Надежность № 2 2016
Dependability no.2 2016

Оригинальная статья
УДК: 519.24; 57.017
DOI: 10.21683/1729-2640-2016-16-2-43-48

Оценка качества малой выборки биометрических данных с использованием более экономичной формы хи-квадрат критерия

Берик Б. Ахметов, *Международная академия информатизации (МАИН), Туркестан, Казахстан, e-mail: berik.akhmetov@ayu.edu.kz*

Александр И. Иванов, *лаборатория биометрических и нейросетевых технологий ОАО «Пензенский научно-исследовательский электротехнический институт», Пенза, Россия, e-mail: ivan@pniei.penza.ru*

Берик Б. Ахметов

Александр И. Иванов

Резюме. Цель. *Поставлена цель повышения мощности хи-квадрат критерия Пирсона с тем, чтобы этот критерий стал работоспособен на тестовых выборках биометрических данных малого объема. Необходимо снизить объем тестовой выборки с 200 примеров до 20 примеров при сохранении вероятностей ошибок первого и второго рода. Сбор 20 примеров биометрических образов рассматривается пользователями как комфортный уровень трудозатрат. Необходимость сбора большего числа примеров воспринимается пользователями негативно.* ***Методы.*** *В статье предложена еще одна (вторая) форма критерия Пирсона гораздо менее чувствительная к объему данных в тестовой выборке. Показано, что традиционная форма хи-квадрат критерия чувствительнее к объему тестовой выборки, чем критерий Крамера фон-Мизеса. Предложенная (вторая) форма хи-квадрат критерия имеет чувствительность к объему тестовой выборки меньше чем чувствительность у классического хи-квадрат критерия и меньше, чем чувствительность у критерия Крамера-фон Мизеса одновременно. Этот эффект достигнут переходом из пространства частот появления событий и вероятностей появления группы похожих событий в пространство более точно оцениваемых младших статистических моментов (математического ожидания и стандартного отклонения). Фрактальная размерность новой синтезированной формы хи-квадрат критерия совпадает с фрактальной размерностью классической формы хи-квадрат критерия.* ***Результаты.*** *Предположительно, что предложенный в статье еще один вариант хи-квадрат критерия является одним из самых мощных из всех существующих статистических критериев. Дано аналитическое описание соотношения стандартных отклонений классической формы хи-квадрат критерия и новой формы хи-квадрат критерия. Стандартное отклонение второй формы хи-квадрат критерия уменьшается примерно в 2 раза при сохранении математического ожидания на выборках одинакового объема. Последнее эквивалентно четырехкратному снижению требований к объему тестовой выборки в интервале от 16 до 20 примеров. Выигрыш по мощности от применения нового критерия растет по мере роста объема тестовой выборки.* ***Выводы.*** *Пирсон при создании в 1900 году классического хи-квадрат критерия ориентировался на ограниченные вычислительные возможности, существовавшей тогда вычислительной техники, и был вынужден опираться на найденные им аналитические соотношения. Сегодня ситуация изменилась, ограничения на привлекаемые вычислительные ресурсы исчезли. Однако мы продолжаем по инерции опираться на то, что было создано под вычислительные ресурсы 1900 года. Видимо, следует пытаться учитывать современные возможности вычислительной техники и строить более мощные варианты статистических критериев. Даже если новые критерии будут требовать перебора большого числа возможных состояний (будут иметь большие заранее вычисленные таблицы вместо аналитических соотношений) это сегодня не является сдерживающим фактором. Когда данных недостаточно (в биометрии, в медицине, в экономике) вычислительная сложность статистических критериев не играет особой роли, если результат оценок оказывается более точным.*

Ключевые слова: *многомерный статистический анализ, хи-квадрат критерий, малые выборки тестовых биометрических данных.*

Формат цитирования: *Ахметов Б.Б., Иванов А.И. Оценка качества малой выборки биометрических данных с использованием более экономичной формы хи-квадрат критерия // Надежность. 2016, №2. С. 43-48. DOI: 10.21683/1729-2640-2016-16-2-43-48*

Введение

Информационное общество предполагает активное использование Интернет-ресурсов. Государственные и частные структуры создают на своих сайтах личные кабинеты пользователей. К сожалению, существующая практика парольной защиты доступа к личным кабинетам обладает существенными уязвимостями. Пользователи не способны запоминать длинные случайные пароли. Владелец информационного ресурса не может быть уверен

в том, что к личному электронному кабинету получил доступ именно его хозяин. Пароль может быть перехвачен программной закладкой, также не составляет проблемы подменить IP адрес Интернет-пользователя.

Для усиления защиты доступа к электронным кабинетам в настоящее время разрабатываются технологии биометрической аутентификации личности путем преобразования личных биометрических данных человека в его криптографический ключ или длинный случайный пароль доступа. Используются такие биометрические образы, как: рисунок отпечатка пальца [1], рисунок радужной оболочки глаза [2], голосовой пароль [3], рукописный пароль [4], рисунок кровеносных сосудов глазного дна или ладони руки [5]. Естественно, что преобразователи биометрия-код не могут быть идеальными и имеют вероятности ошибок первого и второго рода. Возникает необходимость тестирования ошибок первого и второго рода на реальных биометрических данных. Кроме того, при настройке «нечетких экстракторов» [1, 2, 3] и при обучении нейросетевых преобразователей [4, 5] необходимо контролировать отсутствие в биометрических данных грубых ошибок. По сути дела, на небольшом числе примеров биометрического образа необходимо контролировать показатель близости распределения биометрических данных к многомерному нормальному закону [6]. Формально для этой цели может быть использован обычный одномерный хи-квадрат критерий Пирсона [7, 8], однако такой подход далек от оптимального. В данной статье мы попытаемся показать, что классическая форма критерия Пирсона является далеко не единственной, то есть возможна постановка задачи по поиску более эффективных функционалов Пирсона, учитывающих различные особенности их практического применения.

Появление шумов квантования при статистической обработке малых выборок

Рассмотрим простейшую ситуацию, когда тестовая или обучающая выборка представлена 9 примерами образа «Свой». Из-за того, что непрерывная функция вероятности $P(x)$ первого биометрического параметра v_1 – малой выборки, мы вынуждены описывать ее ступенчатой монотонно возрастающей функцией $\tilde{P}(x)$, как это показано в левой части рисунка 1.

Для того, чтобы построить ступенчатое монотонно возрастающее приближение $\tilde{P}(x)$, необходимо осуществить сортировку биометрических данных по их возрастанию:

$$x_i = sort(v_{1,i}) \text{ для } i = 0, 1, 2, \ldots, n, \qquad (1)$$

где n – размер тестовой выборки или число квантов приближения монотонной функции вероятности.

В этом случае монотонно возрастающая ступенчатая функция будет описываться следующим кусочно-постоянным приближением:

$$\tilde{P}(x_i) = \frac{i}{n}. \qquad (2)$$

Ошибка приближения или шум квантования находится как разность непрерывной функции вероятности и ее ступенчатого приближения:

$$\Delta P(x) = P(x) - \tilde{P}(x). \qquad (3)$$

В нижней части рисунка 1 отображены функции ошибки квантования или шумы квантования, возникающие из-за малых тестовых выборок.

В контексте вышеизложенного, статистический критерий Колмогорова-Смирнова [7] следует рассматривать как поиск максимального значения модуля ошибки приближения:

$$\sup_{-\infty<v<+\infty} \left|P(x) - \tilde{P}(x)\right| = \max \left|\Delta P(x_i)\right| \qquad (4)$$

или выбор наибольшего из локальных максимумов шума квантования.

С этих же позиций статистический критерий Крамера-фон Мизеса [7] является оценкой стандартного отклонения шума квантования непрерывной функции вероятности:

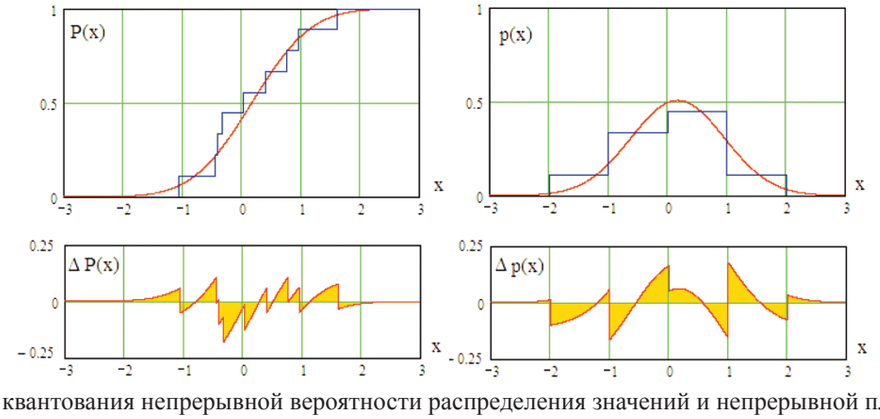

Рис. 1. Эффекты квантования непрерывной вероятности распределения значений и непрерывной плотности распределения значений путем их представления 9 примерами, порождающие непрерывный шум ошибки квантования

$$\int_{-\infty}^{\infty}\left\{P(x)-\tilde{P}(x)\right\}^2\cdot dx=\int_{-\infty}^{\infty}\left\{E(\Delta P(x))-\Delta P(x)\right\}^2\cdot dx=$$

$$=\int_{-\infty}^{\infty}\left\{\Delta P(x)\right\}^2\cdot dx=\sigma^2(\Delta P(x)), \qquad (5)$$

если выполняется условие нулевого математического ожидания шума квантования $E(\Delta P(x))=0$.

Следует подчеркнуть, что статистический критерий Колмогорова-Смирнова (4) всегда имеет меньшую мощность в сравнении с критерием Крамера-фон Мизеса (5). Критерий Колмогорова-Смирнова (4) – точечный, а критерий Крамера-фон Мизеса (5) –интегральный.

Очевидно, что с ростом размеров n тестовой выборки оба эти статистические критерии набирают мощность оценок, однако оценка по интегральному критерию всегда оказывается надежнее, чем оценка по точечному критерию.

Классический вариант критерия хи-квадрат Пирсона

Общая практика проверки статистических гипотез в большинстве отраслей промышленности сводится к построению гистограмм имеющихся данных (правая часть рисунка 1) и вычислению классического хи-квадрат критерия:

$$\chi^2=\sum_{i=1}^{k}\left\{\frac{\left(\frac{n_i}{n}-P_i\right)^2}{P_i}\right\}=\sum_{i=1}^{k}\left\{\frac{\left(\tilde{P}_i-P_i\right)^2}{P_i}\right\}, \qquad (6)$$

где n_i – число отсчетов, попавших в i-тый столбец гистограммы, P_i – вероятность попадания в i-тый столбец гистограммы теоретического распределения, k – число столбцов гистограммы.

Широкое распространение применения хи-квадрат критерия обусловлено тем, что для него известно аналитическое описание плотности распределения:

$$p(\chi^2,m)=\frac{1}{m}\left\{\frac{1}{2^{\frac{m}{2}}\cdot\Gamma\left(\frac{m}{2}\right)}\left\{x^{\frac{m}{2}-1}\cdot\exp\left(\frac{-x}{2}\right)\right\}\right\}, \qquad (7)$$

где $\Gamma(\cdot)$ – гамма функция, m – число степеней свободы.

Число степеней свободы m может быть задано поразному [8]. Например, оно может быть определено через объем n тестовой выборки:

$$m=\sqrt{n}-3=k-3, \qquad (8)$$

если число k столбцов гистограммы выбирается округлением до ближайшего целого величины $\sqrt{n}$:

$$k=round\left(\sqrt{n}\right). \qquad (9)$$

Заметим, что значение числа k столбцов гистограммы и значение числа m степеней свободы для классического хи-квадрат критерия всегда оказывается много меньшими по сравнению с объемом n тестовой выборки. Получается, что погрешность ступенчатого приближения плотности распределения значений $\Delta p(x)=p(x)-\tilde{p}(x)$ (правая часть рисунка 1) всегда оказывается больше, чем ошибка приближения функции вероятности (3). Так, в левой части рисунка 1 приближение функции вероятности строится с использованием 9 ступенек, тогда как функция приближения плотности распределения вероятности строится с использованием только 4 ступенек в правой части рисунка 1. Шумы квантования критерия Крамера-фон Мизеса оказываются меньше, чем шумы квантования классического критерия хи-квадрат (6).

Это означает, что мощность критерия Крамера-фон Мизеса оказывается всегда выше, чем мощность классического хи-квадрат критерия Пирсона (3).

Сравнение по мощности критерия Крамера-фон Мезиса и классического критерия хи-квадрат

Будем исходить из того, что биометрические данные по каждому из контролируемых параметров распределены нормально. Тогда качество данных одного параметра можно оценивать и по критерию Крамера-фон Мезиса и по критерию хи-квадрат [7, 8]. Для сравнения критериев как альтернативу будем использовать распределение данных по равномерному закону. Результаты численного моделирования для выборок из 9 примеров приведены на рисунке 2.

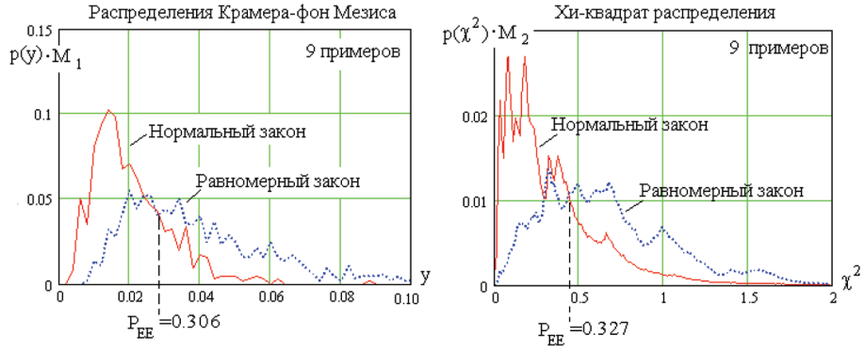

Рис. 2. Распределения значений критерия Крамера-фон Мезиса и хи-квадрат критерия для нормального закона распределения и его альтернативы в виде равномерного закона распределения для выборок из 9 примеров

При принятии решения важным является порог сравнения. Каждый порог сравнения дает свое значение вероятности P_1 ошибок первого рода и вероятности P_2 ошибок второго рода. Для исключения неопределенности порога сравнения будем сравнивать результаты в точке с равной вероятностью ошибок $P_1 = P_2 = P_{EE}$.

Из рисунка 2 видно, что распределение данных, полученных по критерию Крамера-фон Мизеса, дает значение $P_1 = P_2 = P_{EE} = 0{,}306$. При тех же условиях хи-квадрат критерий дает значение равновероятных ошибок $P_1 = P_2 = P_{EE} = 0{,}327$. Результат оказывается хуже примерно на 9%. Это означает, что хи-квадрат критерий требует выборки из 10 примеров, тогда как для критерия Крамера-фон Мизеса потребуется только 9 примеров. Снижение требований к размерам тестовой выборки обусловлено тем, что ошибка квантования функции вероятности $P(x)$ оказывается меньше ошибки квантования плотности распределения $p(x)$ (см. рисунок 1).

Вычислительная процедура критерия Крамера-фон Мизеса примерно в $\sqrt{n}$ раз эффективнее подавляет влияние шумов квантования по сравнению с процедурой вычисления данных по критерию хи-квадрат. Чем больше тестовая выборка, тем сильнее сказывается эффект более сильного подавления влияния шумов квантования. На рисунке 3 приведены данные моделирования для выборки, состоящей из 32 примеров.

Из рисунка 3 видно, что для выборки из 32 примеров критерий Крамера-фон Мезиса дает $P_{EE} = 0{,}111$, что на 43% меньше, чем дает критерий хи-квадрат: $P_{EE} = 0{,}196$. В первом приближении можно ожидать снижения объемов тестовой выборки примерно на 40%, если перейти от применения критерия хи-квадрат к критерию Крамера-фон Мезиса.

Еще один вариант хи-квадрат критерия, более эффективно использующий ограниченный объем тестовой выборки

По сути дела и критерий Крамера-фон Мезиса, и критерий хи-квадрат Пирсона являются некоторыми схемами подавления влияния шумов квантования. То есть, можно попытаться усилить у этих статистических функционалов свойство подавлять шумы квантования. Например, можно использовать дополнительный цифровой фильтр, настроенный на сглаживание скачков кусочно-постоянного приближения функции плотности распределения значений [9, 10].

Еще одним путем является проверка других возможных вариантов вычисления хи-квадрат критерия. В частности, в биометрии достаточно давно используются так называемые функционалы Пирсона (сети функционалов Пирсона [11]) для предварительно нормированных по стандартному отклонению данных:

$$\chi_2{}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\left(E(x)-x_i\right)^2}{\sigma(x)}\right\}, \qquad (12)$$

где $E(x)$ – математическое ожидание данных тестовой выборки, $\sigma(x)\approx 1$ – стандартное отклонение предварительно нормированных данных тестовой выборки.

Следует обратить внимание на то, что при обработке биометрических данных обычно осуществляют их предварительную нормировку по стандартному отклонению, стремясь выполнить условие $\sigma(x)=1$. Однако, на малых выборках выполнить это условие не удается. В частности, относительная ошибка вычисления стандартного отклонения на малых выборках объемом в 20 примеров случайна и может составлять до ±30%, при меньших выборках ошибка может быть еще больше. Для компенсации ошибок нормировки в формуле (12) появляется знаменатель, близкий к единице, но всегда отличный от нее.

Заметим, что выражение (12) осуществляет суммирование квадратичных отклонений по всем отсчетам тестовой выборки, тогда как классический хи-квадрат критерий (6) суммирует квадратичные отклонения только по числу столбцов гистограммы. Так как $n > k$, то можно ожидать более высокой мощности хи-квадрат критерия (12) в сравнении с похожим классическим хи-квадрат критерием (6).

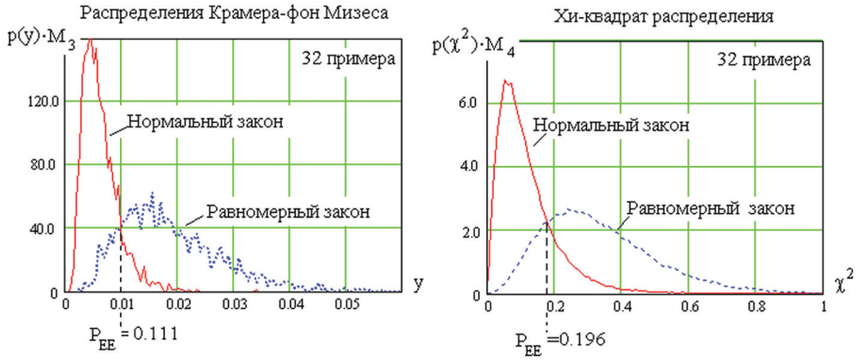

Рис. 3. Распределения значений критерия Крамера-фон Мизеса и хи-квадрат критерия для нормального закона распределения и его альтернативы в виде равномерного закона распределения для выборок из 32 примеров

Сравнение по мощности двух вариантов критерия хи-квадрат

Статистический критерий Крамера-фон Мезиса оказывается мощнее классического критерия хи-квадрат Пирсона из-за того, что он давит шумы квантования меньшей амплитуды (сравниваем левую и правую часть рисунка 1). Однако для критерия Крамера-фон Мизеса нет аналитического описания и это его огромный недостаток. В силу этого обстоятельства далее будем сравнивать между собой только мощности двух модификации критериев хи-квадрат (6) и (12).

Результаты имитационного моделирования при 9 и 32 примерах в тестовой выборке для критерия (12) приведены на рисунке 4.

Если сравнить пересечение распределений на левом графике рисунка 4, дающее $P_{EE} = 0{,}106$, и аналогичное пересечение на правом графике рисунка 2, дающее $P_{EE} = 0{,}327$, то мы получим примерно трехкратный выигрыш по мощности между двумя вариантами критериев (достоверности принимаемым по ним решениям). С ростом объема тестовой выборки выигрыш от применения второй формы хи-квадрат критерия увеличивается. Так, если сравнивать данные пересечения распределений на правом графике рисунка 4, позволяющие достичь $P_{EE} = 0{,}007$, с данными на правом графике рисунка 3, обеспечивающими $P_{EE} = 0{,}196$, то мы получим выигрыш в 28 раз.

Получается, что и критерий Крамера-фон Мизеса, и вторая форма критерия хи-квадрат оказываются мощнее классического хи-квадрат критерия из-за вычисления этих двух критериев по всей выборке. Классический хи-квадрат критерий проигрывает этим двум критериям из-за суммирования им квадратичной ошибки по числу столбцов эмпирической гистограммы. Легко убедиться в том, что критерий Крамера-фон Мизеса по его мощности находится в интервале между мощностями двух форм хи-квадрат критериев.

Аналитическое описание второй формы хи-квадрат критерия

Принципиально важным свойством второй формы хи-квадрат критерия является то, что для независимых данных ее статистические свойства очень хорошо описываются нормальными законами распределения значений. При этом математическое ожидание для распределения формулы (12) при отсутствии в нем нормирующего множителя $1/n$ оказывается близко к размеру тестовой выборки:

$$E(\chi_2^{\,2}) \approx n - 0{,}778. \qquad (13)$$

Это утверждение иллюстрируется положениями максимумов сплошных кривых рисунка 5. То есть значение математического ожидания второй формы хи-квадрат критерия почти наследуется (с точностью до поправки – 0,778) от классического хи-квадрат распределения с числом степеней свободы $m = n$ (точечные кривые на рисунке 5).

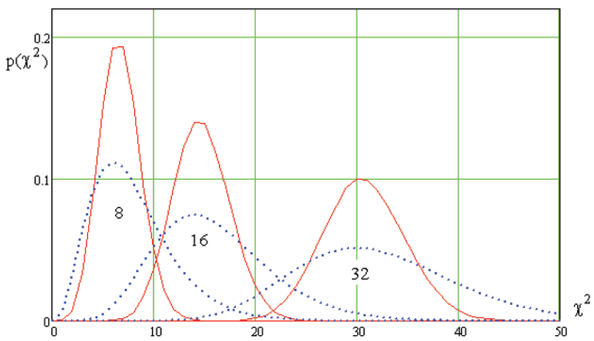

Рис. 5. Распределения значений второй формы хи-квадрат критерия (сплошные кривые) для $n = 8, 16, 32$ и классических хи-квадрат распределений, когда число степеней свободы совпадает с размером выборки (точечные кривые)

Из рисунка 5 видно, что стандартные отклонения нормальных законов распределения оказываются всегда меньше, чем стандартные отклонения классических хи-квадрат распределений Пирсона. При статистических расчетах с инженерной точностью стандартные отклонения распределений описываются следующим простым соотношением:

$$\sigma(n, \chi_2^2) \approx \sqrt{\frac{E(\chi_2^{\,2})}{2}} \approx \sqrt{\frac{n - 0{,}778}{2}} \qquad (14).$$

Следует подчеркнуть, что точность приближения (14) растет по мере нормализации распределений значений

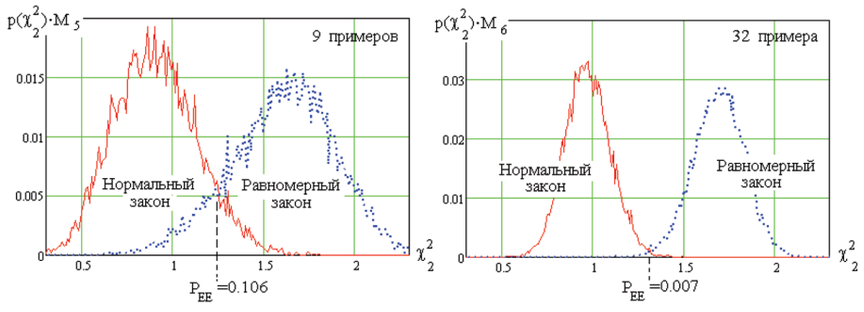

Рис. 4. Гистограммы распределений значений второй более эффективной формы хи-квадрат критерия

второй формы хи-квадрат распределений. Так для $n = 8$, 16, 32 относительная ошибка приближения стандартного отклонения будет составлять $\Delta\sigma = 3{,}10\%, 1{,}70\%, 0{,}75\%$, что вполне приемлемо для инженерной практики статистической обработки биометрических данных.

Заключение

Мы привыкли к тому, что для достоверных статистических оценок требуются выборки, содержащие сотни примеров. Только в том случае, когда мы имеем большую выборку и опираемся на стандартизованные рекомендации [8], появляется уверенность в достоверности проведенного статистического анализа. Такова сложившаяся на данный момент техническая практика.

Как показано в данной статье, мощность хи-квадрат критерия может быть существенно увеличена, то есть достоверные оценки могут быть получены на выборках данных гораздо меньшего объема. Это крайне важно для практики, особенно если производится разрушающий контроль достаточно дорогих изделий. Уже сложившаяся практика статистической обработки биометрических данных подтверждает это важное положение.

Предоставление существенных вычислительных ресурсов сегодня не является проблемой. Мы вполне можем усложнить статистическую обработку реальных данных, например, сделав ее многомерной [11]. На сегодняшний день у нас появилась техническая возможность многократно усложнить используемые нами методы статистической обработки. Если в прошлом веке мы использовали один критерий и вынуждены были довольствоваться его результатами, то сегодня мы можем использовать десятки известных статистических критериев и создавать при необходимости новые критерии под особенности той или иной практической задачи.

Библиографический список

1. Ramírez-Ruiz J., Pfeiffer C., Nolazco-Flores J. Cryptographic Keys Generation Using FingerCodes. // Advances in Artificial Intelligence – IBERAMIA-SBIA 2006 (LNCS 4140), p. 178-187, 2006

1. Monrose F., Reiter M., Li Q., Wetzel S. Cryptographic key generation from voice. In Proc. IEEE Symp. on Security and Privacy, 2001

1. Feng Hao, Ross Anderson and John Daugman. Crypto with Biometrics Effectively, IEEE TRANSACTIONS ON COMPUTERS, VOL. 55, NO. 9, SEPTEMBER 2006.

1. Язов Ю.К. и др. Нейросетевая защита персональных биометрических данных. //Ю.К.Язов (редактор и автор), соавторы В.И. Волчихин, А.И. Иванов, В.А. Фунтиков, И.Г. Назаров // М.: Радиотехника, 2012 г. 157 с. IBSN 978-5-88070-044-8.

1. Ахметов Б.С., Иванов А.И., Фунтиков В.А., Безяев А.В., Малыгина Е.А. Технология использования больших нейронных сетей для преобразования нечетких биометрических данных в код ключа доступа. Монография, Казахстан, г. Алматы, ТОО «Издательство LEM», 2014 г. -144 с.

1. Ахметов Б.С., Волчихин В.И., Иванов А.И., Малыгин А.Ю. Алгоритмы тестирования биометрико-нейросетевых механизмов защиты информации Казахстан, Алматы, КазНТУ им. Сатпаева, 2013 г.- 152 с. ISBN 978-101-228-586-4.

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

1. Р 50.1.037-2002 Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа $\chi 2$. Госстандарт России. Москва-2001 г., 140 с.

1. Ахметов Б.С., Иванов А.И., Серикова Н.И., Фунтикова Ю.В. Алгоритм искусственного повышения числа степеней свободы при анализе биометрических данных по критерию согласия хи-квадрат. Вестник национальной академии наук республики Казахстан. №5, 2014 г. с. 28-:-34.

1. Серикова. Н.И., Иванов А.И., Качалин С.В. Биометрическая статистика: сглаживание гистограмм, построенных на малой обучающей выборке. /Вестник СибГАУ 2014 № 3(55) с.146-150.

1. Ахметов Б.Б., Иванов А.И., Безяев А.В., Фунтикова Ю.В. Многомерный статистический анализ биометрических данных сетью частных критериев Пирсона. // Вестник Национальной академии наук Республики Казахстан. – Алматы, 2015. № 1. С. 5-11.

Сведения об авторах

Берик Б. Ахметов – кандидат технических наук, академик Международной академии информатизации (МАИН), вице-президент «Международного казахско-турецкого университета имени Ходжи Ахмеда Ясави».

Казахстан, 161200 г. Туркестан, проспект Б. Саттарханова, 29, Здание Ректорат, тел. +7 (72533) 3-35-77, e-mail: berik.akhmetov@ayu.edu.kz

Александр И. Иванов – доктор технических наук, доцент – начальник лаборатории биометрических и нейросетевых технологий ОАО «Пензенский научно-исследовательский электротехнический институт».

Россия, 440000 г. Пенза, ул. Советская, 9, тел. +7 (841-2) 59-33-10, e-mail: ivan@pniei.penza.ru