

Всероссийская научно-техническая конференция, посвященная 100-летию со дня рождения Б.И. Рамеева: «Информационно-управляющие и телекоммуникационные системы специального назначения». СЕКЦИЯ: «Биометрическая поддержка криптовалют, блокчейн реестров, облачных сервисов». Доклад состоится 16 мая, 2018 года, с 14⁰⁰-14¹⁵, конференц-зал Технопарка «РАМЕЕВ», ул. Центральная, 1, г. Пенза.

УДК: 519.24; 53; 57.017

ОЦЕНКА СООТНОШЕНИЯ МОЩНОСТЕЙ ХИ-КВАДРАТ НЕЙРОНА И НЕЙРОНА СРЕДНЕГО ГЕОМЕТРИЧЕСКОГО ПРИ ИХ ИСПОЛЬЗОВАНИИ В ПРЕОБРАЗОВАТЕЛЯХ БИОМЕТРИЯ-КОД

К.А. Перфилов, А.И. Газин

Аннотация. Целью представленной работы является описание искусственных нейронов, построенных как аналоги статистического критерия квадрата среднего геометрического плотностей распределения значений многомерных биометрических данных «Свой» и многомерных плотностей распределения значений, предъявленных биометрических данных. Средствами имитационного моделирования на реальных биометрических данных показано, что мощность созданных квадратичных нейронов намного выше, чем мощность классических квадратичных радиально базисных нейронов. При этом важнейшее свойство линейной вычислительной сложности обучения квадратичных нейронов сохранено, что позволяет быстро обучать как угодно большие искусственные нейронные сети среднего геометрического на малых обучающих выборках.

Ключевые слова: нейросетевой преобразователь биометрия-код, биометрические данные, большая размерность данных, мало разрядные вычисления с использованием логарифмов таблиц плотности вероятности.

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, так же 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 длинный случайный пароль доступа. Используются такие биометрические образы как: рисунок отпечатка пальца, рисунок радужной оболочки глаза, голосовой пароль, рукописный пароль, рисунок кровеносных сосудов глазного дна или ладони руки. 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» и при обучении нейросетевых преобразователей 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.

В 1900 году Пирсон предложил хи-квадрат статистический критерий [1], который на сегодняшний день практически стал стандартом [2]. Популярность хи-квадрат критерия Пирсона обусловлена тем, что для больших выборок в 400 и более опытов им была дана аналитическая зависимость плотности распределения значений от числа степеней свободы (от числа столбцов гистограммы экспериментальных данных).

Десятки других статистических критериев [3] на практике куда менее востребованы из-за того, что для них математиками построены таблицы доверительных вероятностей, но нет их точного аналитического описания.

Основная масса таблиц доверительных вероятностей для сотен известных на данный момент статистических критериев перенесены в справочники и стандартизованные рекомендации из первоисточников без независимой серьезной проверки инженерным сообществом. В инженерной среде не принято проверять таблицы доверительных вероятностей, приведенные в справочниках.

В итоге возникает путаница с достоверностью данных, публикуемых в современных статистических справочниках. В этом отношении источник [1] является одним из самых достоверных, так как в нем содержится очень большое число ссылок на первоисточники. По крайней мере, каждый сомневающийся инженер может проследить цепочку ссылок и попытаться найти сведения о независимом подтверждении достоверности таблиц доверительных вероятностей в том или ином первоисточнике.

Тяжесть проблемы состоит в том, что при статистическом анализе биометрических данных приходится настороженно относиться даже к проверенному вдоль и поперек хи-квадрат критерию. Причина состоит в том, что таблицы хи-квадрат критерия для выборки в 16-20 опытов не существует. Еще одной дополнительной проблемой является наличие большого числа предложенных математиками статистических критериев. Часть известных статистических критериев, построенных для интегральных характеристик – сравниваемых функций вероятности приведена в табл. 1. Очевидно, что интегральная функция вероятности - $P(u)$ через дифференциал связана с ее дифференциальным аналогом $p(u)$ - плотностью распределения функции вероятности. В силу линейности операций интегрирования и дифференцирования [4] во всех интегральных статистических критериях таблицы №1 функцию вероятности - $P(u)$ можно заменить на ее дифференциал - $p(u)$. В итоге мы получим табл. 2 дифференциальных статистических критериев.

Таблица 1

Статистические критерии проверки гипотезы о соответствии эмпирической функции вероятности - $P(\tilde{u})$ некоторому ее аналитическому описанию - $\tilde{P}(u)$

№	Название критерия и год создания	Формула вычисления критерия
1	Хи-квадрат критерий Пирсона 1900 г. [1]	$= N \sum_{i=1}^m \left(n_i / N - \tilde{P}_i \right)^2 / \tilde{P}_i$, где N – число опытов, m – число интервалов гистограммы, n_i – число отсчетов в i-том интервале, \tilde{P}_i – теоретическая вероятность попадания в i-тый интервал.
2	Критерий Крамера-фон Мизеса 1928 г. [1]	$= \int_{-\infty}^{+\infty} \left\{ P(\tilde{u}) - \tilde{P}(u) \right\}^2 \cdot du$
3	Критерий Смирнова-Крамера-фон Мизеса 1936 г. [1]	$= \int_{-\infty}^{+\infty} \left\{ P(\tilde{u}) - \tilde{P}(u) \right\}^2 \cdot d\tilde{P}(u)$

4	Критерий Джини 1941 г. [1]	$= \int_{-\infty}^{+\infty} P(\tilde{u}) - \tilde{P}(u) \cdot du$
5	Критерий Андерсона-Дарлинга 1952 г. [1]	$= \int_{-\infty}^{+\infty} \frac{\{P(\tilde{u}) - \tilde{P}(u)\}^2}{\tilde{P}(u) \cdot \{1 - \tilde{P}(u)\}} \cdot d\tilde{P}(u)$
6	Критерий Ватсона 1961 г. [1]	$= \int_{-\infty}^{+\infty} \left\{ \tilde{P}(u) - P(\tilde{u}) - \int_{-\infty}^{+\infty} [\tilde{P}(u) - P(\tilde{u})] \cdot d\tilde{P}(u) \right\}^2 \cdot d\tilde{P}(u)$
7	Критерий Фроцини 1978 г. [1]	$= \int_{-\infty}^{+\infty} P(\tilde{u}) - \tilde{P}(u) \cdot d\tilde{P}(u)$
8	Критерий среднего геометрического, сравниваемых функций вероятности 2014 г. [4]	$= \int_{-\infty}^{+\infty} \sqrt{P(\tilde{u}) \cdot (1 - \tilde{P}(u))} \cdot du$

Подобная замена увеличивает число возможных для использования функционалов обогащения данных. Как показано на рис. 1, в ряде случаев дифференциальные функционалы имеют мощность существенно выше интегральных функционалов, если речь идет о разделении биометрических данных с нормальным законом распределения на фоне альтернативного равномерного закона распределения значений [5, 6].

Как видно из рис. 1, квадрат среднего геометрического сравниваемых функций распределения дает наибольшую мощность (dsg^2), обеспечивая минимальное значение равновероятных ошибок первого и второго рода на малых выборках. Видимо, это самый мощный на текущий момент статистический критерий [4] из известных критериев.

Таблица 2

Статистические критерии проверки гипотезы о соответствии наблюдаемой дифференциальной плотности вероятности $p(u) = \frac{dP(u)}{du}$ некоторому ее аналитическому описанию $\tilde{p}(u)$

№	Название критерия и год создания	Формула вычисления критерия
1	Дифференциальный вариант критерия Крамера-фон Мизеса [4] 2016 г.	$= \int_{-\infty}^{+\infty} \{p(u) - \tilde{p}(u)\}^2 \cdot du$
2	Дифференциальный вариант критерия Смирнова-Крамера-фон Мизеса [4] 2016 г.	$= \int_{-\infty}^{+\infty} \{p(u) - \tilde{p}(u)\}^2 \cdot \tilde{p}(u) \cdot du$
3	Дифференциальный вариант критерия Джини 2006 г. [5, 7, 8, 9] 2006 г.	$= \int_{-\infty}^{+\infty} p(u) - \tilde{p}(u) \cdot du$
4	Интегро-дифференциальный вариант критерия Андерсона-Дарлинга [4] 2016 г.	$= \int_{-\infty}^{+\infty} \frac{\{p(u) - \tilde{p}(u)\}^2}{\tilde{P}(u) \cdot \{1 - \tilde{P}(u)\}} \cdot \tilde{p}(u) \cdot du ;$
5	Дифференциальный вариант критерия Ватсона [4] 2016 г.	$= \int_{-\infty}^{+\infty} \left\{ \tilde{p}(u) - p(u) - \int_{-\infty}^{+\infty} [\tilde{p}(u) - p(u)] \cdot \tilde{p}(u) \cdot du \right\}^2 \cdot \tilde{p}(u) \cdot du$
6	Дифференциальный вариант критерия Фроцини [4] 2016 г.	$= \int_{-\infty}^{+\infty} p(u) - \tilde{p}(u) \cdot \tilde{p}(u) \cdot du$

7	Среднее геометрическое плотностей сравниваемых вероятностей 2016 г. [5, 6]	$= \int_{-\infty}^{+\infty} \sqrt{p(u) \cdot \tilde{p}(u)} \cdot du$
8	Квadrата среднего геометрического плотностей вероятности 2016 г. [5, 6]	$= \int_{-\infty}^{+\infty} p(u) \cdot \tilde{p}(u) \cdot du$

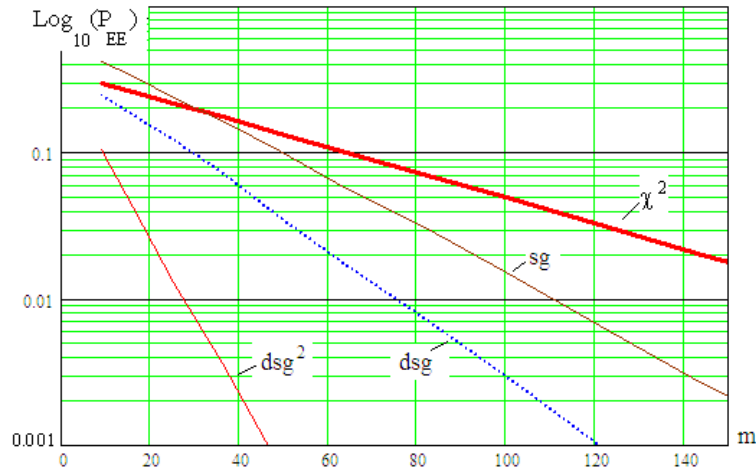


Рис. 1. Эталонная мощность хи-квадрат критерия (толстая линия) в логарифмической шкале равновероятных ошибок, sg – интегральный функционал среднего геометрического (таблица №1 строка 2), dsg – дифференциальный вариант функционала среднего геометрического (таблица №2, строка 8)

Из данных рисунка 1 можно наблюдать, что равновероятные ошибки первого и второго рода $P_1 = P_2 = P_{EE}$ для хи-квадрат критерия достигают значения 0.01 при 160 опытах. Та же самая вероятность ошибок $P_{EE} = 0.01$ для критерия dsg^2 получается на выборке из 27 опытов. Наблюдается 6-ти кратное снижение требований к размеру тестовой выборки, что крайне существенно для биометрических приложений.

При практической реализации многомерного статистического анализа очень удобным оказалось применение искусственных нейронных сетей [10], обучаемых стандартным алгоритмом [11] с линейной вычислительной сложностью и тестируемых после обучения стандартными алгоритмами [12]. Быстрое и абсолютно устойчивое автоматическое обучение может быть организовано не только для сетей из персептронов, но и для иных нейронных сетей, воспроизводящих хорошо исследованные радиально-базисные функции [13] или множество иных, менее изученных, квадратичных функционалов [14 – 20].

Можно представить, что любому известному статистическому критерию (статистическому функционалу) можно поставить в соответствие некоторый нейрон [4]. Их отличие будет состоять только в том, что нейрон требует обучения (настройки) тогда как статистические критерии, как правило, не настраивают (не регулируют) в части предобработки данных. При использовании статистических критериев необходима настройка только порогового элемента (необходимо выбрать значение требуемого показателя доверительной вероятности).

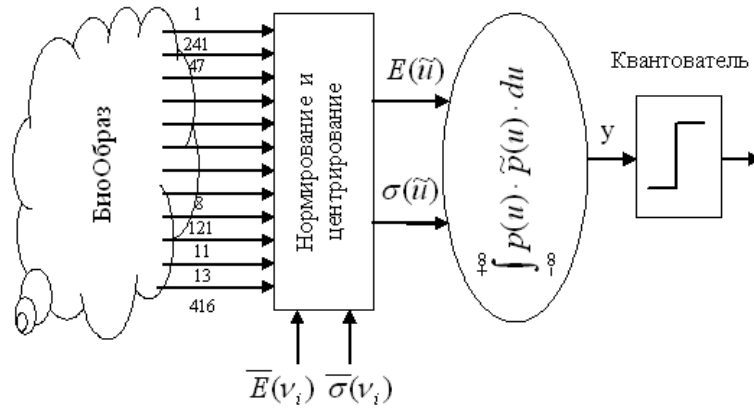


Рис. 2. Структурная схема искусственного нейрона, построенного как эквивалент квадрата среднего геометрического сравниваемых плотностей распределения значений

Так же как все квадратичные функционалы нейрон среднего геометрического со структурой, изображенной на рис. 2, всегда имеет положительный отклик линейной части. Его настройка сводится к нормированию и центрированию m входных биометрических параметров по формуле:

$$u_i = \frac{E(v_i) - v_i}{\sigma(v_i)}, \quad (1)$$

где i - упорядоченные номера входов нейрона; $i = 1, 2, \dots, m$, связанные с 416, контролируемыми биометрическими параметрами БиоОбраза, например, полученного в среде моделирования «БиоНейроАвтограф» [20], таблица связей формируется заранее с использованием генератора псевдослучайных чисел, как это рекомендует стандарт [11].

После нормирования и центрирования (1) для выборки m параметров по n примерам образа «Свой» вычисляют математическое ожидание - $\tilde{E}(u)$ и стандартное отклонение - $\tilde{\sigma}(u)$ для нормального теоретического распределения - $\tilde{p}(u)$.

Если после настройки нейрона dsg^2 подать на его входы тестовые примеры образа «Свой», не участвовавшие в его обучении, то на выходе линейной части получим отклики с малой дисперсией:

$$y = \int_{-\infty}^{+\infty} (\tilde{p}(u))^2 \cdot du = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left\{ \exp\left(\frac{-u^2}{2}\right) \right\}^2 du, \quad (2)$$

где математическое ожидание $E(u) \approx 0$ - практически не является случайной величиной, стандартное отклонение $\sigma(u) \approx 1$, так же практически не является случайной величиной.

Если же на входы обученного нейрона dsg^2 подавать биометрические данные образа «Чужой», то для них нормировка (1) работать не будет:

$$\tilde{u}_i = \frac{E(v_i) - \xi_i}{\sigma(v_i)}, \quad (2)$$

где ξ_i - биометрический параметр образа «Чужой».

Как следствие, математическое ожидание $E(\tilde{u})$ оказывается случайно величиной, а стандартное отклонение $\sigma(\tilde{u})$ принимает большие значения в интервале от 2 до 5. В итоге отклик - \tilde{y} нейрона dsg^2 на воздействие вектором биометрических параметров образа «Чужой» - $\tilde{\xi}$ будет описываться уравнением совершенно не похожим на уравнение (2):

$$\tilde{y} = \int_{-\infty}^{+\infty} p(u) \cdot \tilde{p}(u) \cdot du = \frac{1}{2\pi \cdot \sigma(\tilde{u})} \int_{-\infty}^{+\infty} \left\{ \exp\left(-\frac{u^2}{2}\right) \right\} \cdot \left\{ \exp\left(-\frac{(E(\tilde{u}) - u)^2}{2 \cdot (\sigma(\tilde{u}))^2}\right) \right\} du, \quad (3)$$

где математическое ожидание $E(\tilde{u})$ - случайная величина с нулевым математическим ожиданием $E(E(\tilde{u})) \approx 0.0$ и значительным стандартным отклонением $\sigma(E(\tilde{u})) \approx 1.41$, стандартное отклонение $\sigma(\tilde{u}) \approx 3.8$ самой переменной не случайно и имеет значительную величину.

Кардинальное отличие уравнений состоит в том, что они дают совершенно разные по своей природе отклики. Уравнение (2) является почти детерминированным, тогда как уравнение (3) дает случайную величину с большим стандартным отклонением $-\sigma(\tilde{y})$. Именно это обстоятельство и давало возможность добиваться высокого уровня подавления шумов квантования, возникающих на малых выборках при применении критерия среднего геометрического от двух сравниваемых плотностей распределения значений [6, 21]. Соотношения математических ожиданий распределения математических ожиданий образов «Свой», «Чужой» и их стандартных отклонений приведено на рис. 3 для нейронов dsg^2 с 8 входами (данные среды моделирования «БиоНейроАвтограф» [20]).

В силу того, что выражение (2) дает большое и почти детерминированное значение, а выражение (3) дает малое и случайное значение примеры образа «Свой» и примеры образов «Чужой» оказываются хорошо различимы, если использовать нейрон dsg^2 с 8 входами.

Результат разделения (одинаковая вероятность ошибок первого и второго рода) оказывается намного лучше, чем для линейного нейрона и обычного квадратичного нейрона;

- линейный нейрон с 8 входами $P_1=P_2=P_{EE}=0.45$;
- квадратичный нейрон, имеющий 8 входов, $P_{EE}=0.26$;
- нейрон dsg^2 с 8 входами $P_1=P_2=P_{EE}=0.21$.

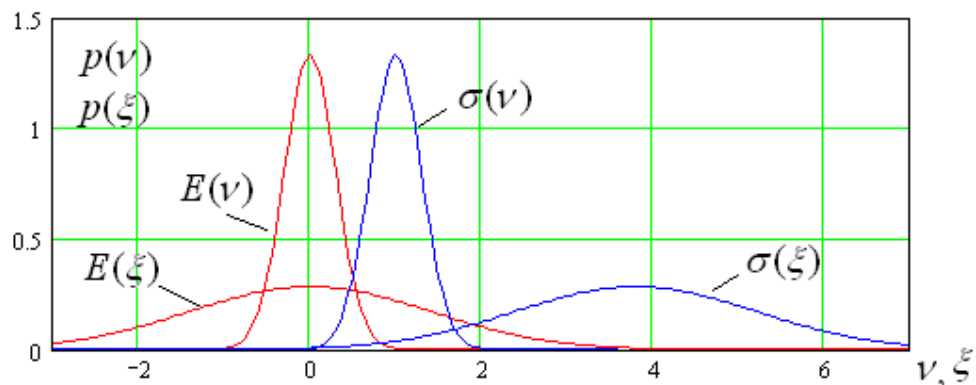


Рис. 3. Распределения математических ожиданий и стандартных отклонений для биометрических параметров образа «Свой» - v и образа «Чужой» - ξ для нейрона dsg^2 с 8 входами

С ростом размерности нейросетевого преобразования выигрыш от замены линейных и квадратичных нейронов на нейроны среднего геометрического квадрата плотностей распределения значений вероятностей усиливается.

Очевидно, что прямые вычисления вида (2), (3) реализовать на низкоразрядных процессорах с малой производительностью трудно. В связи с этим достаточно сложные функции преобразования (2) и (3) следует вычислить заранее и представить в виде двумерной таблицы:

$$\tilde{Y} = -\log_2(\tilde{y}(E(\tilde{u}), \sigma(\tilde{u}))). \quad (4)$$

Проблемы вычислений с использованием очень малых значений вероятностей решаются заранее во время вычисления двумерных таблиц. Эта задача хорошо решается при использовании 64 разрядной сетки вычислительной машины под управлением какой-либо из сред для математических вычислений (например, MathLAB, MathCAD и другие). Сам же нейрон квадрата среднего геометрического может быть реализован программно на любом 8-ми разрядном процессоре низкой производительности.

Заключение. В данной работе впервые сделана попытка показать, что каждому из известных статистических функционалов может быть поставлен в соответствие некоторый нейрон. Особый интерес этот подход представляет при реализации многомерного статистического анализа биометрических данных. При технической реализации нейронов квадрата среднего геометрического проблемы работы с малыми значениями вероятностей легко разрешимы применением заранее вычисленных двумерных таблиц логарифмов вероятностей. Это позволяет реализовывать нейроны квадрата среднего геометрического не только на обычных ПЭВМ, работающих под ОС семейства Windows, но и на любом 8-ми битном процессоре малой производительности.

Библиографические ссылки

1. Кобзарь А. И. Прикладная математическая статистика для инженеров и научных работников. М.: ФИЗМАТЛИТ, 2006. 816 с.
2. ГОСТ Р 50.1.037-2002. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Ч. 1. Критерии типа хи-квадрат. Госстандарт России. М., 2001. 140 с.
3. ГОСТ Р 50.1.037-2002 Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Ч. 2. Непараметрические критерии. Госстандарт России. М., 2002. 123 с.
4. Иванов А.И. Многомерная нейросетевая обработка биометрических данных с программным воспроизведением эффектов квантовой суперпозиции. Монография. Пенза: Издательство АО «ПНИЭИ», 2016. 133 с.
5. Эффект снижения размера тестовой выборки за счет перехода к многомерному статистическому анализу биометрических данных / В.И. Волчихин, А.И. Иванов, Н.И. Серикова и др. // Известия высших учебных заведений. Поволжский регион. Технические науки. Пенза: ПГУ, 2015. №1. С. 50 – 59.
6. Иванов А.И., Перфилов К.А. Оценка соотношения мощностей семейства статистических критериев «среднего геометрического» на малых выборках биометрических данных. XI Всероссийская научно-практическая конференция. «Современные охранные технологии и средства обеспечения комплексной безопасности объектов. Пенза-Заречный, 2016. С. 223-229.
7. Быстрые алгоритмы тестирования нейросетевых механизмов биометрико-криптографической защиты информации / А.Ю. Малыгин, В.И. Волчихин, А.И. Иванов и др. // Пенза: ПГУ, 2006. 161 с.
8. Оценка правдоподобия гипотезы о нормальном распределении по критерию Джини для числа степеней свободы, кратного числу опытов / Н.И. Серикова, А.И. Иванов, Ю.И. Серикова // М.: Вопросы радиоэлектроники. 2015. № 1 (1). С.85–94.
9. Серикова Н.И. Оценка правдоподобия гипотезы о нормальном распределении по критерию Джини для сглаженных гистограмм, построенных на малых тестовых выборках. Вопросы радиоэлектроники. М.: ЦНИИ «Электроника», 2015. № 1. С.85 – 94.
10. ГОСТ Р 52633.0-2006. Защита информации. Техника защиты информации. Требования к средствам высоконадежной биометрической аутентификации. М.: Стандартинформ, 2007. 27 с.

11. ГОСТ Р 52633.5-2011. Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия-код доступа. М.: Стандартинформ, 2012. 16 с.
12. ГОСТ Р 52633.3-2011. Защита информации. Техника защиты информации. Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора. М.: Стандартинформ, 2012. 16 с.
13. Саймон Х. Нейронные сети: полный курс. М.: «Вильямс», 2006. 1104 с.
14. Многомерный статистический анализ биометрических данных сетью частных критериев Пирсона / Б.Б. Ахметов, Иванов А.И., Безяев А.В. и др. // Алматы: Вестник Национальной академии наук Республики Казахстан, 2015. № 1. С. 5-11.
15. Подавление шумов квантования биометрических данных при использовании многомерного критерия Крамера-фон Мизеса / А.И. Иванов, Газин А.И., Вятчанин С.Е. и др. // Проблемы информационной безопасности. Компьютерные системы. Санкт Петербург: ПТУ 2016. № 2. С. 21-28.
16. Ахметов Б., Иванов А. Многомерные статистики существенно зависимых биометрических данных, порождаемые нейросетевыми эмуляторами квадратичных форм. Монография. Казахстан – Алматы: LEM, 2016. 86 с.
17. Снижение требований к размеру тестовой выборки биометрических данных при переходе к использованию многомерных корреляционных функционалов Байеса / А.И. Иванов, П.С. Ложников, А.Е. Сулавко и др. // Инфокоммуникационные технологии. 2017. № 15 (2). С. 186-193.
18. Идентификация подлинности рукописных автографов сетями Байеса-Хэмминга и сетями квадратичных форм / А.И. Иванов, П.С. Ложников, Е.И. Качайкин // «Вопросы защиты информации». 2015 г. № 2. С.28-34.
19. Биометрическая идентификация рукописных образов с использованием корреляционного аналога правила Байеса / А.И. Иванов, П.С. Ложников, Е.И. Качайкин и др. // Вопросы защиты информации. 2015. № 3. С. 48-54.
20. Иванов А.И., Захаров О.С. Среда моделирования «БиоНейроАвтограф». Программный продукт создан лабораторией биометрических и нейросетевых технологий [Электронный ресурс]. URL: <http://пниэи.рф/activity/science/noc.htm> (дата обращения 10.08.17).
21. Оценка качества малых выборок биометрических данных с использованием дифференциального варианта статистического критерия среднего геометрического / А.И. Иванов, К.А. Перфилов, Е.А. Малыгина // Вестник СИБГАУ. 2016. №4 (17). С.864-871.

Данные об авторах:

Перфилов Константин Александрович - аспирант кафедры «Информационная безопасность систем и технологий» ФГБОУ ВО «Пензенский государственный университет». Российская Федерация, 440026, г. Пенза, ул. Красная, 40. E-mail: perfilov58@gmail.com

Газин Алексей Иванович - к.т.н., доцент кафедры Информатики, информационных технологий и защиты информации ФБГОУ ВПО «Липецкий государственный педагогический университет».

Gazin Alexei Ivanovich – PhD in Technical Sciences, assistant professor in professorial chair “Informatics, information technology and information security” in “Lipetsk State Pedagogical University”