DOI: 10.18372/2410-7840.19.12219
УДК 681.322:681.5.015.42 (045)

ПРОБЛЕМЫ РАЗМЕРНОСТИ ЗАДАЧ РАСПОЗНАВАНИЯ ОБРАЗОВ В СИСТЕМАХ БИОМЕТРИЧЕСКОЙ АУТЕНТИФИКАЦИИ

Жулдыз Алимсеитова, Нургуль Сейлова, Сергей Гнатюк

В системах биометрической аутентификации осуществляется процесс доказательства и проверки подлинности заявленного пользователем имени через предъявление пользователем своего биометрического образа и путём преобразования этого образа в соответствии с заранее определенным протоколом аутентификации. Важным вопросом остается преобразование биометрических данных в код. В статье рассматриваются две наиболее известные технологии преобразования биометрии в код, приводится схема преобразования биометрических параметров в код ключа. Показано, что одной из основных причин трудности биометрической аутентификации является высокая размерность задачи. Для решения этой проблемы используются искусственные нейронные сети или «нечеткие экстракторы». Из множества существующих алгоритмов обучения нейронных сетей выбран алгоритм автоматического обучения большой искусственной нейронной сети. Показано применение энтропийного аппарата для снижения размерности задачи преобразования биометрия-код. Для снижения объемов вычислений произведен переход в расстояния Хемминга.

***Ключевые слова**: биометрический образ, аутентификация, искусственная нейронная сеть, нейросетевой преобразователь, размерность задачи преобразования, корреляция, пространство расстояний Хэмминга.*

В настоящее время существует несколько технологий преобразования биометрии в код. США и страны Евросоюза идут по пути связывания биометрических данных человека с его криптографическим ключом через применение так называемых «нечетких экстракторов» [1]. При этом «нечеткие экстракторы» рассматриваются как алгоритмы, выделяющие случайные, равномерно распределенные последовательности битов из биометрических данных в условиях зашумленности и способные компенсировать ошибки, возникающие из-за невозможности абсолютно точного повторного воспроизведения биометрических данных. В России и Казахстане преобразователи биометрия-код строятся с использованием искусственных нейронных сетей большого и сверхбольшого размера, которые обеспечивают частичную конфиденциальность и анонимность биометрического образа [2].

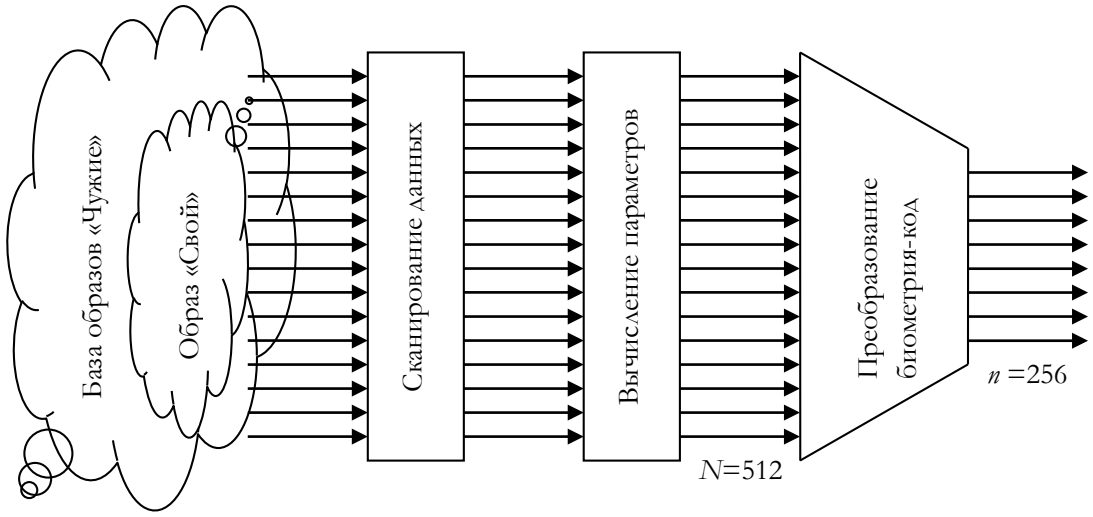

Рис. 1. Преобразование биометрических параметров в код ключа

Независимо от того, какая технология используется, преобразователь биометрия-код всегда имеет больше входных биометрических данных, чем число его выходов (чем длина выходного криптографического ключа). Так, на рис. 1 преобразователь имеет $N = 512$ входных биометрических параметров и только $n = 256$ выходных разрядов криптографического ключа (в соответствии с требованиями российских стандартов на шифрование и формирование электронной цифровой подписи длина криптографического ключа составляет 256 бит). Условие N >n всегда выполняется из-за низкой информативности биометрических данных. Одного биометрического параметра, как правило,

не хватает для получения одного бита криптографического ключа. Обязательно необходимо использовать избыточное число биометрических параметров для того, чтобы исправлять ошибки кодирования в «нечетких экстракторах» или «обогащать» входные биометрические данные нейронной сетью преобразователя [3].

Представление многомерных континуумов биометрических образов малым дискретным числом примеров

Одной из основных причин трудности биометрической аутентификации является высокая размерность задачи (приходится учитывать 512 и более «плохих» биометрических параметров). Именно по этой причине не удается воспользоваться аппаратом классической линейной алгебры и многомерной статистики. Положение усложняется тем, что преобразователи биометрия-код приходится обучать (настраивать) на малом числе примеров в обучающей выборке.

Так, при обучении биометрических средств аутентификации пользователи готовы предъявить от 10 до 20 примеров биометрического образа «Свой». Однако если их попросить предъявить 100 или 200 примеров, то эта работа воспринимается пользователями негативно. На сегодня пользователи не готовы прилагать значительные усилия для обучения своего биометрического программного робота (преобразователя биометрия-код). Последнее означает, что 512-мерные распределения континуумов параметров образа «Свой» мы вынуждены представлять всего 20 примерами по каждому из параметров. Мы можем попытаться представить 512-мерный объем естественно-континуальной формы представления данного образа «Свой» через внутренний объем графа 20 примерами «Свой», однако это будет весьма и весьма слабым приближением. Эта ситуация отображена на рис. 2.

Пользуясь всего 20 примерами, невозможно точно вычислить математическое ожидание биометрических параметров $E(v_i)$, их среднеквадратическое отклонение $\sigma(v_i)$ и коэффициенты корреляции между параметрами $r_{i,j}$. На обучающей выборке из 20 примеров относительная ошибка в оценке математического ожидания может быть от 0 до ±25 %, относительная ошибка среднеквадратического отклонения может составлять от 0 до ±50 %, относительная ошибка оценки коэффициентов корреляции может быть от 0 до ±100 %. При столь неточных данных невозможно построить многомерную аналитическую модель распределения данных образа «Свой».

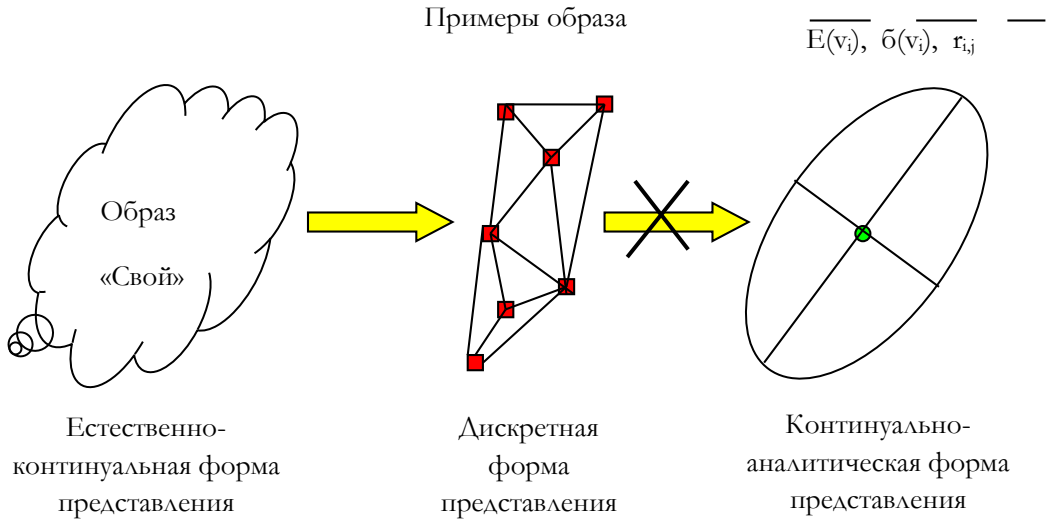

Рис. 2. Представление многомерного образа «Свой» малым числом примеров (7 примеров)

Даже, если придется работать в рамках гипотезы нормального нормированного распределения значений биометрических параметров, необходимо построить для них 512-мерную корреляционную матрицу. Придется обращать 512-мерную матрицу для того, чтобы вычислять поверхность гиперэллипса, разделяющую образ «Свой» от образов «Чужие». Она описывается квадратичными формами линейной алгебры, куда входит обратная ковариационная (корреляционная) матрица высокой размерности, и поэтому построить аналитическое описание разделяющей поверхности в форме гиперэллипсов технически невозможно (правая стрелка рис. 2 зачеркнута).

Описанная выше ситуация получила название «проклятия размерности». Казалось бы, что учет дополнительных биометрических параметров должен приводить к более точным результатам,

однако все происходит с точностью до наоборот. При попытках увеличивать размерность решаемой задачи она решается все хуже и хуже. Возникает эффект плохой обусловленности или накопления погрешностей вычислений. Именно из-за «проклятия размерности» в биометрии не удается пользоваться линейной алгеброй и классической многомерной статистикой. Однако применение искусственных нейронных сетей и «нечетких экстракторов» позволяет ослабить проблему «проклятия» размерности или вообще снять ее.

В работе [4] предложен алгоритм автоматического обучения большой искусственной нейронной сети с 512 входами и 256 выходами с использованием всего 20 примеров образа «Свой». Этот алгоритм успешно справляется с шумами дискретизации биометрических данных, обусловленных тем, что естественная многомерно-континуальная форма представления образа «Свой» замещается приближением из 20 примеров. Всегда, когда непрерывная функция (континуум) представляется малым числом состояний (примеров), возникают очень большие погрешности дискретизации, которые способны накапливаться при вычислениях (при обучении нейронных сетей). Мы привыкли к интегрированию 512 отсчетов измерений во времени и согласны с тем, что операция интегрирования повышает точность вычислений. Для многомерных систем вместо интегрирования по времени можно усреднять данные по 512 разным пространствам, это не менее эффективно давит естественные ошибки вычислений и делает их устойчивыми. Как следствие алгоритм работает тем лучше, чем выше размерность решаемой задачи. Возникает эффект, обратный эффекту «проклятия» размерности, который можно назвать «благодатью» высоких и сверхвысоких размерностей.

Применение энтропийного аппарата для снижения размерности задачи преобразования биометрия-код

Возможности интеллекта (искусственного и естественного) определяются тем, какой сложности (какой размерности) он может решать задачи. С точки зрения обычной математики, любую задачу можно записать как некоторую многомерную функцию $F(x_1, x_2, x_3, ..., x_n)$, где n – число учитываемых переменных. Очевидно, что размерность задачи и ее сложность можно оценивать в штуках по числу переменных. Это весьма и весьма приземленный, но очень эффективный подход.

Если переменные сильно зависимы, то они создают не более чем видимость высокой размерности. Измерение числа входных переменных в штуках и числа выходных состояний преобразователя биометрия-код также в штуках почти ничего не говорит о размерности решаемой задачи. Действительная размерность решаемой задачи преобразования биометрия-код связана исключительно со значением выходной энтропии кодовых состояний преобразователя биометрия-код, вычисленная при условии воздействия на него образами «Все Чужие».

Важность этого показателя легко продемонстрировать на примере шифрования текстов. Текст на русском или казахском языках легко читается носителями этих языков, если он не зашифрован. Разряды кодов открытого естественного языка имеют значительные коэффициенты корреляции. Наличие высоких значений коэффициентов парной корреляции между разрядами кодов – это признак всех биометрических кодов, в том числе кодов коллективной биометрии русскоговорящих и казахов. Напомним, что коэффициент парной корреляции или коэффициент Пирсона – это некоторое число от -1 до 1, показывающее тесноту линейной корреляционной связи между зависимой и независимой случайной величиной. Шифрование разрушает естественные корреляционные связи личных биометрических кодов и кодов парольных фраз, набранных на компьютере. После шифрования коды становятся «белым» шумом, состояния «0» и «1» в их разрядах становятся независимы, парные корреляции отсутствуют ($r_{i,j}$= 0.0 для всех i≠j).

Еще одной важной характеристикой «белого» шума является то, что энтропия каждого разряда таких кодов точно совпадает с одним битом, то есть длина действительно случайного кода, представляющего «белый» шум, совпадает со значением его энтропии. Энтропия случайного кода H длиной 256 бит точно равна 256 битам. Если же код оказывается не случайным, то его энтропия падает:

$$H(256) \langle 256 \; при \; |r_{i,j}| \rangle 0.0, \quad \text{хотя бы для одной пары } i \neq j. \qquad (1)$$

Для всех кодов (с коррелированными и независимыми разрядами) энтропия связана с показателем стойкости к атакам подбора или с вероятностью ошибок второго рода. Если речь идет о симметричном шифровании на ключе длиной 256 бит, то вероятность расшифровывания текста с первой попытки случайной подстановки ключа составит

$$P_2 = 2^{256} \text{ или } H(256) = 256 = -\log_2(P_2). \qquad (2)$$

В общем виде можно записать:

$$H(n) = -\log_2(P_2), \qquad (3)$$

где n – длина биометрического кода; P_2 – вероятность ошибок второго рода преобразователя биометрия-код.

Заметим, что ГОСТ Р 52633.0–2006 [5] накладывает ограничение на корреляционные связи в разрядах биометрических кодов «Чужой». Среднее значение модуля этих корреляционных связей не должно превышать 0.15. В случае, если $E(|r_{i,j}|) = 0.15$, показатель вероятности ошибок второго рода такого преобразователя падает примерно на порядок $P_2 = 2^{25.6}$. В сравнении с шифрованием происходит потеря стойкости к атакам подбора на $2^{31.4}$ порядка. При больших значениях модулей коэффициентов парной корреляции происходит катастрофическое падение стойкости преобразователя биометрия-код к атакам подбора. Если кому-то удастся сделать преобразователь биометрия-код, дающий 256-разрядный код с отсутствием в коде парных корреляций между разрядами, то такой преобразователь будет обладать стойкостью к атакам подбора, совпадающей со стойкостью защиты шифрованием с такой же длиной ключа.

Важным моментом при этом является вопрос вычисления энтропии выходных биометрических кодов. Понятно, что классический метод вычисления многомерной энтропии с использованием формулы Шеннона

$$H(256) = -\sum_{i=1}^{2^{256}} P_i \cdot \log_2(P_i), \qquad (4)$$

где P_i – вероятность появления *i*-го состояния биокода, требует огромных вычислительных затрат и огромных размеров исходных биометрических данных.

Действительно, выполнить оценку энтропии по формуле (4) для преобразователей биометрия-код с 256 выходами (число состояний 2^{256}) технически невозможно. Алфавит возможных выходных состояний биометрических кодов слишком велик, поэтому необходимо предпринять усилия по созданию более экономичных вычислительных процедур. Одним из вариантов значительного снижения объемов вычислений является переход из поля обычных кодов в поле кодов расстояний Хэмминга [6-7].

Переход в пространство расстояний Хэмминга

Входные биометрические данные, как правило, континуальны (непрерывны), а выходные коды преобразователя дискретны. Число возможных состояний кода всегда конечно. Любой преобразователь биометрия-код – это одно из возможных отображений многомерных континуумов в конечное число состояний кода заданной длины. Соответственно для исследования его выходных кодов может быть использовано расстояние Хэмминга. Для вычисления расстояний Хэмминга для двух кодов одинаковой длины их сравнивают поразрядно, далее подсчитывают число не совпавших разрядов. Если представить произвольный выходной код в виде вектора $\bar{x}$ с множеством состояний разрядов «0» и «1», то расстояние Хэмминга до кода «Свой» $\bar{c}$ вычисляется следующим образом:

$$h = \sum_{i=1}^{256} x_i \oplus c_i, \qquad (5)$$

где 256 – длина сравниваемых кодов; I – номер сравниваемых разрядов; $\oplus$ – операция сложения по модулю два.

На рис. 3 показано распределение расстояний Хэмминга для «нечетких экстракторов» и нейросетевых преобразователей для кодов длиной 256 бит.

Расстояние Хэмминга между выходными кодами «Чужие» и выходным кодом «Свой» измеряется в битах и оказалось очень удобной метрикой для статистических исследований [6-9]. Применение этой метрики к «нечетким экстракторам» и нейросетевым преобразователям биометрия-код дает разные распределения расстояний для образов «Свой» и «Все Чужие». «Нечеткие экстракторы» осуществляют квантование биометрических параметров образ до их обогащения. По этой причине коды образа «Свой» ведут себя нестабильно и имеют достаточно широкое распределение значений расстояний Хэмминга (см. верхнюю часть рис. 3).

Из-за того, что нейросетевые преобразователи способны улучшать качество биометрических данных образа «Свой», распределение Хэмминга для кодов «Свой» сжимается (прижимается к «нулевой» оси системы координат в нижней части рис. 3).

Из верхней части рисунка 3 видно, что, контролируя выходные коды «нечеткого экстрактора», необходимо принимать как коды «Свой» множество кодов с расстоянием Хэмминга до действительного кода «Свой» менее 40 бит. На выходе нейросетевого преобразователя биометрия-код коды «Свой» имеют намного меньшую нестабильность до 7 бит.

Еще одной важной особенностью распределения Хэмминга кодов «Все Чужие» является то, что оно очень хорошо описывается нормальным законом распределения значений. Это позволяет рассчитать ожидаемую вероятность ошибок второго рода по формуле (1).

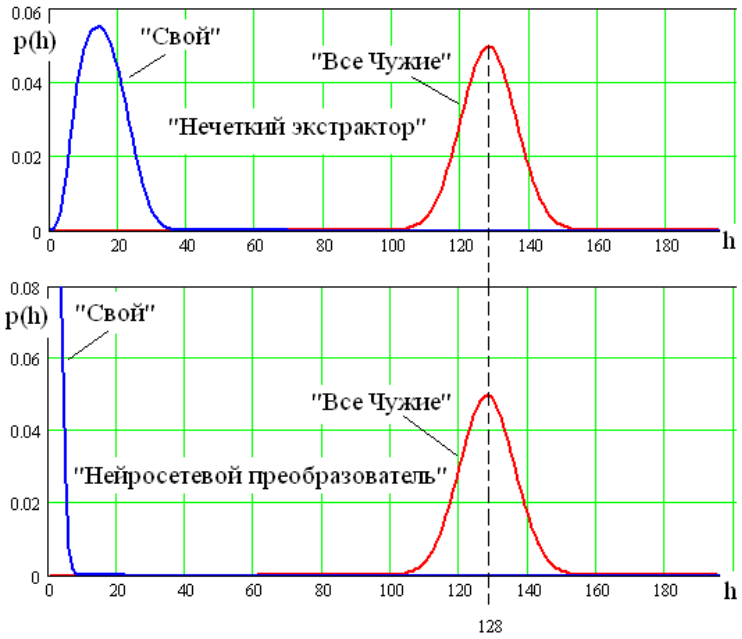

Рис. 3. Распределение расстояний Хэмминга для «нечетких экстракторов» нейросетевых преобразователей для кодов длиной 256 бит

Далее, вычисляя энтропию выходных кодов по формуле (4), мы фактически уменьшаем длину выходного кода до полноценной размерности задачи, рассматривая разницу между длиной кода в штуках и короткого кода в битах как некоторый бесполезный довесок с абсолютно коррелированными состояниями [6-9]. Эта ситуация отображена на рис. 4.

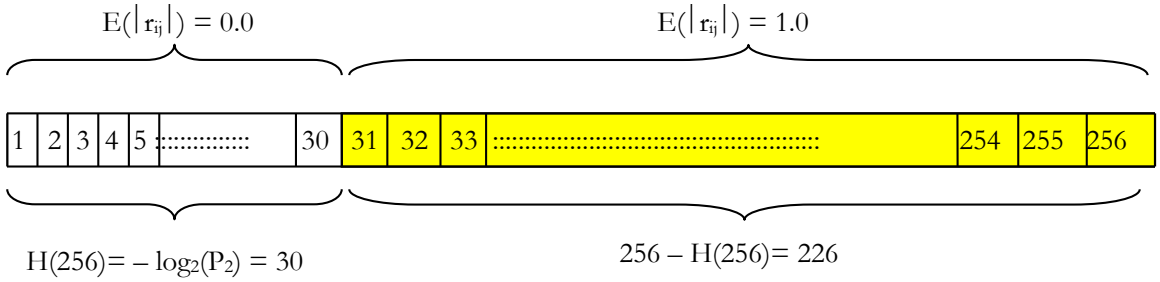

Рис. 4. Представление биокода длиной 256 бит со слабо коррелированными разрядами в виде 30 стойких некоррелированных разрядов и 226 бесполезных полностью коррелированных разрядов

При таком методе оценки размерность решаемой биометрической задачи может быть очень высокой даже при относительно малых длинах выходного кода, если его разряды слабо коррелированны. При росте корреляционных связей между разрядами кода размерность решенной задачи биометрической защиты падает. В случае, когда разряды кода полностью коррелированы при любой длине, его энтропия 1 бит, размерность также составляет 1 бит.

Число связей нейронов, слоев нейронов, входов у преобразователя биометрия-код и алгоритм обучения влияют косвенно на оценку размерности решаемой задачи.

Простое увеличение какого-то из перечисленных выше параметров без соответствующего роста энтропии выходных кодов преобразователя не приводит к росту размерности решенной задачи биометрической защиты информации [6-9].

Выводы. Таким образом, в работе показано, что алгоритм автоматического обучения большой искусственной нейронной сети является эффективным, так как нет необходимости вычислять частные производные и полностью размыкается петля обратной связи по ним. Кроме того, он обеспечивает высокую устойчивость вычислений. При этом устойчивость вычислений будет тем выше, чем больше входов у обучаемого нейрона. Вместо проблемы «проклятия» размерности появляется противоположный эффект «благодати» высоких и сверхвысоких размерностей. Описывается влияние размерности нейронных сетей на качество преобразователей биометрия-код. Рассматривается применение энтропийного аппарата для

снижения размерности задачи преобразования биометрия-код. Показано, что можно достичь снижения размерности входной выборки за счет учета корреляционных связей между выходными сигналами нейронных сетей.

ЛИТЕРАТУРА

[1]. A. Arakala, J. Jeffers, K. J. Horadam, "Fuzzy Extractors for Minutiae-Based Fingerprint Authentication", *Advances in Biometrics (LNCS 4642), Springer*, pp. 760-769, 2007.

[2]. А. Чморра, "Маскировка ключа с помощью биометрии", *Проблемы передачи информации*, № 2(47), С. 128-143, 2011.

[3]. Б. Ахметов, А. Иванов, А. Малыгин, В. Фунтиков, *Основы биометрической аутентификации личности*, Алматы: КазНТУ, 2014.

[4]. ГОСТ Р 52633.5–2011, «Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия-код доступа», М.: Стандартинформ, 2012.

[5]. ГОСТ Р 52633.0-2006, Защита информации. Техника защиты информации. Требования к средствам высоконадежной биометрической аутентификации. М.: Стандартинформ, 2007.

[6]. Б. Ахметов, О. Захаров, Т. Картбаев, А. Малыгин, А. Иванов, И. Огнев, "Метод оценки вероятностей появления ошибок нейросетевых преобразователей биометрия-код, использующий очень малые тестовые выборки", *Вестник КазНТУ имени К.И. Сатпаева*, 2013, №3(97), С. 279-283.

[7]. А. Иванов, Б. Ахметов, А. Безяев, К. Перфилов, Ж. Алимсеитова, "Вычисление энтропии слабо коррелированных и сильно коррелированных длинных биометрических кодов на малых тестовых выборках", *Вестник НАН РК*, 2015, №3, С. 64-70.

[8]. Б. Ахметов, А. Иванов, Т. Картбаев, Д. Надеев, А. Малыгин, И. Огнев, "Энтропийно-корреляционный подход к расчету вероятности совместного появления большого числа зависимых событий", *Вестник КБТУ*, 2013, №2(25), С. 54-58.

[9]. А. Малыгин, Б. Ахметов, В. Волчихин, И. Урнев, "Учет влияния корреляционных связей на результаты тестирования преобразователей биометрия-код", *Информационные и телекоммуникационные технологии: образование, наука, практика: Сборник трудов Международной научно-практической конференции*, Алматы: КазНТУ, 2012, С. 34–37.

REFERENCES

[1]. A. Arakala, J. Jeffers, K. J. Horadam, "Fuzzy Extractors for Minutiae-Based Fingerprint Authentication", *Advances in Biometrics (LNCS 4642), Springer*, pp. 760-769, 2007.

[2]. A. Cimorra, "Masking key using ", *Problems of information transmission,* 2011, no. 2(47), P. 128-143.

[3]. B. Akhmetov, A. Ivanov, A. Malygin, V. Funtikov, *A fundamentals of biometric person authentication*, Almaty: KazNTU, 2014.

[4]. GOST R 52633.5–2011, "Technique of information security. Automatic learning of neural transmitters biometrics access code", M.: STANDARTINFORM, 2012.

[5]. GOST R 52633.0-2006, "Technique of information security. Requirements to the means of highly reliable biometric authentication", M.: STANDARTINFORM, 2007.

[6]. B. Akhmetov, A. Zakharov, T. Karibaev, A. Malygin, A. Ivanov, I. Ognev, "Method of evaluation of probability of error neural network converters biometrics code using very small test samples", *Bulletin of KazNTU named after K. I. Satpayev*, 2013, no. 3(97), P. 279-283.

[7]. A. Ivanov, B. Akhmetov, A. Bazaev, K. Perfilov, Zh. Alimseitova, "The Calculation of the entropy of weakly correlated and strongly correlated long biometric codes on small test samples", *Bulletin of NAS RK*, 2015, no. 3, P. 64-70.

[8]. B. Akhmetov, A. Ivanov, T. Karabaev, D. Nadeev, A. Malygin, I. Ognev, "Entropy-correlation approach to calculating the probability of the joint occurrence of a large number of dependent events", *Vestnik of KBTU*, 2013, no. 2(25), P. 54-58.

[9]. A. Malygin, B. Akhmetov, V. Volchikhin, I. Ornev, "To account for the influence of correlations on the results of testing converters biometrics code", *Information and telecommunication technologies: education, science, practice: proceedings of the International scientific-practical conference*, Almaty: KazNTU, 2012, P. 34-37.

THE PROBLEM OF PATTERN RECOGNITION DIMENSIONALITY IN BIOMETRIC AUTHENTICATION SYSTEMS

In biometric authentication systems, the process of proving and verifying the authenticity of a user-claimed name through the user's presentation of his biometric image is performed and by converting this image in accordance with a predefined authentication protocol. An important issue remains the transformation of biometric data into code. The paper discusses the two most well-known conversion technology of biometrics in the code, the scheme of conversion of biometric parameters in the code key. It is shown that one of the main reasons for the difficulties of biometric authentication is the high dimensionality of the problem. To solve this problem, there are artificial neural networks or "fuzzy extractors". Of the many existing learning algorithms of neural networks selected algorithm for automatic training of large artificial neural networks. Shows the use of entropy device to reduce the dimension of the problem of converting the biometrics-code. To reduce the volume of calculations made the transition to the Hamming distances.

Keywords: biometric image, artificial neural network, neural network Converter, the dimensionality of the task of transformation, correlation, space of distance of the Hamming.

ПРОБЛЕМИ РОЗМІРНОСТІ ЗАДАЧ РОЗПІЗНАВАННЯ ОБРАЗІВ У СИСТЕМАХ БІОМЕТРИЧНОЇ АУТЕНТИФІКАЦІЇ

У системах біометричної аутентифікації здійснюється процес доказу і перевірки автентичності заявленого користувачем імені через пред'явлення користувачем свого біометричного образу і шляхом перетворення цього образу відповідно до заздалегідь визначених протоколів аутентифікації. Важливим питанням залишається перетворення біометричних даних в код. У статті розглядаються дві найбільш відомі технології перетворення біометрії в код, наводиться схема перетворення біометричних параметрів в код ключа. Показано, що однією з основних причин труднощів біометричної аутентифікації є висока розмірність завдання. Для вирішення цієї проблеми використовуються штучні нейронні мережі або «нечіткі екстрактори». З безлічі існуючих алгоритмів навчання нейронних мереж обраний алгоритм автоматичного навчання великої штучної нейронної мережі. Відображено застосування ентропійного апарату для зниження розмірності задачі перетворення біометрія-код. Для зниження обсягів обчислень проведений перехід до відстаней Хеммінга.

Ключові слова: біометричний образ, аутентифікація, штучна нейронна мережа, нейромережевий перетворювач, розмірність завдання перетворення, кореляція, простір відстаней Хеммінга.

Алімсеітова Жулдиз, лектор кафедри «Інформаційна безпека» Сатпаєв Університету.
E-mail: zhuldyz_al@mail.ru
Алимсеитова Жулдыз, лектор кафедры «Информационная безопасность» Сатпаев Университета.
Alimseitova Zhuldyz, lecturer of Academic Department «Information security» at Satpayev University.

Сейлова Нургуль, кандидат технічних наук, завідувач кафедри «Інформаційна безпека» Сатпаєв Університету.
E-mail: seilova_na@mail.ru
Сейлова Нургуль, кандидат технических наук, заведующая кафедры «Информационная безопасность» Сатпаев Университета.
Seilova Nurgul, PhD in Eng, Head of Academic Department «Information security» at Satpayev University.

Гнатюк Сергій Олександрович, кандидат технічних наук, доцент, доцент кафедри безпеки інформаційних технологій Національного авіаційного університету.
E-mail: s.gnatyuk@nau.edu.ua
Гнатюк Сергей Александрович, кандидат технических наук, доцент, доцент кафедры безопасности информационных технологий Национального авиационного университета.
Gnatyuk Sergiy, PhD in Eng, Associate Professor of IT-Security Academic Department, National Aviation University.