Захаров О.С.

Хи-квадрат аппроксимация асимметричных гистограмм распределений значений показателей стабильности биометрических параметров

Показано, что для аппроксимации асимметричных распределений показателей стабильности биометрических параметров рукописного почерка хорошо подходит хи-квадрат распределение. Дана номограмма связи числа степеней свободы хи-квадрат распределения и коррелированности обрабатываемых биометрических данных.

Введение

В настоящее время развиваются две ветви биометрических систем. Идеология создания первой ветви биометрических систем сформировалась в конце прошлого века. Она не предполагает предъявления высоких требований к обеспечению конфиденциальности используемых биометрических данных человека. Кроме того, эта первая ветвь биометрических средств не накладывает жестких требований на уровень сложности (уровень интеллектуальности) используемого алгоритма обработки биометрических данных.

Параллельно с первой ветвью в начале этого века за рубежом и в России начали создаваться высокоинтеллектуальные и высоконадежные средства биометрической защиты информации, которые параллельно с высоким качеством принимаемых решений обеспечивают конфиденциальность используемых биометрических данных. Зарубежные исследователи идут по пути создания нечётких экстракторов биометрических данных [1, 2], в России для этих целей разрабатывается технология использования больших и сверхбольших искусственных нейронных сетей [3].

Увеличение размеров нейронной сети примерно в 100 раз (входной и выходной размерности нейросетевого преобразователя) позволяет снизить вероятность коллизий (ошибочного принятия образа «Чужой» за образ «Свой») примерно в миллиард раз. Очевидно, что столь высокие показатели качества принимаемых большими нейронными сетями решений нуждаются в подтверждении через тестирование.

Для целей тестирования средств высоконадежной, высокоинтеллектуальной биометрии в России введен в действие базовый национальный стандарт [4], положения которого в части формирования тестовых баз биометрических образов уточняются стандартом [5], прошедшим стадию публичного обсуждения.

При формировании представительных баз естественных биометрических образов возникает проблема создания корректной классификации биометрических образов. Используемая зарубежными исследователями классификация доноров биометрии на классы: «Овец», «Коз», «Волков», «Ягнят» и «Хамелеонов» [6] трудно формализуема. В связи с этим, разрабатываемый национальный стандарт [5] предлагает другую классификацию, позволяющую делить все биометрические образы на 7 классов по их стабильности, уникальности, качеству.

Применение распределения хи-квадрат

Одной из нерёшенных проблем корректного описания распределений биометрических образов является то, что распределение стабильности, уникальности и качества любых биометрических параметров описывается асимметричными законами распределения значений. Например, показатель стабильности биометрического параметра вычисляется следующим образом [5]:

$$c(v_i) = \frac{\sigma_{Чужой}(v_i)}{\sigma_{Свой}(v_i)} \quad (1),$$

где $\sigma_{Чужой}(v_i)$ – стандартное отклонение i-го биометрического параметра множества образов «Чужой»;

$\sigma_{Свой}(v_i)$ – стандартное отклонение i-го биометрического параметра множества образов «Свой».

Для показателя стабильности биометрических параметров относительно нестабильного рукописного образа получается распределение значений, приведенное на рисунке 1.

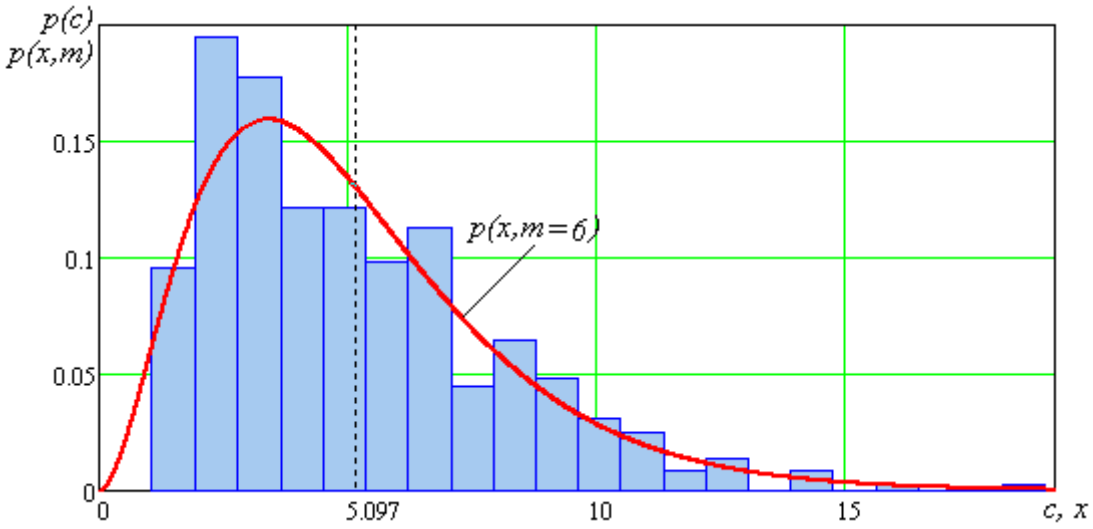

Рис. 1. Гистограмма распределения показателей стабильности 416 параметров биометрического образа с относительно низкой средней стабильностью – 5,097

Пример гистограммы распределения значений показателей стабильности для другого (более стабильного) биометрического образа приведен на рисунке 2.

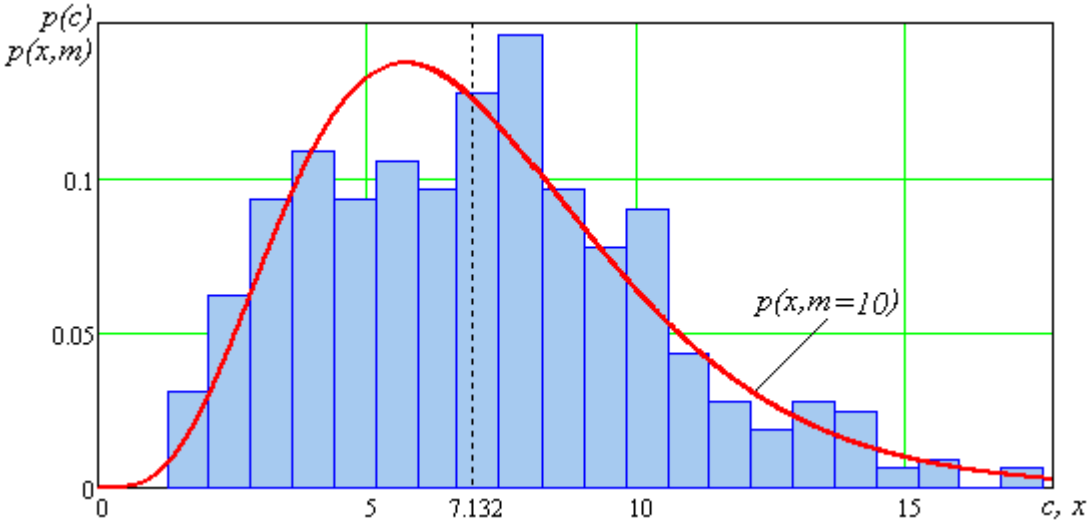

Рис. 2. Гистограмма распределения показателей стабильности 416 параметров биометрического образа с относительно высокой средней стабильностью - 7,132

Из рисунка 1 и рисунка 2 видно, что распределения показателей стабильности биометрических параметров имеют существенную асимметрию. То есть, для их корректного статистического описания использовать аппроксимацию в виде симметричного нормального закона нежелательно. Практика показала, что для описания асимметричных распределений биометрических параметров по показателям их стабильности, уникальности и качества хорошо подходит хи-квадрат распределение [7, 8].

Общая идея применения подобной аппроксимации состоит в использовании классической хи-квадрат плотности распределения значений:

$$p(x,m)=\{2^{\frac{m}{2}}\cdot\Gamma(\frac{m}{2})\}^{-1}\cdot x^{\frac{m}{2}-1}\cdot e^{\frac{-x}{2}} \qquad (2),$$

где m – степень свободы распределения;

x – переменная, находящаяся в пределах от 0 до ∞.

Кривые плотностей распределений хи-квадрат функции (2) приведены на рисунке 3.

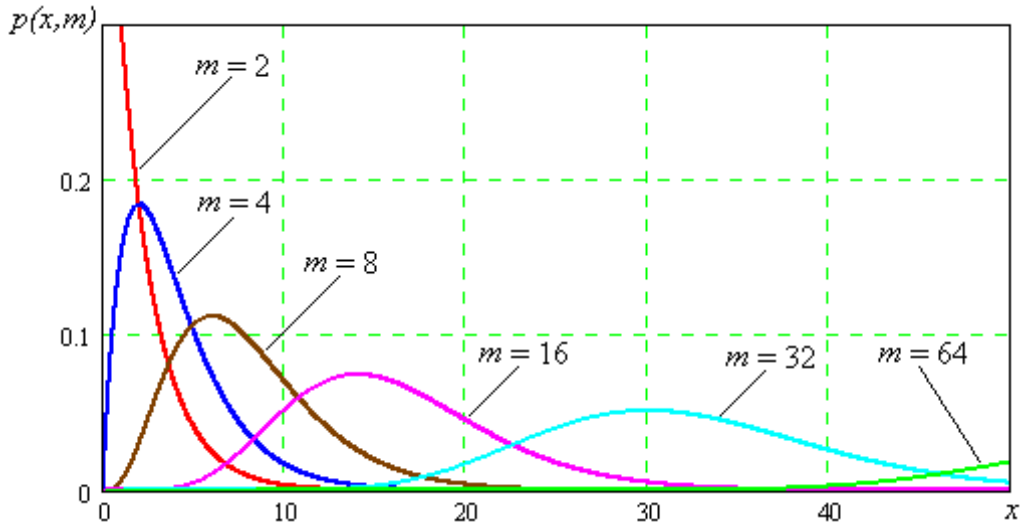

Рис. 3. Плотности классического распределения хи-квадрат для различных степеней свободы

Одной из особенностей распределений рисунка 3 является то, что математическое ожидание каждого распределения точно совпадает с показателем степени свободы – m. При поиске связи показателя степени свободы с коррелированностью биометрических данных [7, 8] оказалось выгодно нормировать переменную (2) по показателю степени свободы. Примеры подобных деформаций хи-квадрат распределений приведены на рисунке 4.

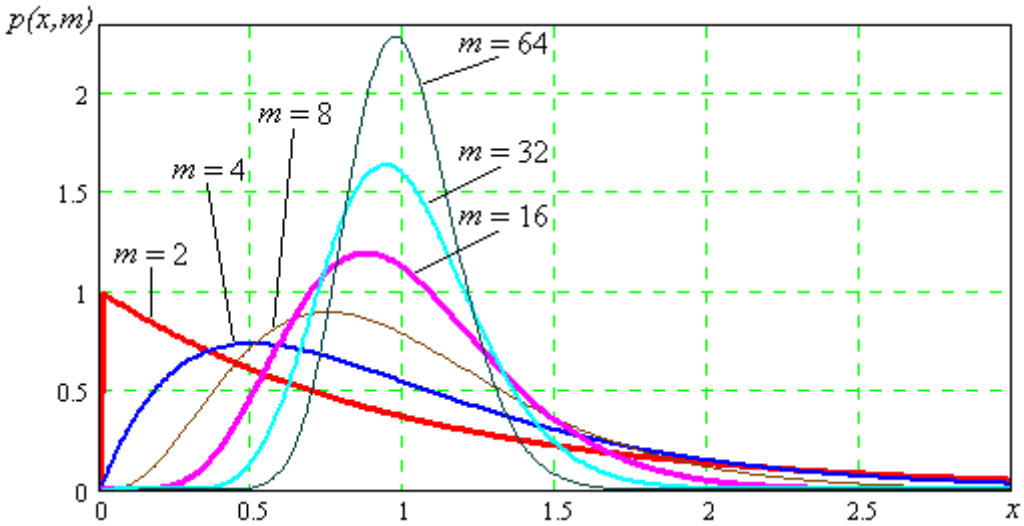

Рис. 4. Эффект от нормирования распределения хи-квадрат по числу степеней свободы

Легко показать, что нормированные по показателю степени свободы распределения, представленные на рисунке 4, будут иметь математические ожидания очень простого вида:

$$E(\frac{x}{m}) = 1 - \frac{1}{m} \quad (2).$$

При росте числа степеней свободы математическое ожидание нормированных распределений стремится к 1, кроме того распределения нормализуются и исчезает их асимметрия.

Для нас принципиально важным является то, что параметр m (число степеней свободы χ^2 распределения) фактически отвечает за форму аппроксимации, а масштабирование можно рассматривать как параметр аппроксимации асимметричного, экспериментально полученного, распределения биометрических параметров.

Как оказалось, математический прием по масштабированию переменной χ^2 распределения и подбору числа степеней свободы является эффективным и им удается хорошо аппроксимировать распределения показателей стабильности, уникальности и качества, введенных стандартом [5].

Следует подчеркнуть, что при подобной аппроксимации биометрических данных рисунка 1 и рисунка 2 мы полностью остаемся в рамках классики, рассматривая только целые показатели числа степеней свободы m = 6 (рисунок 1) и m = 10 (рисунок 2), а все проверяемые биометрические параметры рассматриваются как независимые.

Переход к дробному числу степеней свободы

К сожалению, гипотеза независимости биометрических параметров вообще не работает. Как показала практика, все биометрические параметры оказываются зависимыми. Вторым крайне важным моментом предложенного подхода к аппроксимации является то, что параметр m мы можем рассматривать как (нецелый) дробный параметр. Это позволяет дополнительно снизить погрешность аппроксимации.

Необходимость перехода к дробному показателю размерности иллюстрируется рисунком 5. На этом рисунке отображена гистограмма распределения показателей стабильности рисунка 1 и три ее аппроксимации для 5, 6 и 7 степеней свободы. Каждая из этих аппроксимаций имеет ошибку приближения $\Delta_5 = 0.7$, $\Delta_6 = 0.31$, $\Delta_7 = 0.41$.

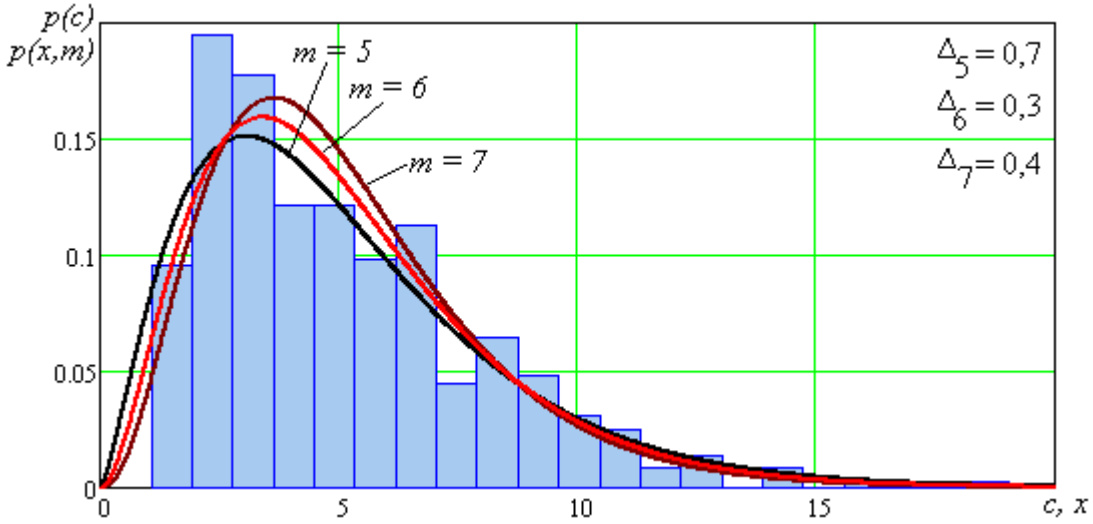

Рис. 5. Три аппроксимации гистограммы распределения значений показателей стабильности биометрических параметров с тремя разными целыми показателями числа степенями свободы

На рисунке 6 дана кривая изменения ошибки аппроксимации хи-квадрат распределением для случая, когда число степеней свободы является непрерывной

величиной (осуществлен переход от дискретных значений m = 5, 6, 7 к непрерывной переменной m).

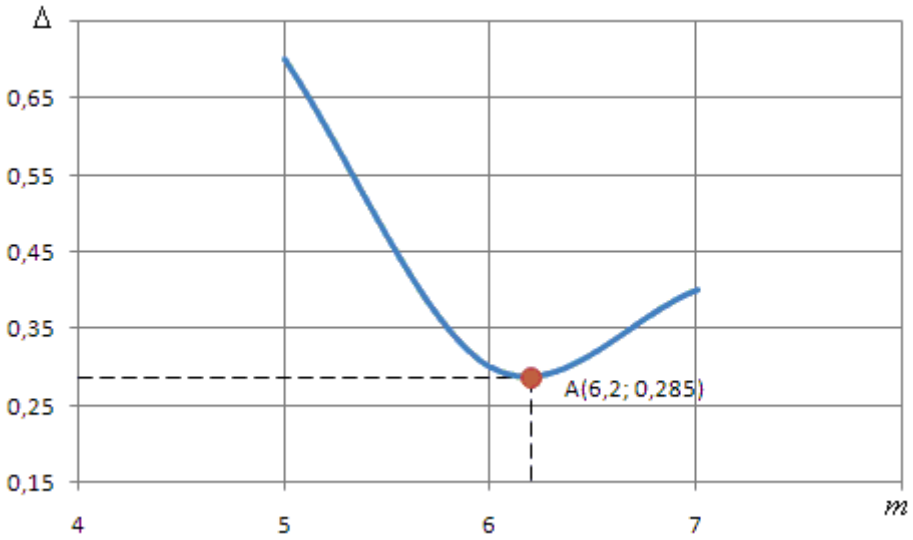

Рис. 6. Минимизация ошибки приближения за счет выбора дробного показателя числа степеней свободы.

Из рисунка 6 видно, что оставаясь в рамках классических представлений и не допуская возможности существования дробных показателей степеней свободы распределения хи-квадрат мы фактически не можем подойти к точке минимума ошибки. Если же мы допустим существование дробных степеней свободы и примем число степеней свободы 6.2, то получим ощутимое снижение погрешности приближения асимметричных распределений значений биометрических данных и их показателей.

Применяя дробные показатели числа степени свободы, мы сталкиваемся с необходимостью доопределения классического распределения хи-квадрат для нецелых значений m в интервале значений больших 1. Это доопределение классического распределения можно осуществлять чисто формально, так как формула аналитического описания (2) содержит функции, определенные как для целых, так и для дробных значений переменных.

Обработка зависимых данных

Еще одним важным моментом является то, что классическое распределение хи-квадрат применительно к обработке биометрических данных нуждается в доопределении на случай зависимых данных. Это может быть сделано, например, через номограмму связи относительной ошибки наблюдения числа степеней свободы $\Delta m/(m\text{-}1)$ со значением коэффициента корреляции между обрабатываемыми биометрическими данными – r (рисунок – 7).

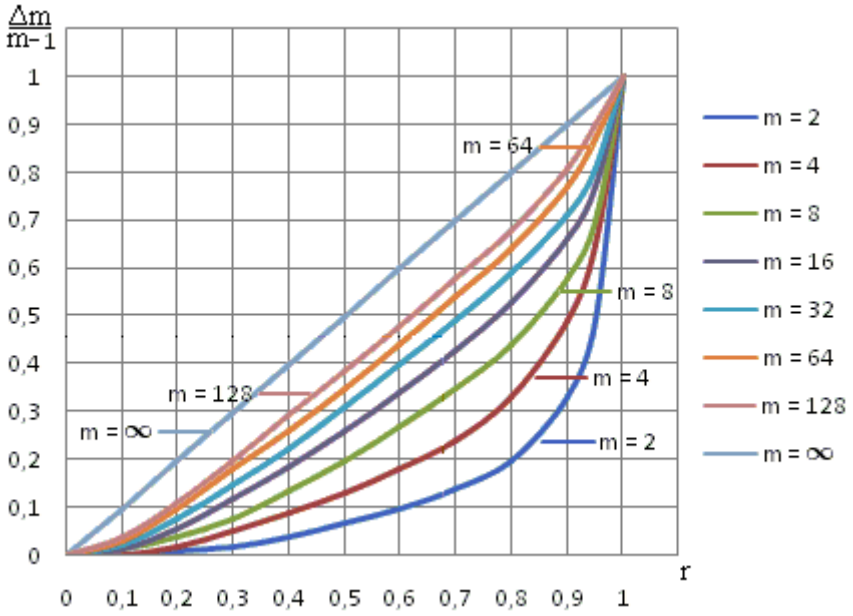

Рис. 7. Номограмма связи относительной ошибки наблюдения числа степеней свободы Δm/(m-1) со значением коэффициента корреляции между обрабатываемыми биометрическими данными – r

Располагая номограммой рисунка 7 мы имеем возможность пересчитать число степеней свободы распределения хи-квадрат для зависимых данных по ошибочно наблюдаемому числу степеней свободы независимых биометрических данных. Например, при аппроксимации распределением хи-квадрат биометрических данных мы наблюдаем $m = 8$, если считать эти данные независимыми. Однако мы знаем, что независимых биометрических данных не бывает. Соответственно мы можем вычислить среднее значение модуля корреляции исследуемых биометрических данных. Предположим, что этот показатель дал значение r = 0.5, тогда по номограмме рисунка 7 относительная ошибка наблюдения числа степеней свободы составит $\Delta m/(m\text{-}1) = 0.2$. Решая уравнение $\Delta m/(m\text{-}1) = 0.2$ для $m = 8$, получаем $\Delta m = 1.4$, то есть число степеней свободы для системы с зависимыми данными r = 0.5 составит $m = 9{,}4$ вместо наблюдаемой размерности $m = 8$ при независимых данных.

Из номограммы рисунка 7 видно, что для малых значений размерности m связь относительной ошибки наблюдения Δm с коэффициентом корреляции сложная (нелинейная), однако по мере увеличения числа степеней свободы эта связь упрощается и становится близка к линейной. При больших значениях коррелированности данных ошибка наблюдения размерности Δm может быть как угодно большой.

Заключение

Таким образом, асимметричные плотности распределения значений биометрических параметров вполне могут быть приближены модифицированным хи-квадрат распределением. При этом переход от целых показателей числа степеней свободы к их дробным значениям не является существенной проблемой. Так же сравнительно легко осуществляется пересчет хи-квадрат распределения с независимыми данными к хи-квадрат распределениям зависимых биометрических данных. Видимо, хи-квадрат распределение является наиболее простым асимметричным законом распределения значений с точки зрения учета влияния корреляционных связей на форму закона распределения.

ЛИТЕРАТУРА:

1. Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data /*Yevgeni Dodis, Leonid Reyzin, Adam Smith* // April 13, 2004. www.cs.bu.edu/~reyzin/fuzzy.html
2. . *Cavoukian, A. Stoianov* Biometric Encryption: A Positive-Sum Technology that Achieves Strong Authentication, Security AND Privacy, March 2007, Canada, Toronto, Ontario, www.ipc.on.ca
3. Волчихин В.И., Иванов А.И., Фунтиков В.А. Быстрые алгоритмы обучения нейросетевых механизмов биометрико-криптографической защиты информации. Монография. Пенза-2005 г. Издательство Пензенского государственного университета, 273 с.
4. ГОСТ Р 52633-2006 «Защита информации. Техника защиты информации. Требования к средствам высоконадежной биометрической аутентификации»
5. ГОСТ Р 52633.1-2009 Окончательная редакция «Защита информации. Техника защиты информации. Требования к формированию баз естественных биометрических образов, предназначенных для тестирования средств высоконадежной биометрической аутентификации». Начало публичного обсуждения с 15.10.08, окончание публичного обсуждения 15.01.09.
6. Болл Руд и др. Руководство по биометрии. / Болл Руд, Коннел Джонатан Х., Панканти Шарат, Ратха Налини К., Сеньор Эндрю У. // Москва: Техносфера, 2007. -368 с., (перевод с английского)
7. Захаров О.С., Иванов А.И. Учет корреляционных связей биометрических данных через дробный показатель степеней свободы закона распределения значений хи-квадрат. Инфокоммуникационные технологии Том 6, № 1, 2008 г., с. 12-15.
8. Захаров О.С., Иванов А.И. Использование закона распределения хи-квадрат для аналитического описания статистик биометрических параметров. Инфокоммуникационные технологии, Том 7, № 1, 2009 г., с. 72-76.