УДК 519.7; 519.66; 57.087.1, 612.087.1

Вестник СибГАУ
2014. № 3(55). С. 146–150

БИОМЕТРИЧЕСКАЯ СТАТИСТИКА: СГЛАЖИВАНИЕ ГИСТОГРАММ, ПОСТРОЕННЫХ НА МАЛОЙ ОБУЧАЮЩЕЙ ВЫБОРКЕ

Н. И. Серикова[1], А. И. Иванов[2], С. В. Качалин[1]

[1]Пензенский государственный университет
Российская Федерация, 440026, г. Пенза, ул. Красная, 40
E-mail: cnit@pnzgu.ru
[2]Пензенский научно-исследовательский электротехнический институт
Российская Федерация, 440000, г. Пенза, ул. Советская, 9
E-mail: ivan@pniei.penza.ru

Рассматривается вопрос улучшения устойчивости статистических вычислений при малой обучающей выборке за счет усложнения статистической обработки исходных данных. Показано, что достоверность оценивания закона распределения малых выборок биометрических данных может быть увеличена за счет сглаживания гистограмм. Предлагается использовать некоторый цифровой фильтр, который будет осуществлять сглаживание традиционной гистограммы, и за счет этой дополнительной обработки улучшится устойчивость статистических вычислений. Корректный выбор окна цифрового фильтра сглаживания и многократное искусственное увеличение числа дискрет, используемых при цифровой фильтрации, дает возможность существенно увеличить мощность критерия согласия Джини и хи-квадрат критерия. Сглаженный критерий Джини менее чувствителен к числу примеров в тестовой выборке. Таким образом, его применение приводит к уменьшению вероятности ошибок второго рода, обусловленных ограниченным числом опытов. Представлена сравнительная таблица вероятности ошибок принятия решений по хи-квадрат критерию и критерию Джини для сглаженных данных. В отличие от критерия хи-квадрат, критерий Джини оказывается работоспособен даже на выборке из малого количества опытов. Таким образом, для задач биометрии мы наблюдаем очевидный выигрыш от применения критерия Джини.

Ключевые слова: биометрия, обучение, закон распределения, статистика, вероятность, гистограмма, критерий согласия.

Vestnik SibGAU
2014, No. 3(55), P. 146–150

BIOMETRIC STATS: SMOOTHING HISTOGRAMS BASED ON SMALL TRAINING SAMPLE

N. I. Serikova[1], A. I. Ivanov[2], S. V. Kachalin[1]

[1]Penza State University
40, Krasnaya St., Penza, 440026, Russian Federation
E-mail: cnit@pnzgu.ru
[2]Penza research Electrotechnical Institute
9, Sovetskaya St., Penza, 440000, Russian Federation
E-mail: ivan@pniei.penza.ru

The question of improving the stability of statistical calculations with small training set by complications statistical processing raw data. It is shown that the accuracy of the estimation of the distribution law of small samples of biometric data can be increased by smoothing the histogram. It is proposed to use some digital filter, which will be smoothed traditional histograms, and due to this additional processing will improve the stability of the statistical calculations. The correct choice of the window digital anti-aliasing filter and multiple artificial increase in the number of discrete, used in digital filtering allows to significantly increase the power of goodness of fit test Gini and chi-square test. Graduated Gini goodness of fit test Gini is less sensitive to the number of examples in the test sample. Thus, its use leads to a decrease in error probability of the second kind, due to the limited number of experiments. The article presents a comparative table of the error probability of decision-making by the chi-square test and the Gini for the smoothed data. In contrast to the chi-square test, the Gini criterion is functional even on a sample of small kolichistvo experiments. Thus, for the problems we are seeing biometrics obvious benefit from the application of the criterion of Gini.

Keywords: biometrics, training, distribution law, statistics, probability, histogram, goodness of fit test.

Нейросетевые преобразователи биометрия–код приходится обучать на выборке из 20–30 примеров образа «Свой», если обучение ведется по стандартному алгоритму [1–3]. Как правило, потребитель неохотно воспроизводит примеры своего биометрического образа. Люди способны обучаться распознаванию образов (и в том числе биометрических образов) на трех-пяти примерах. По этой причине пользователи ожидают от биометрического искусственного интеллекта способности обучаться на меньшем числе примеров [4–6].

В связи с этим возникает задача сократить число примеров биометрического образа «Свой» в обучающей выборке за счет усложнения статистической обработки исходных данных. К сожалению, специальным математическим приемам, позволяющим повысить устойчивость статистических расчетов, уделялось недостаточно внимания. Так, при оценке гипотезы по критерию согласия Джини или критерию согласия хи-квадрат [7–9] желательно выбирать как можно более высокий показатель числа степеней свободы (желательно увеличивать число столбиков гистограммы). Однако при этом быстро растет размер выборки, на которой строятся критерии. Возникает противоречие между желанием выбирать как можно большее число степеней свободы статистических критериев согласия и ограниченным размером исходных биометрических данных.

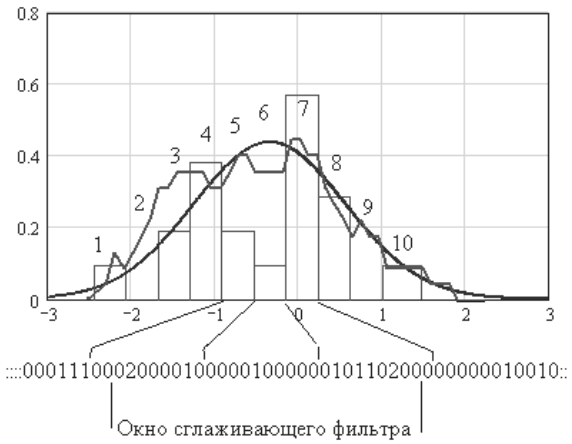
В связи с этим противоречием возникает задача расчета и использования некоторого цифрового фильтра, который будет осуществлять сглаживание традиционной гистограммы, и за счет этой дополнительной обработки улучшится устойчивость статистических вычислений.

Классические гистограммы обычно строятся исходя из того, что в каждом столбике находится от 2 до 5 зафиксированных в опыте результатов. Так, если мы имеем выборку $x_1 \le x_2 \le x_3 \le , \ldots, \le x_{21}$ из 21 примера упорядоченных по возрастанию данных, то, рассчитывая на среднее число в 2 попадания в каждый столбик классической гистограммы, мы должны выбрать ширину столбиков по формуле

$$\Delta x = \frac{x_{21} - x_1}{10} = \frac{\max(x) - \min(x)}{10}. \qquad (1)$$

Пример реализации такой гистограммы, состоящей из 10 столбиков, приведен на рисунке (данные получены от генератора нормального шума с единичной дисперсией и нулевым математическим ожиданием). Из рисунка видно, что 2-й и 9-й столбики классической гистограммы оказались пустыми (данные в них вообще не попали). Максимальный скачок наблюдается между 6-м и 7-м столбиками классической гистограммы, он составляет 0,51.

Фактически столь значительные колебания данных гистограммы обусловлены тем, что использовалась достаточно малая выборка исходных данных и применена очень грубая их обработка путем простого подсчета числа попаданий в заданные интервалы. Усложним предварительную обработку данных, разбив каждый из интервалов Δx на 10 меньших интервалов $0{,}1 \cdot \Delta x$. Всего получится 100 микроинтервалов внутри динамического диапазона входных данных. Кроме того, мы можем добавить справа и слева от обнаруженного динамического диапазона изменения данных по 100 интервалов.



Классическая гистограмма и сглаженная гистограмма, полученная цифровым фильтром с нулевыми фазовыми искажениями

В итоге получается 300 микроинтервалов, на которых будет работать цифровой фильтр, сглаживающий исходные данные. Далее будем заполнять исходные данные по тому же правилу, по которому строятся классические гистограммы. Если в микроинтервал отсчетов не попало, то вводится состояние «0». Если в тот или иной микроинтервал попадают реальные данные 1, 2, 3... раза, то в этих интервалах должны появиться состояния «1», «2», «3» и т. д. В итоге появляется цифровая последовательность, пример которой приведен в нижней части рисунка.

Для осуществления цифрового сглаживания данных необходимо выбрать некоторое нечетное окно цифрового сглаживающего фильтра (на рисунке выбранная ширина окна составляет 27 микродискрет). На выбранной ширине скользящего окна может быть построен любой сглаживающий колебания низкочастотный цифровой фильтр. В частности, может быть использован фильтр простого усреднения (прямоугольное окно) с подстановкой результата в центр окна. В нашем случае для ширины окна 27 отсчетов, длины кодовой последовательности в 300 микродискрет кодовая последовательность z_1, z_2, ..., z_i, ..., z_{200} сглаживается простым усреднением:

$$y_{i+14} = \frac{1}{27}\sum_{j=1}^{27} z_{i+j}. \qquad (2)$$

Пример сглаженной гистограммы представлен на рисунке. Колебания данных сглаженной гистограммы намного меньше, чем скачки у классической гистограммы. Именно этот эффект и должен приводить к росту устойчивости статистических оценок по критерию

подобия Джини и по критерию правдоподобия хи-квадрат. Наиболее просто оценивается выигрыш в стабильности статистических вычислений по критерию Джини. Этот критерий удобен тем, что вероятность подобия экспериментального распределения данных и гипотетического близка к половине показателя Джини:

$$P_0 \approx 1 - \frac{1}{2}\int_{-\infty}^{+\infty} |p(x) - \tilde{p}(x)|dx, \qquad (3)$$

где P_0 – вероятность того, что теоретическое распределение $p(x)$ и экспериментальное распределение $\tilde{p}(x)$ совпадают.

Интеграл в выражении (3) является критерием Джини, он очень чувствителен к числу столбцов гистограммы и объему выборки исходных данных. Ошибка дискретизации данных при вычислении критерия Джини может быть оценена по максимальному отклонению шума дискретизации от ожидаемой кривой распределения значений $p(x)$. На рисунке ожидаемой теоретической плотностью является нормальный закон. Классическая гистограмма дает максимальное значение ошибки отклонения 0,25, сглаживание данных гистограммы снижает максимальную ошибку до 0,15. Мы получаем примерно 40%-ное снижение ошибок квантования за счет увеличения в 10 раз числа квантовых интервалов и сглаживания данных цифровым фильтром. Если говорить в терминах вероятности проверки статистических гипотез, то примерно на 40 % возросла мощность критерия Джини и уровень доверия в оценке вероятности (3).

Очевидно, что после сглаживания гистограммы исходных данных должны возрастать и мощности других широко используемых на практике статистических критериев проверки гипотез. Однако для оценки достижимых улучшений необходимо проводить специальные численные эксперименты. Предположительно в ближайшее время будет проведен численный эксперимент по влиянию цифрового сглаживания гистограмм на широко используемый в биометрии критерий хи-квадрат проверки статистических гипотез.

Описанная выше процедура сглаживания исходных данных цифровым усредняющим фильтром является незначительным программным усложнением обработки данных. Очевидно, что выполнять подобное сглаживание исходных данных в конце XIX в. и в начале XX в. технологически было сложно из-за отсутствия программируемой вычислительной техники. В конце XX века описанные выше процедуры уже могли быть реализованы при программной обработке экспериментальных данных.

При реализации программной обработки биометрических данных применение простейшего сглаживающего фильтра (2) приводит к незначительному усложнению программ (добавляется несколько строк кода), дающему ощутимый рост доверия к статистическим оценкам. Для того чтобы реализовывать программную фильтрацию, необходимо задать число столбцов классической гистограммы (ширину столбика Δx), коэффициент увеличения числа дискрет и оптимальную ширину окна усредняющего фильтра. Фактически необходимо иметь решение оптимизационной задачи по выбору ширины окна сглаживающего фильтра при разном числе столбцов классической гистограммы и разном числе внутренних дискрет каждого столбца.

Предварительные расчеты показали, что ширина окна сглаживающего цифрового фильтра сильно зависит от числа биометрических данных в исследуемой выборке. Результаты оптимизационного моделирования сведены в табл. 1, построенную для 10 столбцов классической гистограммы и 10-кратном увеличении числа квантов при цифровом сглаживании.

Очевидно, что масштабное увеличение числа дополнительных дискрет будет приводить к пропорциональному росту длины окна сглаживающего цифрового фильтра. Изменение числа столбиков классической гистограммы также связано линейно с оптимальной длиной окна сглаживающего фильтра.

Таблица 1

Связь длины окна сглаживающего цифрового фильтра с размерами обучающей выборки

Число примеров в выборке	12	16	20	24	28	32	36	40
Ширина окна усреднения	83	71	61	53	49	45	41	39
Число примеров в выборке	44	48	52	56	60	64	68	72
Ширина окна усреднения	37	35	33	31	29	27	27	25

Таким образом, табл. 1 может рассматриваться как универсальная. Ее данные легко пересчитываются на любое число столбцов классической гистограммы и любое число дискрет дополнительного разбиения данных перед их сглаживанием. Вместо равномерного взвешивания может быть использована более сложная взвешивающая функция цифровой свертки. При этом оптимальная ширина окна будет описываться другой таблицей. При необходимости для каждой взвешивающей функции цифрового сглаживания может быть построена соответствующая таблица оптимальных значений скользящего окна.

Следует подчеркнуть, что сегодня на практике наиболее часто используют хи-квадрат критерий проверки статистических гипотез. При использовании этого критерия число столбцов гистограммы выбирают так, чтобы в среднем на каждый столбец гистограммы приходилось по 5 опытов [10]. Если придерживаться этой рекомендации, то применение хи-квадрат критерия на малых тестовых выборках приводит к появлению ошибок принятия решений с вероятностью второго рода (табл. 2).

Таблица 2

Вероятности ошибок принятия решений по хи-квадрат критерию и критерию Джини для сглаженных данных

Число опытов	8	12	16	20	24	32	48	64
Квантиль доверия P_1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1
$P_2(\chi^2)$	–	–	–	0,46	0,32	0,20	0,17	0,13
$P_2(Dg)$	0,41	0,35	0,25	0,19	0,17	0,11	0,06	0,04

Сглаженный критерий Джини менее чувствителен к числу примеров в тестовой выборке, по этой причине его применение приводит к меньшим значениям вероятности ошибок второго рода, обусловленных ограниченным числом опытов.

Табл. 2 строилась исходя из доверительной вероятности $P_1 = 0,1$ к принимаемому решению, далее использовался критерий хи-квадрат и критерий Джини с 10-кратным увеличением числа столбцов гистограммы и окном сглаживания в 11 отсчетов. Проверка велась на равномерном законе распределения, поверялась гипотеза нормальности закона распределения.

Из табл. 2 видно, что классический хи-квадрат критерий не работает, если выполнить условие попадания в среднем 5 опытов в каждый столбец гистограммы. Происходит это из-за необходимости выбора числа степеней свободы хи-квадрат критерия на 3 единицы меньше, чем число столбцов гистограммы. Фактически хи-квадрат критерий начинает работать при 20 опытах с вероятностью ошибки второго рода 0,46. С такой вероятностью равномерный закон распределения ошибочно принимается за нормальный закон.

Иная ситуация наблюдается для критерия Джини. Этот критерий оказывается работоспособен даже на выборке из 8 опытов. Вероятность ошибки второго рода для критерия Джини оказывается ниже, чем для классического хи-квадрат критерия при любом размере выборки реальных данных. Для задач биометрии мы наблюдаем очевидный выигрыш от применения критерия Джини. Табл. 2 демонстрирует эффективность применения одномерного критерия Джини $Dg(x)$.

Биометрии приходится иметь дело с множеством сильно зависимых биометрических параметров. Так, среда моделирования «БиоНейроАвтограф» [11] учитывает 416 биометрических параметров. Это означает, что от одномерного варианта критерия Джини необходимо перейти к его многомерному варианту $Dg(x_1, x_2, x_3, \ldots, x_{416})$. Построить одну 416-мерную функцию обобщенного критерия Джини не удается. В этом плане следует использовать опыт создания сетей максимального правдоподобия Байеса [12]. Обычно сети Байеса (вероятностные рассуждения) легко реализуемы для 3–9 учитываемых параметров. В частности, Хайкиным [13] было показано, что решения Байеса легко формализуются квадратичными формами, содержащими обратную ковариационную матрицу. Именно проблемы обращения ковариационных матриц при малом числе исходных данных являются сдерживающим фактором.

Предположительно, что синтез обобщенного критерия Джини столкнется с теми же проблемами, что и синтез высокоразмерных сетей наибольшего правдоподобия Байеса [12; 13]. Предполагается решить проблему объединением частных критериев Джини, имеющих размерность от 3 до 9, обобщающим правилом Хэмминга. Возникающие при этом сверхмалые вероятности появления ошибок второго рода предполагается учитывать по методике ГОСТ Р 52633.3–2011 [14] с учетом влияния коэффициента равной коррелированности выходных состояний частных решений [15].

Библиографические ссылки

1. ГОСТ Р 52633.0–2006. Защита информации. Техника защиты информации. Требования к средствам высоконадежной биометрической аутентификации. М. : Стандартинформ. 2007. 24 с.

2. ГОСТ Р 52633.5–2011. Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия – код доступа. М. : Стандартинформ, 2012. 20 с.

3. Нейросетевая защита персональных биометрических данных / В. И. Волчихин [и др.]. М. : Радиотехника, 2012. 160 с.

4. Волчихин В. И., Иванов А. И., Фунтиков В. А. Быстрые алгоритмы обучения нейросетевых механизмов биометрико-криптографической защиты информации : монография. Пенза : Изд-во Пензенского гос. ун-та, 2005. 273 с.

5. Технология использования больших нейронных сетей для преобразования нечетких биометрических данных в код ключа доступа : монография / Б. С. Ахметов [и др.]. Алматы : ТОО «Издательство LEM», 2014. 144 с.

6. Biometric Technology in Securing the Internet Using Large Neural Network Technology / B. Akhmetov, [et al.] // World Academy of Science, Engineering and Technology. 2013, Iss. 79. P. 129–138, pISSN 2010-376X, eISSN 2010-3778. URL: www.waset.org.

7. Кобзарь А. И. Прикладная математическая статистика для инженеров и научных работников. М. : Физматлит, 2006. 816 с.

8. Крамер Г. Математические методы статистики. М. : Мир, 1975. 516 с.

9. Уилкс С. С. Математическая статистика. М. : Наука, 1967. 632 с.

10. Денисов В. И., Лемешко Б. Ю., Постовалов С. Н. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим : метод.

рекомендации. Ч. I. Критерии типа χ2. Новосибирск : Изд-во НГТУ, 1998. 126 с.

11. Среда моделирования «БиоНейроАвтограф» : архив с программным обеспечением размещен на сайте ОАО «Пензенский научно-исследовательский электротехнический институт» [Электронный ресурс]. URL: http://пниэи.рф/activity/science/noc.htm.

12. Стюард Рассел, Питер Норвиг. Искусственный интеллект. Современный подход. М.; СПб.; Киев : Вильямс, 2006. 1407 с.

13. Саймон Хайкин. Нейронные сети. Полный курс. М.; СПб.; Киев : Вильямс, 2006. 1103 с.

14. ГОСТ Р 52633.3–2011. Защита информации. Техника защиты информации. Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора. М. : Стандартинформ, 2011. 16 с.

15. Оценка рисков высоконадежной биометрии : монография / Б. С. Ахметов [и др.]. Алматы : Из-во КазНТУ им. К. И. Сатпаева, 2014. 108 с.

References

1. *GOST R 52633.0–2006 Zashhita informacii. Tehnika zashhity informacii. Trebovanija k sredstvam vysokonadezhnoy biometricheskoy autentifikacii.* [Information protection. Information protection technology. Requirements for means of highly reliable biometric authentication]. Moscow, Standartinform Publ., 2007, 24 p.

2. *GOST R 52633.5–2011. Zashhita informacii. Tehnika zashhity informacii. Avtomaticheskoe obuchenie nejrosetevyh preobrazovatelej biometrija-kod dostupa*. [Information protection. Information protection technology. The neural net biometry-code convertor automatic training]. Moscow, Standartinform Publ., 2012, 20 p.

3. Volchihin V. I., Ivanov A. I., Funtikov V. A., Nazarov I. G, Yazov Y. K. *Nejrosetevaya zashhita personalnyh biometricheskix dannyh.* [Neural network protection of personal biometric data]. Moscow, Radiotekhnika Publ., 2012, 160 p.

4. Volchihin V. I., Ivanov A. I., Funtikov V. A. *Bystrye algoritmy obucheniya nejrosetevyh mehanizmov biometriko-kriptograficheskoj zashhity informacii.* [Fast learning algorithms of neural network tools biometriko-cryptographic information security]. The Monograph. Penza, Penza State University Publ., 2005, 273 p.

5. Akhmetov B. S., Ivanov A. I., Funtikov V. A., Bezyaev A. V., Malygina E. A. *Tekhnologiya ispol'zovaniya bol'shikh neyronnykh setey dlya preobrazovaniya nechetkikh biometricheskikh dannykh v kod klyucha dostupa.*[The technology of using large neural networks to convert the fuzzy biometric data in the key code access]. Kazakhstan, Almaty, TOO “Izdatel'stvo LEM” Publ., 2014, 144 p.

6. Akhmetov B., Doszhanova A., Ivanov A., Kartbaev T. and Malygin A. Biometric Technology in Securing the Internet Using Large Neural Network Technology. *World Academy of Science, Engineering and Technology*. July, 2013, Issue 79, p. 129–138.

7. Kobzar A. I. *Prikladnaya matematicheskaya statistika dlya inzhenerov i nauchnyh rabotnikov.* [Applied mathematical statistics for engineers and scientists]. Moscow, Fizmatlit Publ., 2006, 816 p.

8. Kramer G. *Matematicheskie metody statistiki.* [Mathematical methods of statistics]. Moscow, Mir Publ, 1975, 516 p.

9. Uilks S. S. *Matematicheskaya statistika*. [Mathematical statistics]. Moscow, Nauka Publ, 1967, 632 p.

10. Denisov V. I., Lemeshko B. Yu., Postovalov S. N. *Prikladnaya statistika. Pravila proverki soglasiya opytnogo raspredeleniya s teoreticheskim. Metodicheskie rekomendatsii. Chast' I. Kriterii tipa χ2.* [Applied statistics. Validation rules experienced distribution agreement with the theoretical. Guidelines. Part I. Criteria type χ2]. Novosibirsk, NGTU Publ, 1998, 126 p.

11. *Sreda modelirovaniya “BioNeyroAvtograf” : arkhiv s programmnym obespecheniyem razmeshchen na sayte OAO “Penzenskiy nauchno-issledovatel'skiy elektrotekhnicheskiy institut”* [The modeling environment “BioNeyroAvtograf”. Developed by “Penza research Electrotechnical Institute”]. (In Russ.) Available at: http://пниэи.рф/activity/science/noc.htm.

12. Styuard Rassel, Piter Norvig *Iskusstvennyy intellekt. Sovremennyy podkhod.* [Artificial Intelligence. The modern approach]. Moscow, St. Petersburg, Kiev, 2006, Vil'yams Publ, 1407p.

13. Saymon Khaykin. *Neyronnye seti. Polnyy kurs.* [Neural networks. Full course]. Moscow, St. Petersburg, Kiev, 2006, Vil'yams Publ, 1103p.

14. *ГОСТ Р 52633.3–2011. Zashchita informatsii. Tekhnika zashchity informatsii. Testirovanie stoykosti sredstv vysokonadezhnoy biometricheskoy zashchity k atakam podbora*. [Information protection. Information protection technology. The high reliability biometric protection means endurance testing from matching attacks]. Moscow, Standartinform Publ., 2011, 16 p.

15. Akhmetov B. S., Nadeev D. N., Funtikov V. A., Ivanov A. I., Malygin A. Yu. *Otsenka riskov vysokonadezhnoy biometrii.* [Risk assessment of highly reliable biometrics]. Kazakhstan, Almaty, KazNTU im. Satpaeva Publ., 2014, 108 p.