УДК 519.24; 53; 57.017
DOI 10.21685/2072-3059-2017-4-2

*В. И. Волчихин, А. И. Иванов, А. Г. Банных*

# РЕГУЛЯРИЗАЦИЯ ВЫЧИСЛЕНИЯ ЭНТРОПИИ ВЫХОДНЫХ СОСТОЯНИЙ НЕЙРОСЕТЕВОГО ПРЕОБРАЗОВАТЕЛЯ БИОМЕТРИЯ-КОД, ПОСТРОЕННАЯ НА РАЗМНОЖЕНИИ МАЛОЙ ВЫБОРКИ ИСХОДНЫХ ДАННЫХ

**Аннотация**.

*Актуальность и цели.* Целью работы является регуляризация вычисления энтропии длинных кодов с зависимыми разрядами на малой тестовой выборке.

*Материалы и методы.* Алгоритм построен на предсказании вероятности появления редких событий, использующий гипотезу нормального распределения расстояний Хэмминга. Предложено отказаться от вычисления математического ожидания и стандартного отклонения на малых выборках. Предложено размножать исходные данные через добавление к ним мутаций, полученных от генератора псевдослучайного шума. Даны ограничения на амплитуду генератора шума мутаций.

*Результаты.* Показано, что при переходе в пространство расстояний Хэмминга наблюдается логарифмическое сжатие размеров исходного алфавита спектра выходных состояний нейросетевой молекулы преобразователя биометрия-код. Это в итоге позволяет выполнять быстрое тестирование преобразователей через вычисление их энтропии.

*Выводы.* Предложенная процедура регуляризации вычислений позволяет получить точность оценки значения энтропии нейросетевого преобразователя на выборке в 21 опыт такую же, как точность вычислений, обеспечиваемую стандартными процедурами вычислений по ГОСТ Р 52633.3 на выборке из 2100 опытов. Наблюдается примерно 100-кратный рост устойчивости вычислений.

**Ключевые слова**: статистический анализ малых выборок, предсказание вероятности появления редких событий, искусственные нейронные сети, энтропия.

*V. I. Volchikhin, A. I. Ivanov, A. G. Bannykh*

# REGULARIZING CALCULATIONS OF THE OUTPUT ENTROPY OF A NEURAL NETWORK “BIOMETRICS-CODE” CONVERTER THOUGH MULTIPLICATION OF A SMALL SAMPLE OF ORIGINAL DATA

**Abstract**.

*Background*. The aim of the paper is to regularize calculations of the entropy of long codes with dependent bits on a small testing sample.

*Materials and methods.* The algorithm is based on predicting the probability of occurrence of rare events using the hypothesis of normal distribution of Hamming distances. It is suggested to abandon calculations of mathematical expectation and standard deviation in small samples, and proposed to multiply the original data by adding mutations obtained from a pseudorandom noise generator to them. Limitations to the mutation noise generator’s amplitude are given.

*Results.* It is shown that the transition to the Hamming distances' space leads to a logarithmic compression of the initial alphabet's size of the output states spectrum of a "biometrics-code" converter's molecule. This ultimately allows rapid testing of converters through calculation of their entropy.

*Conclusions.* The procedure of regularization of calculations proposed in the article makes it possible to obtain the same accuracy of evaluating the entropy of a neural network converter on a sample of 21 experiments as the accuracy of calculations provided by standard computational procedures in accordance with GOST R 52633.3 on a sample of 2100 experiments. There is a 100-fold increase in the stability of computations.

**Key words**: statistical analysis of small samples, prediction of the probability of rare events occurrence, artificial neural networks, entropy.

## 1. Стандартный метод вычисления энтропии длинных кодов с зависимыми разрядами

Обычно энтропию объекта исследования оценивают через наблюдение вероятности появления возможных состояний заранее заданного алфавита по Шеннону [1]. Когда число состояний мало, проблем с оценкой энтропии не возникает. Этот классический подход вполне применим к оценке энтропии состояний преобразователей биометрия-код. Например, такие преобразователи за рубежом строят с использованием так называемых «нечетких экстракторов» [2–5]. Убедиться в этом можно, воспользовавшись достоверными биометрическими данными, самостоятельно получаемыми в среде моделирования «БиоНейроАвтограф» [6]. Этот продукт позволяет получать 416-мерные вектора биометрических параметров и соответствующие им 256-битные коды.

Проблема оценки энтропии «нечетких экстракторов» технически разрешима из-за того, что «нечеткие экстракторы» имеют короткий выходной код. Так, если в «нечетком экстракторе» при обработке 416 параметров применен код с 35-кратной избыточностью, то длина его информационной части составит не более 12 бит. Для 12-битных кодов проблем с оценкой их энтропии по Шеннону не возникает.

Однако, если мы воспользуемся российско-казахстанским опытом и будем применять нейросетевые преобразователи биометрия-код [7, 8], то длина выходного кода составит 256 бит и применить алгоритм Шеннона для вычисления энтропии уже нельзя.

Выход из этого положения состоит в переходе из пространства вероятности появления обычных кодов в пространство появления расстояний Хэмминга [9, 10]. На рис. 1 дана схема вычислений, рекомендуемая ГОСТ Р 52633.3.

Нейронная сеть среды моделирования «БиоНейроАвтограф» [6] преобразует 416 биометрических параметров в код длиной 256 бит. То есть его выходные коды могут давать $2^{256}$ состояний, что не позволяет оценить вероятности их появления. Однако, если перейти в пространство расстояний Хэмминга между кодами образов «Чужой-1», «Чужой-2»… «Чужой-21» и кодом образа «Свой», то мы получим спектр, имеющий всего 256 возможных состояний. Наблюдается логарифмическое свертывание числа возможных спектральных состояний $\log_2(2^{256}) = 256$ и, следовательно, экспоненциальное упрощение задачи оценки энтропии.

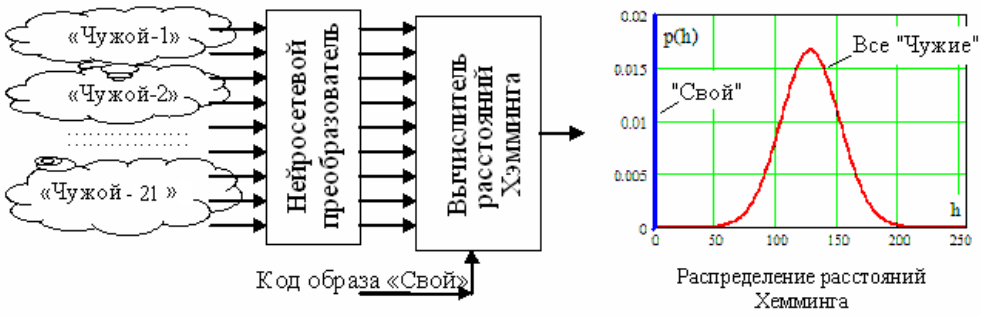


Рис. 1. Тестирование преобразователя биометрия-код
при переходе в пространство расстояний Хэмминга

Все криптографические операции создаются таким образом, чтобы давать данные, близкие к «белому шуму». То есть разряды кодов независимые, а энтропия кодов точно совпадает с длиной кода. Для биометрических данных все иначе, их разряды зависимы (коррелированны). Теоретически 256 зависимых бит можно заменить на меньшее число независимых бит, например путем применения процедуры декорреляции.

Стандартная процедура вычислений по ГОСТ Р 52633.3 построена на том, что, подав на нейросеть 21 образ «Чужой», мы получаем 21 код «$\overline{x}$». Образ «Свой» дает код «$\overline{c}$», соответственно мы можем вычислить расстояние Хэмминга сложением по модулю два их разрядов:

$$h = 256 - \sum_{i=0}^{255} ("c_i") \oplus ("x_i") . \qquad (1)$$

Далее по выборке в 21 опыт мы можем вычислить математическое ожидание $E(h)$ и стандартное отклонение $\sigma(h)$. Знание о двух статистических моментах позволяет оценить вероятность угадывания кода «Свой» $P_2$, подставляя случайные образы «Чужой»:

$$P_2 \approx \frac{1}{\sigma(h)\sqrt{2\pi}} \int_0^1 \exp\left\{ \frac{-\left(E(h)-u\right)^2}{2\left(\sigma(h)\right)^2} \right\} \cdot du . \qquad (2)$$

В этом случае энтропия нейросетевого преобразователя оценивается следующим образом:

$$H("x_1, x_2, ..., x_{256}") \approx -\log_2(P_2) . \qquad (3)$$

### 2. Проблема плохой обусловленности, возникающая при прогнозировании энтропии по данным малой выборки

Каждая достаточно сложная вычислительная процедура, как правило, накапливает ошибки исходных данных. Для оценки показателя обусловленности последовательности процедур (2), (3) необходимо использовать большую базу образов в 300 выборок по 21 образу «Чужой». В результате тестирования на таком объеме данных математическое ожидание составляет

$E(h)$ = 107,26 бит, а его стандартное отклонение – $\sigma\{E(h)\}$ = 6,77 бит. Это означает, что на выборке в 21 опыт относительная ошибка вычисления математического ожидания может составлять до 18,9 % при нормальном законе распределения значений.

Математическое ожидание стандартного отклонения составляет $E\{\sigma(h)\} = 27{,}84$ бит, а его стандартное отклонение $\sigma\{\sigma(h)\} = 4{,}04$ бит. Это означает, что стандартное отклонение на выборке в 21 опыт может оцениваться с относительной ошибкой до 43,5 % при нормальном законе распределения значений. Все это приводит к итоговым относительным ошибкам вычисления энтропии до 271 %. Мы имеем дело с плохо обусловленной задачей.

На рис. 2 приведен пример гистограммы распределения значений энтропии, вычисленной по 300 выборкам, каждая из которых состоит из 21 опыта.

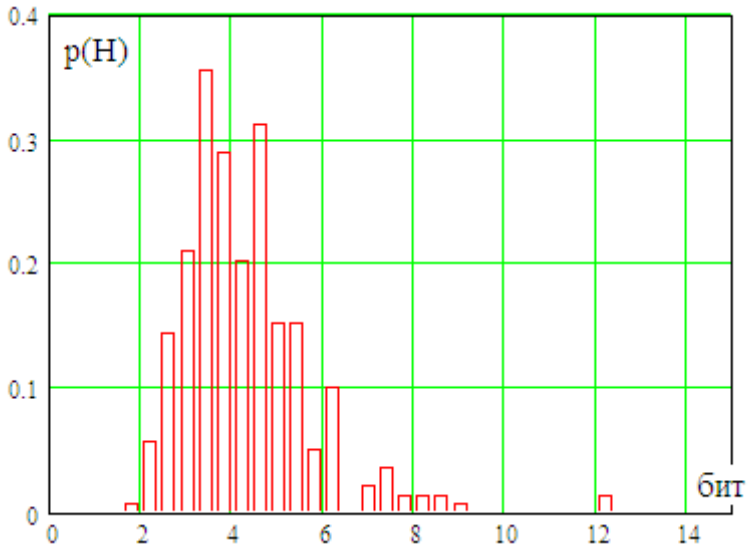


Рис. 2. Гистограмма распределения значений энтропии, вычисленной на выборке в 21 опыт

Из рис. 2 видно, что при математическом ожидании энтропии $E(H) = 4{,}52$ бита (данные численного эксперимента) минимальное значение энтропии составляет 1,8 бита, а максимальное значение энтропии составляет 12,2 бита. Максимальная оценка оказывается выше математического ожидания на 271 %, что еще раз подтверждает плохую обусловленность решаемой задачи. Относительная ошибка вычислений нелинейно усиливается и оказывается намного больше, чем сумма средних значений максимумов относительных ошибок исходных данных:

$$cond(H) \approx \frac{2\max(\left|E(H) - H_i\right|)}{(\max(\left|E(h) - h_i\right|) + \max(\left|E(\sigma(h)) - \sigma(h_i)\right|)} \approx 8{,}96. \qquad (4)$$

Мы наблюдаем почти 9-кратное усиление ошибок исходных данных при стандартных вычислениях по ГОСТ Р 52633.3 [10] на малой выборке в 21 опыт. Для снижения столь существенного эффекта накопления ошибок на

порядок потребуется увеличить объем тестовой выборки на два порядка – до 2100 опытов.

### 3. Регуляризация вычисления значения энтропии на малой тестовой выборке путем размножения исходных данных

При переходе от обычных кодов к кодам расстояний Хэмминга возникает эффект логарифмического сжатия числа спектральных линий и эффект нормализации положения их распределения. На рис. 3 приведены положения 19 обнаруженных спектральных линий расстояний Хэмминга, имеющих нормальную плотность интенсивности (вероятности появления).

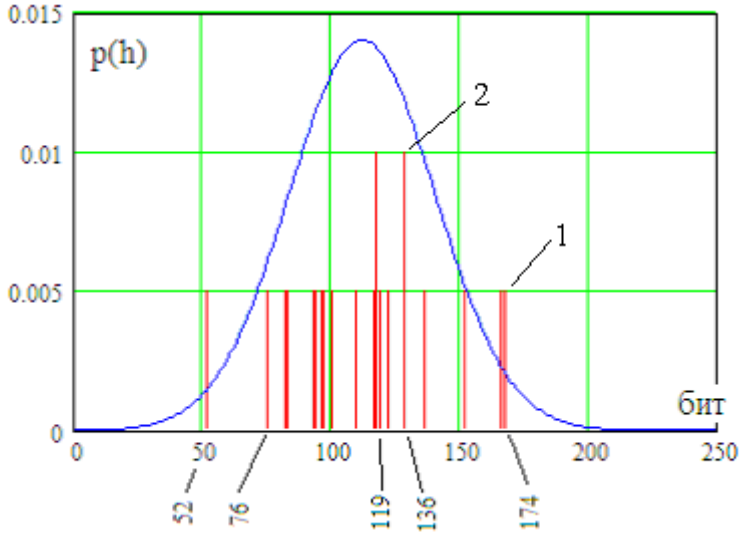


Рис. 3. Пример реального распределения значений положения спектральных линий расстояний Хэмминга (21 опыт) на фоне непрерывного нормального распределения их яркости (интенсивности)

При большом объеме выборки в 2100 опытов интенсивность (яркость) спектральных линий Хэмминга будет точно описываться нормальным законом распределения значений. Яркость спектральных линий на рис. 3 иная только в силу малого объема тестовой выборки. В численном эксперименте всего наблюдаются 19 линий спектра, две из которых в 2 раза интенсивнее остальных 17 линий. Минимальное положение линии спектра 52 бита, максимальное значение линий спектра 174 бита. По сути дела мы наблюдаем сильно прореженную гистограмму положения спектральных линий расстояний Хэмминга. У этой прореженной гистограммы 19 столбцов заполнены и 174 – 50 – 19 = 105 столбцов пустые. Существует ряд технических приемов регуляризации вычислений над данными малых выборок с прореженными гистограммами [11–13].

Одним из самых эффективных приемов регуляризации вычислений является размножение исходных данных малой выборки процедурами ГОСТ Р 52633.2–2010 [14]. На рис. 4 дана блок-схема проведения численного эксперимента, построенная на размножении малой выборки исходных данных мутациями, полученными от программного генератора псевдослучайных чисел.

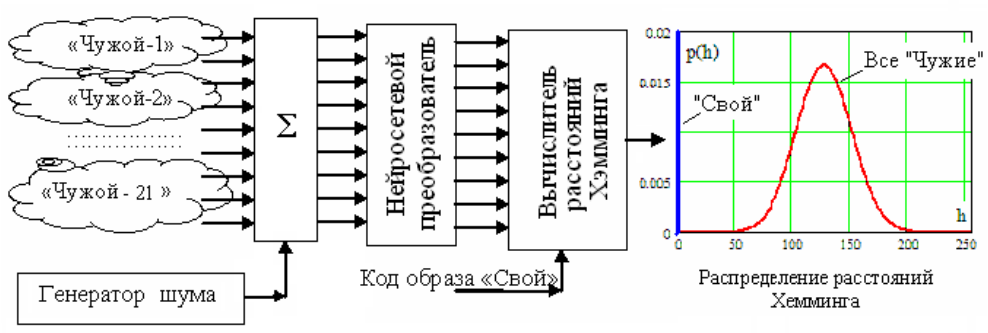


Рис. 4. Схема численного эксперимента, построенная с учетом размножения биометрических образов добавлением мутаций

В том случае, когда от каждого из 21 образа «Чужой» будет получено 99 близких синтетических образов, итоговая гистограмма расстояний Хэмминга практически уже не будет иметь пустых столбцов, как это показано на рис. 5.

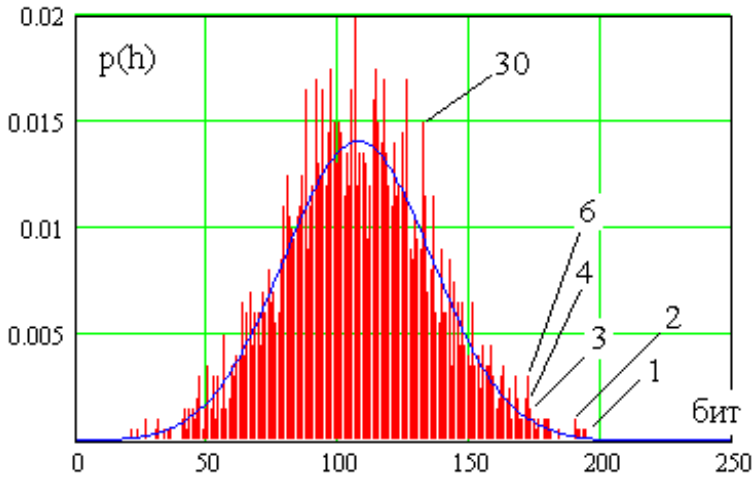


Рис. 5. Гистограмма распределения расстояний Хэмминга, полученная на выборке в 2100 синтетических образов, полученных из 21 образа

Из рис. 5 видно, что на краях нормального распределения могут появляться столбцы гистограммы с нулевым заполнением, однако их намного меньше, чем столбцов с заполнением в 1, 2, 3, 4,… опытов. В центре распределения столбцы с нулевым заполнением полностью отсутствуют, их заполнение находится в интервале от 24 до 40 опытов.

При создании близких копий естественных биометрических образов может возникнуть так называемый «кошмар Дженкинса» [15], когда разрушаются внутренние корреляционные связи в исходных данных в следующих поколениях. Кроме того, слишком большой объем мутаций также приводит к вырождению данных в следующих поколениях. При проводимых нами численных экспериментах стандартное отклонение генератора шума мутаций выбиралось исходя из следующих ограничений:

$$\sigma_{ш} = \frac{1}{\sqrt{2}} E(\sigma(\nu_i)), \quad (5)$$

где $E(\sigma(\nu_i))$ – математическое ожидание значений стандартных отклонений параметров части или полной выборки из 21 исходного биометрического образа.

Результаты численного эксперимента приведены на рис. 6.

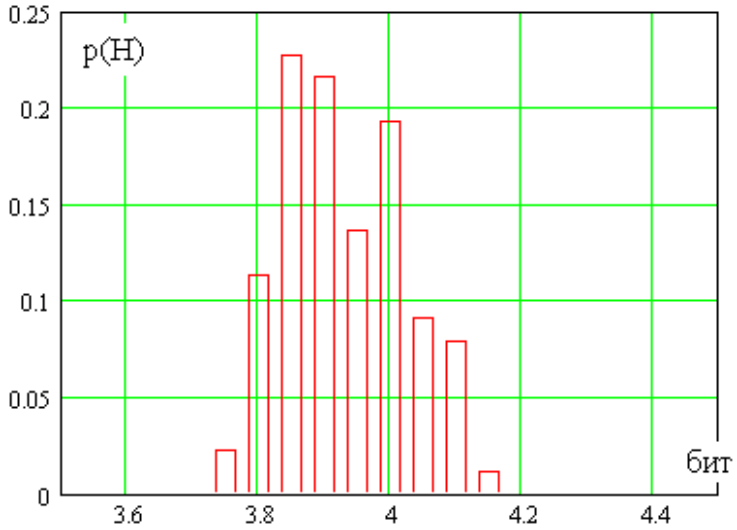


Рис. 6. Гистограмма значений энтропии по 100 выборкам из 2100 опытов

Для того чтобы убедиться в эффективности регуляризации, следует сравнить данные рис. 2 и рис. 4. На рис. 2 оценки энтропии лежат в интервале от 1,8 до 12,2 бита. Те же самые оценки для тех же данных после регуляризации вычисления оказываются в интервале от 3,68 до 4,16. Мы имеем 16-кратное снижение неопределенности вычислений, что эквивалентно снижению примерно в 16 раз числа обусловленности вычислительных процедур. При этом наблюдается некоторое смещение математического ожидания, что может свидетельствовать о появлении незначительной методической погрешности.

**Заключение**

Переход от обычных кодов в пространство расстояний Хэмминга позволяет экспоненциально свертывать число спектральных линий выходных состояний нейросетевого преобразователя биометрия-код. Это обеспечило возможность в 2006 г. начать работы по созданию стандарта ГОСТ Р 52633.3–2011 [9]. На данный момент уже созданный и введенный в действие стандарт нуждается в корректировке своих основных положений. В данной статье мы попытались показать, что заложенные в действующий стандарт вычислительные процедуры имеют плохую обусловленность. Их обусловленность может быть значительно улучшена в новой версии стандарта, что позволит на малой выборке в 21 опыт получать ошибку вычислений такую же, как на большой выборке в 2100 опытов для действующего стандарта.

***Библиографический список***


1. **Яглом, А. М.** Вероятность и информация / А. М. Яглом, И. М. Яглом. – М. : Дом Книги, 2007. – 512 с.
2. **Juels, A.** A Fuzzy Commitment Scheme / A. Juels, M. Wattenberg // Proc. ACM Conf. Computer and Communications Security (1–4 November, 1999). – Singapore, 1999. – P. 28–36.
3. **Monrose, F.** Cryptographic key generation from voice / F. Monrose, M. Reiter, Q. Li, S. Wetzel // In Proc. IEEE Symp. on Security and Privacy (14–16 May, 2001). – Okland, California USA, 2001. – P. 202–213.
4. **Dodis, Y.** Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy / Y. Dodis, L. Reyzin, A. Smith // EUROCRYPT. – 2004. – Data April 13. – P. 523–540.
5. **Hao, F.** Crypto with Biometrics Effectively / Feng Hao, Ross Anderson, and John Daugman // IEEE Transactions on computers. – 2006. – Vol. 55, № 9.
6. **Иванов, А. И.** Среда моделирования «БиоНейроАвтограф». Программный продукт создан лабораторией биометрических и нейросетевых технологий, размещен с 2009 г. на сайте АО «ПНИЭИ» / А. И. Иванов, О. С. Захаров. – URL: http://пниэи.рф/activity/science/noc.htm (для свободного использования университетами России, Белоруссии, Казахстана).
7. **Волчихин, В. И.** Быстрые алгоритмы обучения нейросетевых механизмов биометрико-криптографической защиты информации : монография / В. И. Волчихин, А. И. Иванов, В. А. Фунтиков. – Пенза : Изд-во ПГУ, 2005. – 273 с.
8. Технология использования больших нейронных сетей для преобразования нечетких биометрических данных в код ключа доступа : монография / Б. С. Ахметов, А. И. Иванов, В. А. Фунтиков, А. В. Безяев, Е. А. Малыгина. – Алматы, Казахстан : LEM, 2014. – 144 с. – URL: http://portal.kaznu.kz/files/publicate/2014-06-27-11940.pdf
9. **Малыгин, А. Ю.** Быстрые алгоритмы тестирования нейросетевых механизмов биометрико-криптографической защиты информации / А. Ю. Малыгин, В. И. Волчихин, А. И. Иванов, В. А. Фунтиков. – Пенза : Изд-во ПГУ, 2006. – 161 с.
10. ГОСТ Р 52633.3–2011. Защита информации. Техника защиты информации. Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора. – М., 2011.
11. **Серикова, Н. И.** Оценка правдоподобия гипотезы о нормальном распределении по критерию Джини для сглаженных гистограмм, построенных на малых тестовых выборках / Н. И. Серикова, А. И. Иванов, Ю. И. Серикова // Вопросы радиоэлектроники. Сер.: СОИУ. – 2015. – Вып. 1. – С. 85–94.
12. **Иванов, А. И.** Сравнение мощности хи-квадрат критерия и критерия Крамера-фон Мезиса для малых тестовых выборок биометрических данных / А. И. Иванов, А. И. Газин, С. Е. Вятчанин, К. А. Перфилов // Надежность и качество сложных систем. – 2016. – № 2 (14). – С. 67–72.
13. **Иванов, А. И.** Многомерный статистический анализ качества биометрических данных на предельно малых выборках с использованием критериев среднего геометрического, вычисленного для анализируемых функций вероятности / А. И. Иванов, К. А. Перфилов, Е. А. Малыгина // Измерение. Мониторинг. Управление. Контроль. – 2016. – № 2 (16). – С. 58–66.
14. ГОСТ Р 52633.2–2010. Защита информации. Техника защиты информации. Требования к формированию синтетических биометрических образов, предназначенных для тестирования средств высоконадежной биометрической аутентификации. – М., 2010.
15. Solving the inverse task of neural network biometrics without mutations and Jenkins "nightmare" in the implementation of genetic algorithms / B. Akhmetov, S. Kachalin,

A. Ivanov, A. Bezyaev, K. Mukapil // Computational and Informational Technologies in Science, Engineering and Education (CITech-2015) : International Conference. – Almaty, Kazakhstan, 2015 – P. 89–92.

### *References*

1. Yaglom A. M., Yaglom I. M. *Veroyatnost' i informatsiya* [Probability and information]. Moscow: Dom Knigi, 2007, 512 p.
2. Juels A. A, Wattenberg M. *Proc. ACM Conf. Computer and Communications Security (1–4 November, 1999)*. Singapore, 1999, pp. 28–36.
3. Monrose F., Reiter M., Li Q., Wetzel S. *In Proc. IEEE Symp. on Security and Privacy (14–16 May, 2001)*. Okland, California USA, 2001, pp. 202–213.
4. Dodis Y., Reyzin L., Smith A. *EUROCRYPT*. 2004, April 13, pp. 523–540.
5. Hao F., Anderson R, Daugman J. *IEEE Transactions on computers*. 2006, vol. 55, no. 9.
6. Ivanov A. I., Zakharov O. S. *Sreda modelirovaniya «BioNeyroAvtograf». Programmnyy produkt sozdan laboratoriey biometricheskikh i neyrosetevykh tekhnologiy, razmeshchen s 2009 g. na sayte AO «PNIEI»* ["BioNeyroAvtograph" simulation environment. The product was established by the laboratory of biometric and neural network technologies and announced in 2009 on the PNIEI website]. Available at: http://pniei.rf/activity/science/noc.htm (dlya svobodnogo ispol'zovaniya universi-tetami Rossii, Belorussii, Kazakhstana).
7. Volchikhin V. I., Ivanov A. I., Funtikov V. A. *Bystrye algoritmy obucheniya neyrosetevykh mekhanizmov biometriko-kriptograficheskoy zashchity informatsii: monografiya* [Fast learning algorithms for neural network mechanisms of biometric cryptographic data protection: monograph]. Penza: Izd-vo PGU, 2005, 273 p.
8. 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: monografiya* [The technology of large neural networks' implementation for fuzzy biometric data transformation into access code: monograph]. Almaty, Kazakhstan: LEM, 2014, 144 p. Available at: http://portal.kazntu.kz/files/publicate/2014-06-27-11940.pdf
9. Malygin A. Yu., Volchikhin V. I., Ivanov A. I., Funtikov V. A. *Bystrye algoritmy testirovaniya neyrosetevykh mekhanizmov biometriko-kriptograficheskoy zashchity informatsii* [Fast testing algorithms for newural network mechanisms of biometric cryptographic data protection]. Penza: Izd-vo PGU, 2006, 161 p.
10. *GOST R 52633.3–2011. Zashchita informatsii. Tekhnika zashchity informatsii. Testirovanie stoykosti sredstv vysokonadezhnoy biometricheskoy zashchity k atakam podbora* [Data protection. Data protecting technology. Testing of highly reliable biometric protective means' resistance to breaking attacks]. Moscow, 2011.
11. Serikova N. I., Ivanov A. I., Serikova Yu. I. *Voprosy radioelektroniki. Ser.: SOIU* [Problems of radioelectronics. Series: SOIU]. 2015, iss. 1, pp. 85–94.
12. Ivanov A. I., Gazin A. I., Vyatchanin S. E., Perfilov K. A. *Nadezhnost' i kachestvo slozhnykh system* [Reliability and quality of complex systems]. 2016, no. 2 (14), pp. 67–72.
13. Ivanov A. I., Perfilov K. A., Malygina E. A. *Izmerenie. Monitoring. Upravlenie. Kontrol'* [Measurement. Monitoring. Management. Control]. 2016, no. 2 (16), pp. 58–66.
14. *GOST R 52633.2–2010. Zashchita informatsii. Tekhnika zashchity informatsii. Trebovaniya k formirovaniyu sinteticheskikh biometricheskikh obrazov, prednaznachennykh dlya testirovaniya sredstv vysokonadezhnoy biometricheskoy autentifikatsii* [Data protection. Data protecting technology. Requirements to creation of synthetic biometric images intended for highly reliable biometric authentication means testing]. Moscow, 2010.

15. Akhmetov B., Kachalin S., Ivanov A., Bezyaev A., Mukapil K. *Computational and Informational Technologies in Science, Engineering and Education (CITech-2015): International Conference*. Almaty, Kazakhstan, 2015, pp. 89–92.

***Волчихин Владимир Иванович***
доктор технических наук, профессор, президент Пензенского государственного университета (Россия, г. Пенза, ул. Красная, 40)

E-mail: president@pnzgu.ru

***Volchikhin Vladimir Ivanovich***
Doctor of engineering sciences, professor, President of Penza State University (40 Krasnaya street, Penza, Russia)

***Иванов Александр Иванович***
доктор технических наук, доцент, начальник лаборатории биометрических и нейросетевых технологий, Пензенский научно-исследовательский электротехнический институт (Россия, г. Пенза, ул. Советская, 9)

E-mail: ivan@pniei.penza.ru

***Ivanov Aleksandr Ivanovich***
Doctor of engineering sciences, associate professor, head of the laboratory of biometric and neural network technologies, Penza Research Institute of Electrical Engineering (9 Sovetskaya street, Penza, Russia)

***Банных Андрей Григорьевич***
аспирант, Пензенский государственный университет (Россия, г. Пенза, ул. Красная, 40)

E-mail: ibst@pgzgu.ru

***Bannykh Andrey Grigor'evich***
Postgraduate student, Penza State University (40 Krasnaya street, Penza, Russia)

УДК 519.24; 53; 57.017

**Волчихин, В. И.**

**Регуляризация вычисления энтропии выходных состояний нейросетевого преобразователя биометрия-код, построенная на размножении малой выборки исходных данных** / В. И. Волчихин, А. И. Иванов, А. Г. Банных // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2017. – № 4 (44). – С. 14–23. DOI 10.21685/2072-3059-2017-4-2