышает доверительную вероятность принимаемых критерием статистических решений, что подтвердилось результатами моделирования дифференциального варианта критерия Джини.

Использование группы статистических критериев через их нейросетевой перевод в единую обобщенную шкалу, когда каждый бинарный нейрон построен на квантовании данных своего статистического критерия, позволил значительно повысить снижения вероятности ошибок первого и второго рода.

При этом в связи с ростом кодовой избыточности откликов нейронной сети появляется возможность свернуть данные путем обнаружения ошибок и их исправления.

Самым простым алгоритмом в данном случае является «голосование» разрядов. По этому алгоритму подсчитывается число состояний «0» и число состояний «1» избыточного кода.

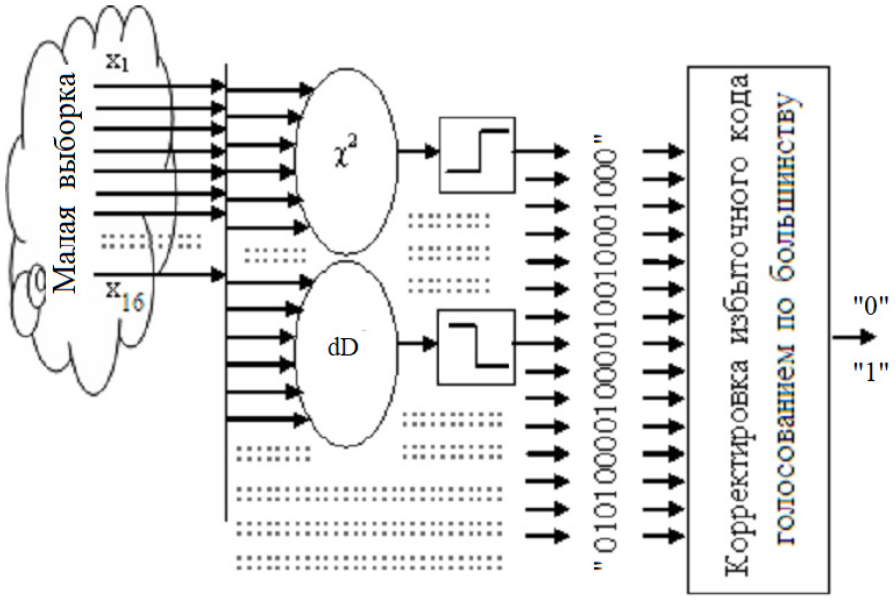

Рис. 5. Нейросетевое обобщение нескольких статистических критериев, позволяющее получать выходной бинарный код с высокой избыточностью

Если число состояний «0» превышает число разрядов с состоянием «1», то выходной блок свертывания кодовой избыточности откликается состоянием «0».

Обычно при организации нейросетевого обобщения с избыточным выходным кодом используют нечетное число искусственных нейронов. В этом случае не возникает неопределенности, когда число состояний «0» и число состояний «1» совпадает.

В случае использования 9 новых статистических критериев, созданных в этом веке, удается добиться снижения вероятности ошибок первого и второго рода до значения 0,031 [4, 9, 10]. Если в этой группе заменить наихудший критерий на рассмотренный в данной статье критерий dD, то вероятность появления ошибок удается снизить до значения 0,027.

Список литературы

1. Р 50.1.037-2002. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа χ2. Госстандарт России. М., 2001. 140 с.
2. Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников. М. : Физматлит, 2006. 816 с.
3. Иванов А. И. Бионика: обучение «на лету» с использованием генетически поразному предобученных искусственных нейронов // Системы безопасности. 2023. № 4. С. 122–125.
4. Иванов А. П., Иванов А. И., Малыгин А. Ю., Безяев А. В. [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
5. Иванов А. П., Иванов А. И., Безяев А. В., Куприянов Е. Н. [и др.]. Обзор новых статистических критериев проверки гипотезы нормальности и равномерности

распределения данных малых выборок // Надежность и качество сложных систем. 2022. № 2. С. 33–44.
6. Волчихин В. И., Иванов А. И., Серикова Ю. И. Снижение требований к объему выборки при нейросетевом объединении классического критерия Эджуорта – Эдлтона – Пирсона и двух его фрактальных аналогов при проверке гипотезы независимости данных // Известия высших учебных заведений. Поволжский регион. Технические науки. 2023. № 1. С. 5–13. doi: 10.21685/2072-3059-2023-1-1
7. Волчихин В. И., Иванов А. И., Иванов А. П., Еременко Р. В., Савинов К. Н. Номограммы для сравнения корректирующих способностей бинарных и троичных нейронов, используемых при многокритериальной проверке гипотезы независимости данных малых выборок // Известия высших учебных заведений. Поволжский регион. Технические науки. 2022. № 4. С. 5–16. doi: 10.21685/2072-3059-2022-4-1
8. 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.
9. Иванов А. И. Нейросетевой многокритериальный статистический анализ малых выборок : справочник. Пенза : Изд-во ПГУ, 2022. 160 с.
10. 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

References

1. *R 50.1.037-2002. Recommendations for standardization. Applied statistics. Rules for proving agreement between experimental and theoretical distribution. Part 1. Criteria of type χ2. State Standard of Russia*. Moscow, 2001:140. (In Russ.)
2. Kobzar' A.I. *Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnykh rabotnikov = Applied mathematical statistics. For engineers and scientists*. Moscow: Fizmatlit, 2006:816. (In Russ.)
3. Ivanov A.I. Bionics: Learning on the fly using genetically differently engineered artificial neurons. *Sistemy bezopasnosti = Security systems*. 2023;(4):122–125. (In Russ.)
4. Ivanov A.P., Ivanov A.I., Malygin A.Yu., Bezyaev A.V. et al. An album of nine classical statistical tests for testing the hypothesis of normal or uniform distribution of small sample data. *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
5. Ivanov A.P., Ivanov A.I., Bezyaev A.V., Kupriyanov E.N. et al. Review of new statistical criteria for testing the hypothesis of normality and uniformity of distribution of small sample data. *Nadezhnost' i kachestvo slozhnykh system = Reliability and quality of complex systems*. 2022;(2):33–44. (In Russ.)
6. Volchikhin V.I., Ivanov A.I., Serikova Yu.I. Reducing sample size requirements for neural network combining the classical Edgerworth – edleton – Pearson test and its two-fractal counterparts when testing the data indenpendance hypothesis. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki = University proceedings. Volga region. Engineering sciences*. 2023;(1):5–13. (In Russ.). doi: 10.21685/2072-3059-2023-1-1
7. Volchikhin V.I., Ivanov A.I., Ivanov A.P., Eremenko R.V., Savinov K.N. Nomograms for comparing the corrective abilities of binary and ternary neurons used in multicriteria testing of the hypothesis of small sample data independence. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki = University proceedings. Volga region. Engineering sciences*. 2022;(4):5–16. (In Russ.). doi: 10.21685/2072-3059-2022-4-1

8. 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 (programs in the MathCAD language): Preprint*. Penza: Izd-vo PGU, 2020:36. (In Russ.)
9. Ivanov A.I. *Neyrosetevoy mnogokriterial'nyy statisticheskiy analiz malykh vyborok: spravochnik = Neural network multicriteria statistical analysis of small samples: handbook*. Penza: Izd-vo PGU, 2022:160. (In Russ.)
10. Volchikhin V.I., Ivanov A.I., Bezyaev A.V., Filipov I.A. Recognition of small choices with a given distribution of data using sophisticated neurons that predict the reliability of 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

Информация об авторах / Information about the authors

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

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

E-mail: 1996svetlanaserikova@gmail.com

Автор заявляет об отсутствии конфликта интересов / The author declares no conflicts of interests.

Поступила в редакцию / Received 15.04.2024

Поступила после рецензирования и доработки / Revised 27.05.2024

Принята к публикации / Accepted 30.07.2024