BULLETIN OF NATIONAL ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
ISSN 1991-3494
Volume 5, Number 5(2014), 28 – 34

UDC: 519.7; 519.66; 612.087.1

ALGORITHM OF ARTIFICIALLY INCREASING THE NUMBER OF DEGREES OF FREEDOM IN THE ANALYSIS OF BIOMETRIC DATA BY CHI-SQUARED CONSENT

B.S.Akhmetov[1], A.I.Ivanov[2], N.I. Serikova[3] , Yu.V. Funtikova[3]
b_akhmetov@ntu.kz, ivan@pniei.penza.ru
[1]Kazakh national technical university named after K.I.Satpayev, Almaty
[2]Penza scientific-research electrotechnical institute, Russia
[3]Penza university, Penza, Russia

Key words: the selection of the biometric data, artificial neural network, assessment of the reliability of statistical hypotheses, the Chi-square.

Abstract. The procedure of "smoothing" histograms in assessing the reliability of statistical hypotheses is considered. It is shown that for small samples, classical histograms poorly approximate the observed law of distribution of biometric parameters values. Smoothing of histograms by digital filter can theoretically make the number of degrees of freedom of the chi-square consent higher than the number of examples in the test sample of biometric data. This allows to increase the power of the chi-square consent and, consequently, increase the accuracy of decisions.

УДК: 519.7; 519.66; 612.087.1

АЛГОРИТМ ИСКУССТВЕННОГО ПОВЫШЕНИЯ ЧИСЛА СТЕПЕНЕЙ СВОБОДЫ ПРИ АНАЛИЗЕ БИОМЕТРИЧЕСКИХ ДАННЫХ ПО КРИТЕРИЮ СОГЛАСИЯ ХИ-КВАДРАТ

Б.С. Ахметов[1], А.И. Иванов[2], Н.И. Серикова[3], Ю.В. Фунтикова[3]

[1]Казахский национальный технический университет имени К.И. Сатпаева, г. Алматы
[2]Пензенский научно-исследовательский электротехнический институт, Россия
[3]Пензенский государственный университет, Россия

Ключевые слова: выборка биометрических данных, искусственные нейронные сети, оценка достоверности статистических гипотез, критерий хи-квадрат.

Аннотация. Рассматривается процедура «сглаживания» гистограмм при оценке достоверности статистических гипотез. Показано, что при малых выборках классические гистограммы плохо приближают наблюдаемый закон распределения значений биометрического параметра. Сглаживание гистограмм цифровым фильтром теоретически позволяет сделать число степеней свободы хи-квадрат критерия согласия выше, чем число примеров в исследуемой выборке биометрических данных. Это позволяет увеличивать мощность хи-квадрат критерия согласия и, соответственно, увеличить достоверность принимаемых решений.

Введение. Классическая статистика создавалась в конце 19 века и начале 20 века. В это время не было возможности создавать сложные алгоритмы обработки данных из-за отсутствия ЭВМ. Ситуация изменилась в конце 20 века, однако ряд заложенных ранее стереотипов в статистических пакетах обработки данных сохранились.

1 Статья подготовлена в рамках выполнения проекта «Исследование вариантов реализации и разработка действующего лабораторного образца ON-LINE системы биометрического обезличивания электронных историй болезней для медицинского учреждения» в соответствии с Приказом Председателя Комитета науки МОН РК №17-нж от 08.04.2013 г

В биометрии, как и в других областях знаний, активно используется хи-квадрат критерий проверки статистических гипотез. Если идти по пути создания классических гистограмм с последующим их использованием для проверки гипотезы нормальности, то требуется выборка от 50 до 200 данных [1], например, полученных предъявлением, соответствующей, базы биометрических образов «Чужой» и/или «Свой» [2].

При минимальном размере 50 примеров можно ожидать, что в динамическом диапазоне наблюдаемого параметра может быть размещено 10 столбиков гистограммы со средним числом попаданий в каждый из интервалов 5 раз.

Если учитывать, что математическое ожидание и среднеквадратическое отклонение вычисляются по этой же выборке, то возможно использование хи-квадрат критерия с 8 степенями свободы.

Мощность критерия хи-квадрат согласия растет с ростом числа степеней свободы (с ростом, числа столбиков гистограммы). Возникает вопрос о том, можно ли на той же самой выборке статистических данных увеличить число степеней свободы хи-квадрат критерия или снизить требования к размерам исходной выборки.

Ответ на этот вопрос положителен, так как люди способны обучаться распознавать образы весьма и весьма эффективно. Человеку достаточно увидеть три-четыре раза один биометрический образ, и он начинает эффективно распознавать его в различных ситуациях. Это эквивалентно тому, что человек обучился (запомнил и может эффективно экстраполировать многомерные статистики биометрических данных) при сложной обработке информации естественными нейронными сетями головного мозга.

При обработке биометрических данных искусственными нейронными сетями [2–4], обученными по ГОСТ Р 52633.5-2011 [5] возникает аналогичная ситуация. Современные нейросетевые преобразователи биометрия-код способны обучаться на 20 примерах образа «Свой» и принимать решения сопоставимые по ошибкам первого и второго рода с решением принимаемым человеком. Это является следствием создания (обучения) и применения сложной нейросетевой обработки данных. Более того, ставится задача снизить размеры обучающей выборки с 20 примеров образа «Свой» до 10 примеров, что делает возможность алгоритмов обучения искусственных нейронных сетей с алгоритмами обучения естественных нейронных сетей, используемых людьми. При этом сдерживающим фактором становятся ставшие классическими каноны традиционной статистической обработки.

Используемые сегодня процедуры статистической обработки, просты, понятны, но дают плохие результаты при исследовании малых обучающих выборок из 8 – 10 примеров.

Снижение ошибки дискретизации статистических данных «сглаживанием» случайных скачков столбиков гистограмм

Будем полагать, что в исследуемой выборке содержатся только данные 8 примеров биометрического параметра. Если предположить, что среднее число попаданий в один из столбиков гистограммы должно быть 2, то динамический диапазон исследуемых данных следует разбивать на 4 интервала. Пример расположения обрабатываемых данных и, соответствующей им гистограммы приведен на рисунке 1.

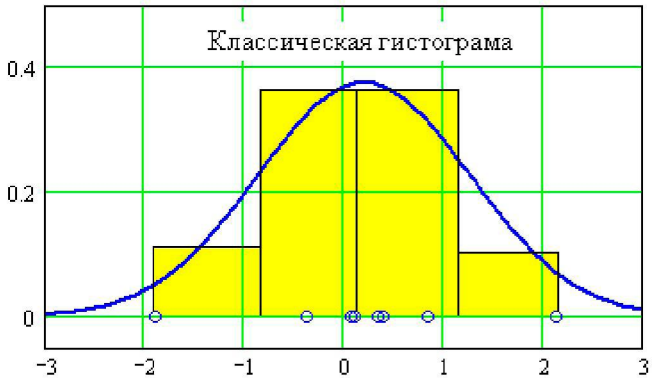

Рисунок 1– Классическая гистограмма представления 8 примеров биометрических данных
(данные получены от генератора случайных чисел с нормальным законом распределения значений)

Из рисунка 1 видно, что шаг квантования данных слишком велик (динамический диапазон данных разбит всего на 4 интервала), как следствие гистограмма имеет сильно отличающиеся по высоте соседние столбики. Для сглаживания данных создадим цифровой фильтр усредняющий результат по окну из 5 наблюдений и размещающий результат в центре окна (в 3 отсчет). Для того, чтобы осуществлять «сглаживание» каждый интервал гистограммы разобьем на четыре интервала внутренний дискретизации. Далее каждому интервалу будем присваивать состояние «0», если он пустой, состояние «1», если туда попал один отсчет. Будем присваивать состояние «2», если в интервал попали два отсчета. Правый и левый полуинтервалы вне динамического диапазона наблюдаемых данных так же разобьем на микро интервалы. В итоге мы получим некоторую цифровую последовательность состояний «0», «1», «2», которую можно подать на сглаживающий данные усреднением цифровой фильтр. Процедура введения дополнительной (не традиционной) дискретизации данных, полученная цифровая последовательность и результат сглаживания приведен на рисунке 2. Из рисунка 2 видно, что после сглаживания результирующая гистограмма будет иметь меньшие ступенчатые скачки, растет так же и число столбцов «сглаженной» гистограммы (число столбцов увеличивается с 4-х до 20). Из-за увеличения числа столбцов с 4 до 20 теоретически возможно увеличить число степеней свободы хи-квадрат критерия согласия с 2 до 18, то есть появляется теоретическая возможность увеличить достоверность статистических оценок без роста размеров исходной выборки.

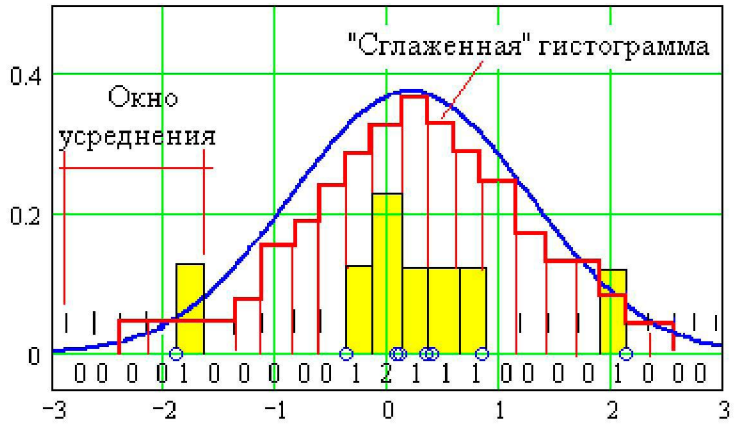

Рисунок 2 – Результат сглаживания потока данных, полученных дополнительным 4-кратными квантованием динамического диапазона и примыкающих полуинтервалов

Казалось бы, что увеличивая число степеней свободы при оценке статистической гипотезы мы получаем некоторою дополнительную информацию. Чем больше мы используем дополнительных искусственных микро квантов, тем больше будет выигрыш. К сожалению,

линейной зависимости нет. Наблюдается некоторая экспоненциальная зависимость с насыщением. Только первые шаги по увеличению числа квантов оказываются эффективны и дают ощутимый выигрыш в получаемой информации, далее наступает участок насыщения, и прирост информации прекращается. При наступлении некоторого предела дальнейший рост числа микро квантов приводит к росту ошибки (к снижению получаемой информации). Имеется явно выраженный максимум информативности, описанных выше, процедур цифрового сглаживания.

Синтез хи-квадрат распределений для зависимых данных

Основной причиной ошибок является то, что классический хи-квадрат критерий введен Пирсоном в 1904 году для независимых данных. Насколько критерий независимости экспериментальных данных работает, обычно не проверяют, однако считается хорошим тоном придерживаться именно гипотезы независимости. К сожалению, для биометрических данных, гипотеза независимости не работает. Даже если формировать случайные биометрические образы «Чужой», воспроизводя случайные рукописные пароли их биометрические данные, оказываются коррелированными (зависимыми) [6, 7]. Пожалуй, только специалисты в области криптографии, имеют полное и безоговорочное право применять гипотезу независимости данных, если эти данные предварительно зашифрованы или осуществлено их криптографическое хеширование. В иных приложениях наблюдаются остаточные корреляционные связи, нуждающиеся в учете.

Причиной появления ошибок «сглаживания» данных является то, что они становятся зависимыми. Чем длиннее окно усредняющего цифрового фильтра, тем сильнее связаны (коррелированны) его выходные данные. В первом приближении коррелированность отсчетов можно оценивать через отношение среднего значения ступенек входных и выходных данных цифрового фильтра:

$$r \approx \left\{1 - \left\{\frac{E(\Delta_{ВЫХ})}{E(\Delta_{ВХ})}\right\}^2\right\}, \qquad (1)$$

где $E(\Delta_{ВЫХ})$ - математическое ожидание скачков столбцов выходной «сглаженной» гистограммы; $E(\Delta_{ВЫХ})$ - математическое ожидание скачков столбцов входной классической гистограммы.

Если коррелированность данных (1) значительна, то использовать классический критерий хи-квадрат Пирсона нельзя, так как он работает только в рамках гипотезы независимости. Выход из создавшегося положения состоит в синтезе хи-квадрат распределений зависимых данных [8, 9]. Для этой цели необходимо создать машину имитации зависимых данных, состоящую из n программных генераторов нормального белого шума - ξk, данные которых связываются между собой путем умножения на матрицу одинаковых элементов с единичной диагональю:

$$\begin{bmatrix} 1 & a & \cdots & a \\ a & 1 & \cdots & a \\ \cdots & \cdots & \cdots & \cdots \\ a & a & \cdots & 1 \end{bmatrix} \times \begin{bmatrix} \xi_{1,i} \\ \xi_{2,i} \\ \cdots \\ \xi_{n,i} \end{bmatrix} = \begin{bmatrix} y_{1,i} \\ y_{2,i} \\ \cdots \\ y_{n,i} \end{bmatrix} \quad (2), \qquad R = \begin{bmatrix} 1 & r & \cdots & r \\ r & 1 & \cdots & r \\ \cdots & \cdots & \cdots & \cdots \\ r & r & \cdots & 1 \end{bmatrix} \quad (3).$$

В конечном итоге получается, что случайные выходные данные оказываются равно коррелированными по отношению друг другу (3). Значение равной коррелированности оказывается монотонной функции единственного регулируемого параметра – a. Если теперь возвести в квадрат центрированные и номерованные случайные данные и просуммировать их, мы получим случайную величину, распределенную по закону Пирсона для зависимых данных - $\chi^2(m, r)$. Примеры изменения формы распределения Пирсона для зависимых данных от значения коэффициента равной корреляции для 3 и 4 степеней свободы даны на рисунке 3.

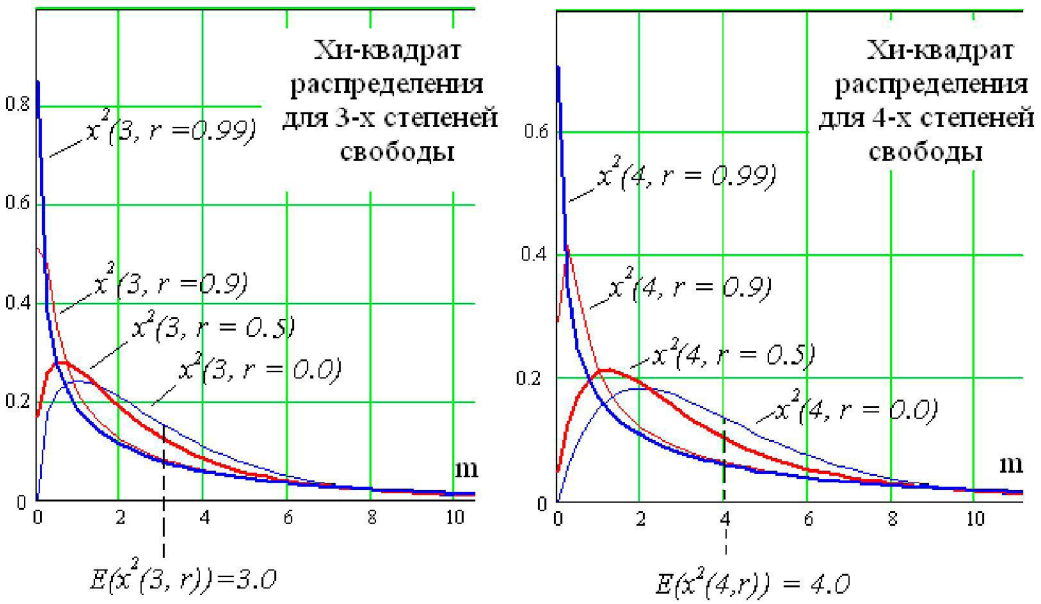

Рисунок 3 – Кривые хи-квадрат распределений для трех и четырех степеней свободы при разных значениях коррелированности биометрических данных

Из рисунка 3 видно, что хи-квадрат распределения зависимых данных имеют математические ожидания точно совпадающее с числом его степеней свободы:

$$E(\chi^2(m,r)) = m \qquad (4)$$

Свойство (4) распределения $\chi^2(m,r)$ сохраняется для любых значений коэффициентов равной коррелированности. Во всем остальном поведение плотностей зависимых распределений $p(\chi^2(m,r \neq 0))$ и не зависимых плотностей $p(\chi^2(m,r=0))$ сильно отличаются. По мере увеличения коррелированности данных фактически происходит частичная утрата числа степеней свободы.

Тем не менее, зная число степеней свободы (число столбцов у классической или «сглаженной» гистограммы) и, соответствующий ей коэффициент равной коррелированности данных мы всегда можем скорректировать критерий хи-квадрат проверки статистических гипотез. Тем самым мы можем существенно увеличить достоверность оценок проверки статистических гипотез.

Решение обратной задачи определения коррелированности данных заданным «сглаживающим» фильтром.

Получается, что число степеней свободы может быть существенно увеличенно, что является положительным фактором. Однако при этом растет коррелированность данных, что отрицательным фактором снижающим эффективность расчетов. Необходимо искать оптимум. Оптимум может быть найден, если мы будем менять параметры сглаживающего фильтра (число микро квантов, ширину окна усреднения) и параллельно просчитывать квантили хи-квадрат критерия проверки статистических гипотез для разного значения степеней свободы.

Задача оптимизации была бы простой, если бы мы умели оценивать значения равной коррелированность данных на выходе сглаживающего фильтра. Выражение (1) для этой цели не подходит из-за его приближенного характера. Более точно решить задачу оценки равной коррелированности данных удается, если осуществить многократное имитационное моделирование работы «сглаживающего» гистограмму цифрового фильтра. Схема численного эксперимента приведена на рисунке 4.

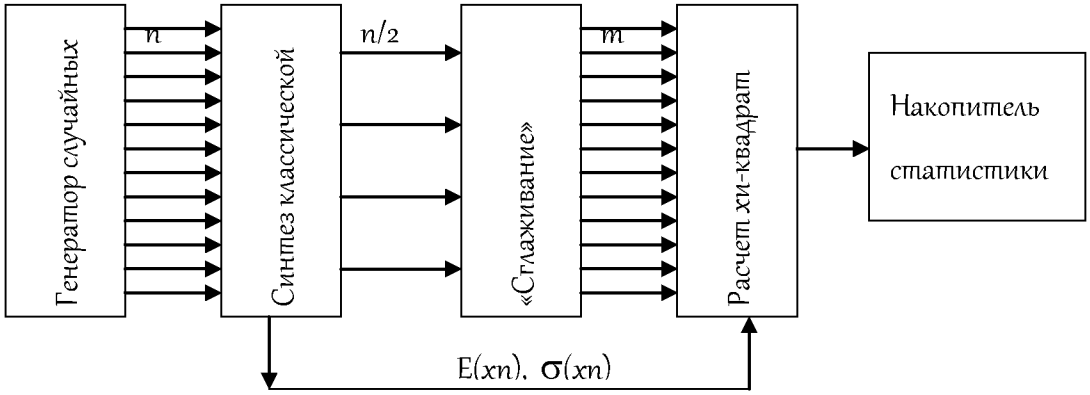

Рисунок 4– Накапливание данных, позволяющих оценить коэффициент равной коррелированности «сглаживающего» фильтра

В соответствии с блок-схемой рисунка 4 необходимо многократно синтезировать выборку нормальных данных (n=8 для рассмотренного выше случая). Далее по 8 отсчетам следует синтезировать классическую гистограмму. Далее следует осуществить сглаживание данных, получив гистограмму с m=20 степенями свободы (рис.2). Параллельно необходимо вычислять математическое ожидание - E(xn) и среднеквадратическое отклонение σ(xn) входных данных. Располагая этими данными, следует найти значение хи-квадрат критерия. Повторив многократно перечисленные выше операции, получим плотность распределения значений p(χ2(m, r)) отличную от классического распределения Пирсона. После этого достаточно обратиться к заранее вычисленным таблицам распределений p(χ2(m, r)) для поиска ближайшего распределения.

В итоге мы имеем оценку равнокоррелированности данных, вносимых усредняющим фильтром, работа которого иллюстрируется рисунком 2.

Выводы. Таким образом, задача искусственного увеличения числа степеней свободы при использовании хи-квадрат критерия согласия является корректной и вполне может быть реализована в современных программных пакетах статистической обработки биометрических данных.

ЛИТЕРАТУРА

[1] Кобзарь А.И. Прикладная математическая статистика для инженеров и научных работников. М.: ФИЗМАТЛИТ, 2006 г., 816 с.

[2] Волчихин В.И., Иванов А.И., Фунтиков В.А., Назаров И.Г., Язов Ю.К. Нейросетевая защита персональных биометрических данных. // М.: Радиотехника, 2012, 157 с., ISBN 978-5-88070-044-8.

[3] B. Akhmetov, A. Doszhanova, A. Ivanov, T. Kartbaev and A.Malygin. Biometric Technology in Securing the Internet Using Large Neural Network Technology. // World Academy of Science, Engineering and Technology. Issue 79, July, 2013, Singapore, p. 129-138, pISSN 2010-376X, eISSN 2010-3778, www.waset.org.

[4] Иванов А.И., Кисляев С.Е., Гелашвили П.А. Искусственные нейронные сети в биометрии, медицине, здравоохранении. Самара: ООО «Офорт», 2004, 236 с.

[5] ГОСТ Р 52633.5-2011 «Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия-код доступа»

[6] B. Akhmetov, A. Ivanov, V. Funtikov, I. Urnev. Evaluation of Multidimensional Entropy on Short Strings of Biometric Codes with Dependent Bits. // «Progress in Electromagnetics Research Symposium» PIERS Proceedings, August 19-23, Moscow, RUSSIA 2012, p.66-69.

[7] Ахметов Б.С., Надеев Д.Н., Урнев И.В. Сериков И.В. Аппроксимация биномиального зависимого закона композициями нормального, равномерного, арксинусного распределения значений. М.: Радиотехника, «Нейрокомпьютеры: разработка, применение», №3, 2012. С. 17-20.

[8] Ахметов Б.С., Иванов А.И., Урнев И.В., Сериков И.В., Газин А.И. Оценка значений числа степеней свободы статистик описания выходного кода преобразователя биометрия-код при использовании распределения χ2. Алматы: Изд-во КазНТУ имени К.И. Сатпаева, http://portal.kaznu.kz/files/publicate/2013-04-05-elbib.pdf

[9] Фунтикова Ю.В., Иванов А.И., Захаров О.С. Гипотеза χ2 распределения расстояний Хэмминга для кодов биометрической аутентификации примеров образа «Свой». // Труды научно-технической конференции кластера

пензенских предприятий, обеспечивающих БЕЗОПАСНОСТЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ. Пенза: Изд-во Пензенского гос. ун-та, 2014. Том 9, , с. 7-8., http://www.pniei.penza.ru/RV-conf/T9/C7.

REFERENCES

[1] Kobzar' A.I. Prikladnaya matematicheskaya statistika dlya inzhenerov i nauchnyh rabotnikov. M.: Fizmatlit, 2006 g., 816 s.

[2] Volchihin V.I., Ivanov A.I., Funtikov V.A., Nazarov I.G., Yazov Yu.K. Neirosetevaya zawita personalnih biometricheskih dannyh. // M.: Radiotehnika, 2012, 157 p., ISBN 978-5-88070-044-8.

[3] B. Akhmetov, A. Doszhanova, A. Ivanov, T. Kartbaev and A.Malygin. Biometric Technology in Securing the Internet Using Large Neural Network Technology. // World Academy of Science, Engineering and Technology. Issue 79, July, 2013, Singapore, p. 129-138, pISSN 2010-376X, eISSN 2010-3778, www.waset.org.

[4] Ivanov A.I., Kislyaev S.E., Gelashvili P.A. Iskusstvennie neironnye seti v biometrii, medicine, zdravoohranenii. Samara: OOO «Ofort», 2004, 236 p.

[5] GOST R 52633.5-2011 «Zashhita informacii. Tehnika zashhity informacii. Avtomaticheskoe obuchenie nejrosetevyh preobrazovatelej biometrija-kod dostupa»

[6] B. Akhmetov, A. Ivanov, V. Funtikov, I. Urnev. Evaluation of Multidimensional Entropy on Short Strings of Biometric Codes with Dependent Bits. // «Progress in Electromagnetics Research Symposium» PIERS Proceedings, August 19-23, Moscow, RUSSIA 2012, p.66-69.

[7] Ahmetov B.S., Nadeev D.N., Urnev I.V. Serikov I.V. Approksimacija binomial'nogo zavisimogo zakona kompozicijami normal'nogo, ravnomernogo, arksinusnogo raspredelenija znachenij. M.: Radiotehnika, «Nejrokomp'jutery: razrabotka, primenenie», №3, 2012. S. 17-20.

[8] Ahmetov B.S., Ivanov A.I., Urnev I.V., Serikov I.V., Gazin A.I. Ocenka znachenij chisla stepenej svobody statistik opisanija vyhodnogo koda preobrazovatelja biometrija-kod pri ispol'zovanii raspredelenija $\chi 2$. Almaty: Izd-vo KazNTU imeni K.I. Satpaeva, http://portal.kazntu.kz/files/publicate/2013-04-05-elbib.pdf

[9] Funtikova Ju.V., Ivanov A.I., Zaharov O.S. Gipoteza $\chi 2$ raspredelenija rasstojanij Hjemminga dlja kodov biometricheskoj autentifikacii primerov obraza «Svoj». // Trudy nauchno-tehnicheskoj konferencii klastera penzenskih predprijatij, obespechivajushhih BEZOPASNOST" INFORMACIONNYH TEHNOLOGIJ. Penza: Izd-vo Penzenskogo gos. un-ta, 2014. Tom 9, , s. 7-8., http://www.pniei.penza.ru/RV-conf/T9/C7.

ХИ-КВАДРАТ КЕЛІСІМ КРИТЕРИІ БОЙЫНША БИОМЕТРИЯЛЫҚ ДЕРЕКТЕРДІ ТАЛДАУ КЕЗІНДЕ БОСТАНДЫҚ ДӘРЕЖЕЛЕР САНЫН ЖАСАНДЫ ЖОҒАРЛАТУ АЛГОРИТМІ

Б.С. Ахметов1, А.И. Иванов2, Н.И. Серикова3, Ю.В. Фунтикова3

Тірек сөздер: биометриялық деректерді таңдау, жасанды нейронды желілер, статистикалық гипотезалардың шынайлығын бағалау, хи-квадрат критерийі.

Аннотация. Хи-квадрат критериін қолдану кезінде еркіндіктің дәрежесінің санын жасанды көбейту есебі қарастырылған.

Поступила 08.08.2014 г.