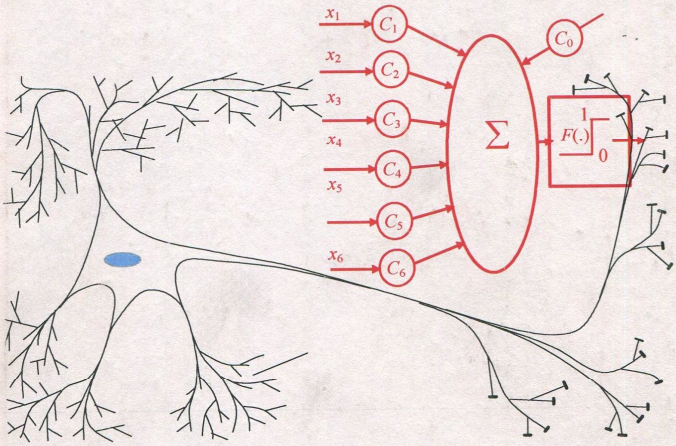
БЫСТРЫЕ АЛГОРИТМЫ ТЕСТИРОВАНИЯ ВЫСОКОНАДЕЖНЫХ НЕЙРОСЕТЕВЫХ МЕХАНИЗМОВ БИОМЕТРИКО-КРИПТОГРАФИЧЕСКОЙ ЗАЩИТЫ ИНФОРМАЦИИ



ПЕНЗА 2006

Быстрые алгоритмы тестирования высоконадежных нейросетевых механизмов биометрико-криптографической защиты информации

Издательство
Пензенского государственного
университета
Пенза 2006

УДК: 519.7+621

Б95

Р е ц е н з е н т ы :

Доктор технических наук, профессор,
член-корреспондент Академии качества России, заместитель директора по науке
Пензенского филиала ФГУП НТЦ «Атлас» ФСБ России

Г. Н. Чижухин

Доктор технических наук, заведующий кафедрой
«Вычислительные системы и моделирование» Пензенского государственного
педагогического университета им. В. Г. Белинского

В. И. Горбаченко

Б95 Быстрые алгоритмы тестирования высоконадежных нейросетевых механизмов биометрико-криптографической защиты информации: монография / А. Ю. Малыгин, В. И. Волчихин, А. И. Иванов, В. А. Фунтиков – Пенза: Изд-во Пенз. гос. ун-та, 2006. – 160 с. : ил. – Библиогр. : с. 149–154.

Излагаются основы процедур ускоренного тестирования высоконадежных нейросетевых механизмов биометрико-криптографической защиты информации. Рассматриваются как нейросетевые механизмы с одним дискретным выходом, так и многомерные нейросетевые механизмы, дающие на выходе дискретный вектор-решение. Показано, что знание функции закона распределения значений выходных кодов многомерных нейросетевых преобразователей позволяет на десятки порядков снизить число тестовых примеров, используемых при численных статистических экспериментах. Формулируются требования к представительным тестовым выборкам примеров реальных биометрических образов. Обсуждается проблема увеличения баз тестовых примеров реальных биометрических образов за счет их увеличения тестовыми синтетическими биометрическими образами.

Монография предназначена для специалистов, занимающихся приложением искусственных нейронных сетей при решении задач защиты информации в платежных системах, системах электронного документооборота, в системах электронных паспортов и удостоверений личности с автоматической биометрической идентификацией личности.

УДК: 519.7+621

О Г Л А В Л Е Н И Е

Предисловие

В настоящее время идут активные процессы информатизации современного общества. Одним из направлений информатизации является использование биометрических технологий идентификации личности. Современная биометрическая идентификация личности [1] позволяет передать от людей искусственному нейросетевому интеллекту способность с высокой надежностью узнавать конкретного человека. Особую важность биометрические технологии идентификации получают в связи с введением в ближайшем будущем нового поколения загранпаспортов и визовых документов. С этой целью в декабре 2002 г. создан специальный подкомитет ISO/IEC JTC1 SC37 (Biometrics), призванный подготовить в ближайшее время несколько десятков международных биометрических стандартов. Более 37 международных стандартов касаются различных особенностей биометрии. Они описывают то, как нужно снимать, хранить, обрабатывать отпечатки пальцев, изображения лица, рукописные образы. Как минимум, пять международных стандартов [2, 3, 4, 5, 6] планируется посвятить вопросам тестирования биометрических устройств и технологий.

Международные стандарты [2, 3, 4, 5, 6] касаются очень важных вопросов тестирования, несомненно, что перевод этих стандартов на русский язык и введение в ближайшем будущем их как национальных стандартов (ГОСТ Р) станет для России важным шагом в направлении глобальной информатизации. Однако ожидать решения всех проблем тестирования биометрии от международных документов не приходится. Это, прежде всего, связано с тем, что разрабатываемые международным подкомитетом ISO/IEC JTC1 SC37 стандарты относятся к относительно «слабой» биометрии, способной идентифицировать личность человека только локально, под прямым кон-

тролем проверяющего. Проверяющий (например, пограничник) должен обязательно контролировать действия проверяемого при его автоматизированной биометрической идентификации. Разрабатываемый ISO/IEC JTC1 SC37 пакет международных биометрических стандартов нельзя использовать для дистанционной биометрической идентификации человека, например, через Интернет.

Для того, чтобы обеспечить высоконадежную дистанционную идентификацию человека, необходимо привлекать биометрические технологии, способные безопасно взаимодействовать с криптографическими механизмами. В настоящее время практически все страны, имеющие значимый национальный научно-технический потенциал, пытаются решать эти задачи. Две страны – Россия и США, являясь лидерами технологий защиты информации, открыто публикуют результаты своих исследований. США идут по пути использования нечеткой математики [7], ученые этой страны предлагают мировому сообществу специализированные «fuzzy» обогатители (экстракторы), превращающие бедную неоднозначную размытую биометрическую информацию в сильный личный ключ пользователя.

Россия предлагает мировому сообществу иной путь использования больших и сверхбольших искусственных нейронных сетей, которые заранее обучаются преобразовывать размытые биометрические данные пользователя в его личный криптографический ключ. Теория создания подобных преобразователей биометрия/код [1, 8] позволяет надеяться на их высокую стойкость по отношению к атакам подбора и попыткам изучения. Правильно построенный преобразователь биометрия/код ведет себя как классическая необратимая хэш-функция. Случайные входные биометрические образы нейросетевой преобразователь биометрия/код перемешивает (хэширует), а заранее известное множество нечетких образов «Свой» преобразователь свертывает в единственное значение личного криптографического ключа. При этом достаточно сложная нейронная сеть с 256 выходами позволяет обеспечивать стойкость к атакам случайного подбора на уровне 10^{22} (22-я степень) попыток. Для того, чтобы проверить качество российских биометрико-нейросетевых преобразователей, потребуется сформировать огромные базы случайных биометрических образов.

Практика показывает, что на воспроизведение одного случайного биометрического образа, например в форме случайного рукописного пароля, у человека уходит порядка 10 с. Соответственно, на формирование достаточно большой базы случайных биометрических образов уйдет 10^{23} с. Один год составляет всего лишь 10^6 с, т. е. на формирование нужной базы случайных биометрических образов силами одного добровольца уйдет 10^{17} лет, что существенно больше возраста Земли. Даже если привлечь к работе все население Пензенской области (1,42 млн жителей), сроки создания базы сократятся до 10^{11} лет, что также превышает возраст Земли, солнечной системы и нашей галактики. Из подобных достаточно простых расчетов вытекает бесперспективность всех попыток оценки стойкости современных биометрико-нейросетевых преобразователей прямым численным экспериментом. Эта задача сопоставима по своей сложности с задачей подбора криптографического ключа [9].

Необходимо подчеркнуть, что по мере информатизации современного общества проблемы его информационной безопасности будут усиливаться. В частности, по прогнозам специалистов уже в ближайшее время обществом будет ощущаться проблема «цифрового неравенства» граждан. Предполагается, что уже в очень близком будущем электронный документооборот получит широкое распространение. Как следствие, повсеместно станет использоваться электронная цифровая подпись (ЭЦП).

Широкое использование ЭЦП ставит всех в неравные условия. Например, если рассматривать цифровые права банкира и рядового гражданина, то юридически они равны, однако практически это далеко не так. Доверие к электронной цифровой подписи банкира намного выше доверия к ЭЦП рядового гражданина [10]. Это связано с тем, что у банкира есть сейф, охрана, таким образом, он может обеспечить надежное хранение своего личного криптографического ключа, формирующего электронную цифровую подпись электронного документа. В отличие от банкира, ключ формирования ЭЦП которого всегда хранится в сейфе, рядовой гражданин не может себе этого позволить. Его ключ формирования ЭЦП, скорее всего, будет храниться в кошельке (или кармане), что автоматически ставит его в более уязвимое положение.

Именно это обстоятельство и называется действительным «цифровым неравенством», когда декларированная для всех одинаковая юридическая значимость электронной цифровой подписи на деле будет иметь разный уровень доверия.

Ликвидировать подобное неравенство может только государство, предпринимая специальные меры, уравнивающие цифровые права всех граждан независимо от их социального статуса. Предвидя возникновение и усиление «цифрового неравенства», государству необходимо создавать специальные механизмы, сглаживающие изначальное неравенство.

В плане противодействия «цифровому неравенству» своих граждан Россия по праву занимает лидирующее положение, формальным подтверждением является разработка ею своего национального стандарта [11], регламентирующего требования к средствам высоконадежной биометрии. Одним из главных требований ГОСТа [11] является наличие средств встроенного контроля вероятности ошибок средств защиты или вероятности удачи атаки подбора. Проблема состоит в том, что биометрический пароль (рукописный или голосовой) пользователь должен сохранить в тайне от всех. Пользователь должен сам придумать удобную для него цифровую комбинацию, слово, фразу. При этом пользователю нельзя доверять среднестатистическим характеристикам, заявленным производителем.

Тайный биометрический образ может оказаться слабым и обеспечивать низкую стойкость защиты к атакам подбора. Чтобы убедиться в стойкости нейросетевой защиты на конкретном биометрическом образе, после обучения нейросети необходимо протестировать стойкость преобразователя. Для этого необходимо использовать специальные методы ускоренного тестирования стойкости биометрико-нейросетевой защиты [9, 12, 13].

Возникает целый комплекс вопросов, связанный с ускоренным тестированием средств биометрико-нейросетевой защиты. Кроме того, к этому комплексу примыкают вопросы сертификации и полного (неускоренного) тестирования средств защиты самим производителем или некоторым независимым (например, государственным) органом сертификации (испытаний). В свою очередь, испытательный центр (лаборатория) должен иметь соответствующие методики ис-

пытаний и большие базы случайных биометрических образов, верно отражающих реальную статистику распределения биометрических параметров «Своих» пользователей и наиболее вероятную статистику (тактику) организации потенциальным злоумышленником атак подбора.

Для решения перечисленных выше вопросов при факультете военного обучения Пензенского государственного университета была создана межведомственная лаборатория тестирования биометрических устройств и технологий [12]. Изложенные ниже материалы во многом являются результатами совместного труда коллектива этой лаборатории.

Г л а в а 1

Тестирование относительно «слабых» биометрико-нейросетевых механизмов с низкой размерностью выходного вектора принимаемых решений

1.1. Классический подход к тестированию относительно «слабых» нейросетевых решений низкой размерности

Переход от парадигмы классического программирования (от парадигмы переноса в программу полностью детерминированных знаний) к парадигме обучения программ нечетким знаниям (например, обучения программного эмулятора нейросети на примерах) приводит к резкому возрастанию роли тестирования полученных нейросетевых решений. Тестирование нейросетевых решений становится неотъемлемым элементом их обучения (программирования). Классической рекомендацией является дробление всех примеров пополам и использование первой половины для обучения, а второй половины – для тестирования [14, 15, 16, 17].

С одной стороны, тратить половину примеров обучения на промежуточное тестирование – это расточительство, однако, с другой стороны, такой подход является гарантией статистической сбалансированности процедур обучения и тестирования. Естественно, что с ростом числа примеров в обучающей выборке проблема их статистической балансировки ослабляется. Для иллюстрации этой связи на рисунке 1.1 приведены аппроксимации распределения примеров рукописных образов «а» (центр – m1), рукописных образов «в» (центр – m2) и тестовой выборки образов «а» (центр – mt).

Из рисунка 1.1 видно, что тестовая выборка из 6 образов «а» расходится примерно на 35 % с аналогичной обучающей выборкой этого же образа. Если мы увеличим тестовую и обучающую выборку до 16 примеров, то получим ситуацию, отображенную на рисунке 1.2.

По мере роста размеров обучающей и тестовой выборок ошибочное расхождение между их распределениями падает.

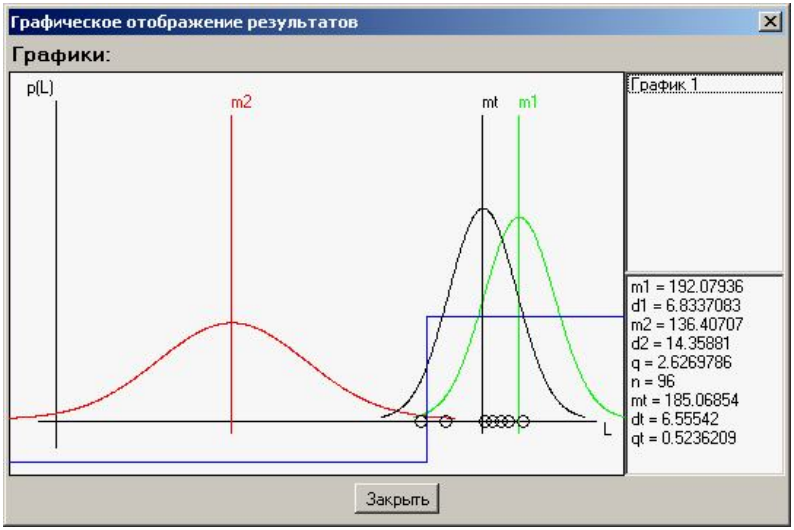


Рисунок 1.1 – Расхождение порядка 35 % от площадей распределений обучающей выборки из 6 образов «а» с центром m1 и тестовой выборки из 6 образов «а» с центром mt

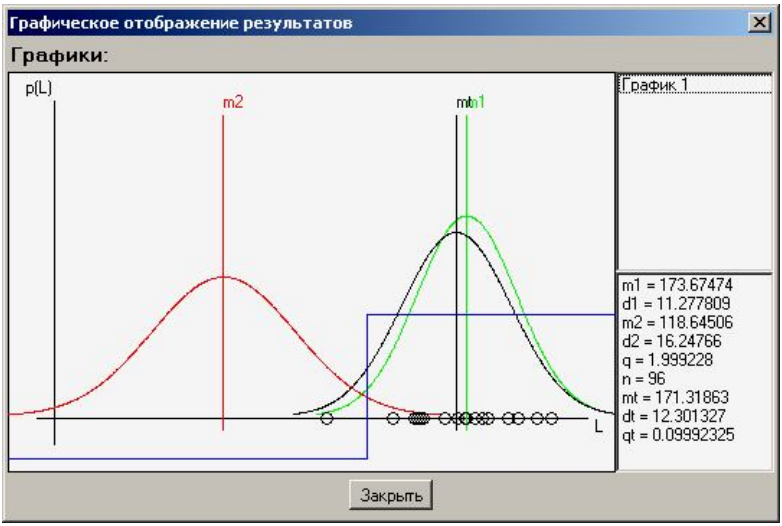


Рисунок 1.2 – Расхождение порядка 7 % от площадей распределений обучающей выборки из 16 образов «а» с центром m1 и тестовой выборки из 16 образов «а» с центром mt

Приведенные выше рисунки получены в среде моделирования нейронных сетей «Нейропреподаватель», предназначенной для проведения студентами лабораторных работ. В этом программном продукте для отображения законов распределения значений используется их аппроксимация в виде нормального закона распределения значений. Эта форма представления не случайна. Если рассматривать один из выходов любой нейронной сети, то мы будем иметь некоторый пороговый элемент на выходе сумматора. Сумматор, как известно, является нормализующим элементом. В соответствии с известной предельной теоремой статистики [18, 19] суммирование множества независимых (слабо зависимых) случайных величин с произвольными законами распределений в итоге дает распределение, близкое к нормальному. Эта связь асимптотическая, чем больше входов у сумматора, тем сумматор лучше нормализует выходные данные. Естественно, что у сумматора последнего нейрона весовые коэффициенты будут разными, однако это никак не отражается на содержании асимптотической связи. При любых весовых коэффициентах сумматор остается нормализующим элементом, что дает формальное право широко использовать гипотезу нормальности выходных законов распределения его данных.

Гипотеза нормальности законов распределения данных на выходе сумматора нейрона является весьма эффективным инструментом предсказания ожидаемого качества принимаемых последним нейроном решений. Используя эту гипотезу, мы можем оценить вероятности ошибок первого рода и второго рода вообще без привлечения дополнительных тестовых примеров по следующей формуле:

$$P_1 = P_2 \approx 0{,}5 - \frac{1}{\sqrt{2\pi}} \int_0^q \exp\left(\frac{-x^2}{2}\right) \mathrm{d}x = 0{,}5 - \Phi_0(q), \qquad (1.1)$$

где $\Phi_0(.)$ – односторонняя функция Лапласа (односторонний интеграл Лапласа);

$q = \dfrac{|m_1 - m_2|}{\sigma_1 + \sigma_2}$ – логарифмический показатель качества обучения [1],

где m_1, m_2 – математические ожидания первого и второго разделяемых множеств; σ_1, σ_2 – среднеквадратические отклонения первого и второго разделяемых множеств.

Если воспользоваться соотношением (1.1) для оценки вероятностей ошибок, то для ситуации на рисунке 1.1 мы получим $P_1 = P_2 = 1 - \Phi_0(2,6) = 0,004$, а для ситуации с большим числом примеров, отображенной на рисунке 1.2, мы получим худшие результаты предсказания $P_1 = P_2 = 1 - \Phi_0(1,9) = 0,024$.

Сравнивая приведенные выше прогнозы, можно допустить ошибку, предположив, что на малых обучающих выборках получаются лучшие результаты обучения. То, что это далеко не так, показывают тестовые выборки. Учет поправки на тестирование осуществляется следующим образом:

$$P_1 = P_2 \approx 0,5 - \frac{1}{\sqrt{2\pi}} \int_0^{q-2q_t} \exp\left(\frac{-x^2}{2}\right) dx = 0,5 - \Phi_0(q - 2q_t), \qquad (1.2)$$

где q_t – логарифмический показатель расхождения распределений примеров обучающей выборки и аналогичной тестовой выборки на линейном выходе обученного нейрона.

Соотношение (1.2) для ситуации рисунка 1.1 дает $P_1 = P_2 = 1 - \Phi_0(2,6 - 1) = 0,05$, а для ситуации на рисунке 1.2 получим $P_1 = P_2 = 1 - \Phi_0(1,9 - 0,2) = 0,036$.

Учет независимого тестирования дает гораздо более корректные предсказания.

1.2. Особенности обучения и тестирования относительно «слабых» биометрико-нейросетевых решений

В предыдущем параграфе была рассмотрена общая постановка задачи разделения двух сопоставимых по размерам классов образов. Такая постановка задачи не характерна для биометрии. Биометрические системы обучаются таким образом, чтобы выделять из общего широкого класса случайных образов «Все чужие» узкий подкласс «Свой». Для этой цели при обучении нейросети используется несколько примеров образов «Свой», примером могут служить образы, отображенные на рисунке 1.3.

Очевидно, что число примеров образов «Свой», используемых при обучении, не должно быть слишком большим. Это связано не только с обеспечением мер безопасности в процессе обучения нейросети, но и с учетом интересов самого пользователя. Поэтому производители стараются создавать максимально дружественные к пользователю биометрические системы.

Рисунок 1.3 – Пример обучающей выборки «Свой»,
состоящей из 9 рукописных образов слова-пароля «*Пенза*»

Проведенное реальное тестирование показало, что обычный пользователь, прошедший краткий ознакомительный курс с устройствами ввода рукописного слова-пароля: планшетом или экраном карманного персонального компьютера и программным обеспечением, обычно без особого напряжения способен воспроизвести до 20 примеров пред-

ложенного им слова-пароля, затрачивая на это примерно 2–3 мин времени.

Практическая работа по сбору баз натуральных биометрических образов написания слова-пароля показывает, что если требовать от пользователя воспроизводить большее число примеров своих биометрических образов, то они воспроизводят их с явной неохотой. Видимо, системы, требующие воспроизведения 40, …, 60 примеров для обучения, будут иметь существенно ограниченное использование при доступе только к очень ответственным приложениям.

Численность примеров случайных образов «Все чужие» может быть любой. Эти образы необязательно воспроизводить рукою в процессе обучения, они могут быть зашиты в программу обучения, и, соответственно, их число определяется только технологическими потребностями конкретного алгоритма обучения. На рисунке 1.4 приведены 9 примеров таких случайных рукописных образов, которые вполне могут являться частью более обширной обучающей выборки.

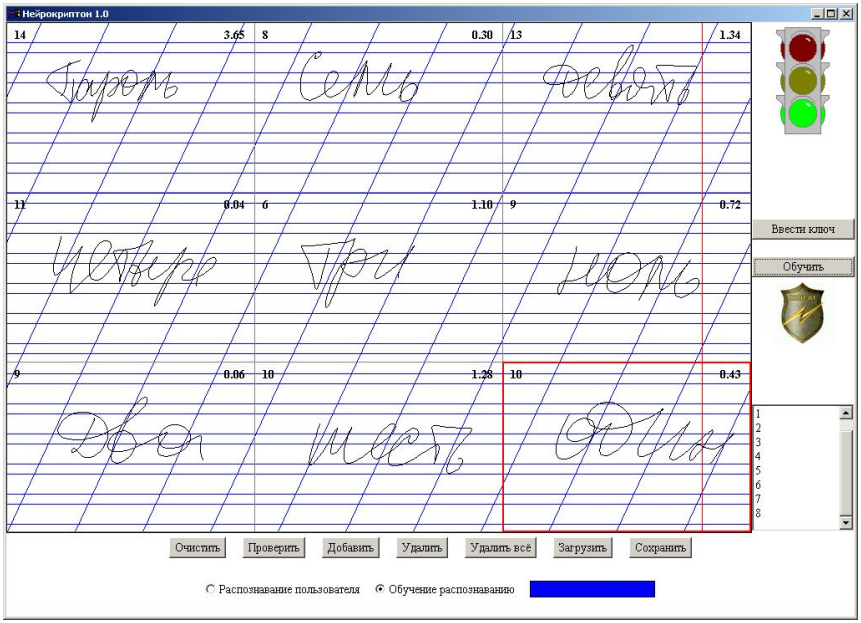


Рисунок 1.4 – Пример обучающей выборки «Чужие»,
состоящей из 9 случайно выбранных рукописных образов

Обычно после обучения относительно «слабые» биометрико-нейросетевые системы с одним выходом имеют вероятность ошибок первого и второго рода на уровне от 0,01, …, 0,05. В качестве примера на рисунке 1.5 даны выходные распределения подобной системы биометрической защиты. В результате обучения нейронов быстрыми декорреляционными алгоритмами [1, 20, 21, 22, 23] происходит выталкивание выделяемого узкого биометрического образа «Свой» на периферию более общего (более широкого) множества случайных образов «Все чужие».

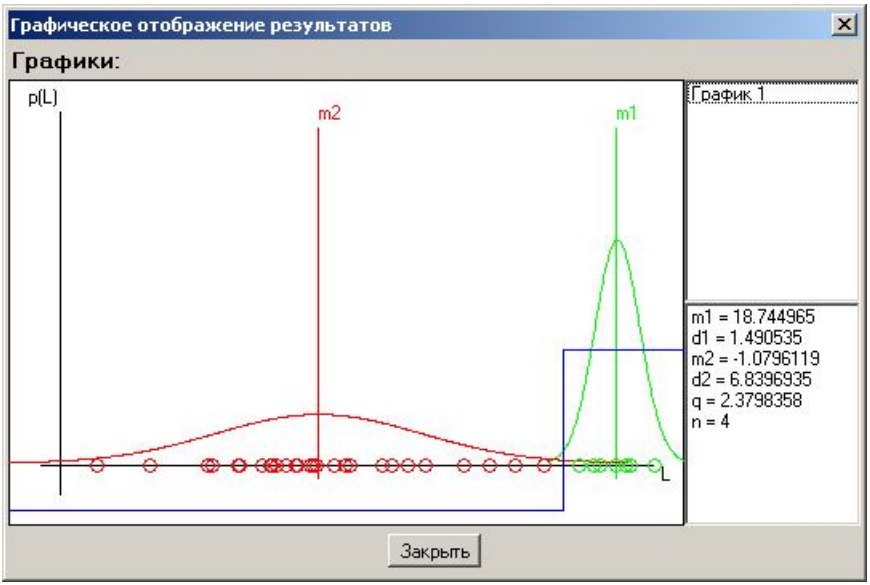


Рисунок 1.5 – Результат обучения одного нейрона выделять рукописный образ «Свой» из множества случайных рукописных образов «Чужие» ($P_{EE} = 0{,}008$)

Одной из важных особенностей биометрических систем защиты информации является то, что для них не существует проблемы оценки вероятности ошибок первого рода (ошибочного отказа «Своему»). Во-первых, параметр P_1 не является критическим для работоспособности биометрической защиты и может изменяться в широких пре-

делах от 0,01 до 0,25. Значение этого параметра имеет скорее психологическое, чем практическое значение. Даже в наихудшем случае $P_1 = 0{,}25$ пользователь получает отказ в доступе с вероятностью 0,016, если система предоставляет ему хотя бы три попытки. Если система относится к классу высоконадежных [11], то она вообще может не иметь ограничений по числу попыток доступа, т. е. реальное значение P_1 для «настырного» легального пользователя высоконадежной биометрии всегда является нулевым.

В связи с вышеизложенным в системах биометрической защиты стараются вообще не жертвовать примерами образов «Свой» на тестирование. Все примеры дефицитных образов «Свой» используются для обучения. Проблема прогноза значения параметра P_1 решается путем вычисления параметров нормального закона распределения выходных значений образов «Свой» и расчетами по формуле (1.1).

Тестирование на образах «Чужие» для «слабых» систем биометрической защиты также не является особо трудной проблемой [24, 25, 26]. При вероятностях ошибок второго рода на уровне $P_2 \approx 0{,}001$ для прямых статистических оценок достаточно базы случайных биометрических образов порядка 100000 примеров. Сбор, хранение и использование подобных баз тестовых образов при современных технических возможностях не представляет сколько-нибудь значимых трудностей.

1.3. Оценка размеров независимых тестовых испытаний, необходимых при прямых вычислениях вероятностей

В том случае, если о законе распределения значений разделяемых множеств ничего неизвестно, сокращать тестовые выборки нельзя в принципе. При полном отсутствии знания о виде законов распределения разделяемых нейронной сетью множеств корректное тестирование может быть осуществлено только прямым экспериментом. При этом погрешность определения вероятностей ошибок первого и второго рода полностью определяется числом тестовых примеров. При нулевом числе тестовых примеров погрешности (бесконечны) полностью неопределенны. По мере роста числа использованных тестовых

примеров модуль погрешности оценки значений вероятностей ошибок первого и второго рода асимптотически уменьшается до нуля.

Рассмотрим эту ситуацию более подробно на примере оценки вероятностей ошибок первого рода. Для этой цели требуется от пользователя последовательно воспроизводить свой биометрический образ. Например, если пользователь воспроизвел свой биометрический образ 20 раз и ни разу не получил отказа в доступе, то мы можем утверждать, что вероятность ошибок первого рода не хуже 1/21 = 0,047. Формальная логика построения этой оценки сводится к тому, что прямое вычисление вероятности дает неправдоподобно хороший результат 0/20 = 0. Верить слишком хорошему результату $P_1 = 0$ нельзя. Мы вынуждены использовать наихудший прогноз, состоящий в том, что следующая 21-я попытка будет неудачной. Этот наихудший прогноз и дает более правдоподобный результат $P_1 = 0{,}047$. Рассчитывая вероятность по конечной тестовой выборке, мы всегда решаем дискретную задачу со своей ошибкой дискретизации.

Если попытаться учесть дискретный характер расчетов вероятности ошибки конечной тестовой выборки из N примеров, то мы получим номограмму оценки относительной ошибки дискретизации в виде двухмерной функции $\Delta P(P,N)$. Номограмма функций относительных ошибок из-за дискретности задачи приведена на рисунке 1.6.

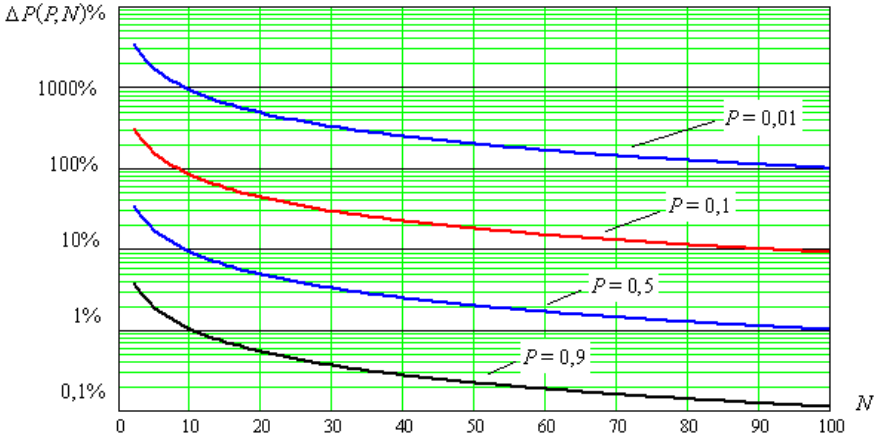


Рисунок 1.6 – Номограмма оценки относительных ошибок дискретизации, обусловленных конечностью (дискретностью) тестовой выборки

Из номограммы видно, что большие значения вероятностей оцениваются намного точнее, чем малые значения вероятностей. Очевидно, что 100 % относительная ошибка при прямом вычислении вероятности мало кого устраивает. Относительная ошибка должна быть хотя бы на уровне 10 % .

Для обеспечения 10 % относительной ошибки при прямых расчетах вероятностей потребуется примерно в 10 раз больше тестовых испытаний, чем величина, обратная оцениваемой вероятности. Чем меньше оцениваемая величина, тем точнее выполняется это правило:

– при $P = 0{,}1$ требуется 89 тестов;

– при $P = 0{,}01$ требуется 990 тестов;

– при $P = 0{,}001$ требуется 9998 тестов;

– при $P = 0{,}0001$ требуется 99999 тестов.

Таким образом, при оценке очень малых вероятностей прямым численным экспериментом требуется выполнять примерно в 10 раз больше испытаний, чем обратная величина оцениваемой вероятности. При условии десятикратной избыточности числа тестов нет необходимости дожидаться положительного исхода хотя бы в одном тесте. Это очень важное условие тестирования. При оценках очень малых вероятностей можно вообще никогда не дождаться первого положительного исхода, однако подсчет уже выполненных отрицательных тестов дает нам право обоснованно заявлять о достижении гарантированной оценки сверху контролируемой величины. Естественно, что такие гарантии появляются только при условии независимости привлеченных к испытаниям тестовых примеров.

Очевидно, что если допустимо увеличение погрешности оценки, то можно сократить избыточность тестовой выборки, т. е. для достижения 30 % относительной погрешности вычислений вероятности можно в три раза сократить размер тестовой выборки. При точном совпадении обратной величины оцениваемой вероятности и размеров тестовой выборки погрешность расчетов может достигать 100 %.

1.4. Корректное сокращение необходимого числа тестовых примеров

за счет гарантированно нормального закона распределения выходных данных

Заметим, что описанная выше избыточность числа тестовых примеров биометрических образов при испытаниях высоконадежных систем защиты становится технически не реализуемой. Например, для высоконадежной биометрии рядовой является ситуация, когда ожидаемая вероятность ошибки второго рода составляет величину, близкую к 10^{-17}. Это означает, что для прямого вычисления столь малой величины потребуется использовать не менее 10^{18} тестовых примеров. Если это биометрическая защита, анализирующая рукописное слово-пароль (время написания порядка 10 с), то на написание 10^{18} слов одним человеком потребуется 10^{13} лет. Очевидно, что такой тест выполнить нереально [27].

Гораздо более реальным является воспроизведение всего 70 случайно выбранных слов из книги на случайно открытой странице. На рукописное воспроизведение уйдет не более 12 мин. При такой тестовой выборке удается вычислить математическое ожидание откликов с относительной ошибкой 2,7 % (m = 127,4 ± 3,4). Дисперсию удается вычислить с существенно большей относительной ошибкой 21 % (σ = = 13,6 ± 2,9). Это означает, что в наихудшем случае мы можем оценить снизу значение показателя степени стойкости защиты с относительной ошибкой 28 %. Эта ситуация отображена на рисунке 1.7.

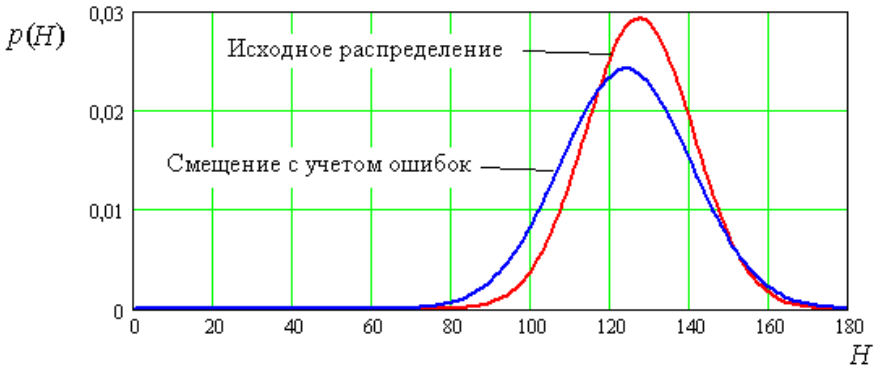


Рисунок 1.7 – Смещение закона, описывающего распределение выходных данных преобразователя, полученное с учетом влияния ошибок определения двух первых статистических моментов

Для ошибок вычисления математического ожидания наихудшими являются те, которые смещают распределение значений множества «Чужие» в левую сторону (к нулю). Соответственно, наихудшее для системы биометрической защиты значение математического ожидания составит $m = 127{,}4 - 3{,}4 = 124$. Наихудшее значение дисперсии будет наибольшим, т. е. $\sigma = 13{,}6 + 2{,}9 = 16{,}5$.

Исходное распределение значений (без учета конечности тестовой выборки) в рамках гипотезы нормального закона распределения дает оценку стойкости биометрической защиты $10^{16{,}84}$. После сдвига влево и расширения дисперсии (левое множество примеров рисунка 1.7) его распределение дает существенно меньшую оценку стойкости – $10^{13{,}93}$, т. е. знание закона распределения контролируемых данных на выходе нейронной сети позволяет сократить размеры тестовой выборки на 16 порядков. При этом мы гарантируем относительную ошибку вычисления показателя степени стойкости к атакам подбора – 28 %. По сравнению с изначальной оценкой, стойкость снизилась на 2 порядка, но при этом мы имеет инженерную гарантию, обусловленную знанием закона распределения значения. Если гарантий нормального закона распределения значений на выходе нейросети нет, то, соответственно исчезают и гарантии снижения стойкости только на 2 порядка, в сравнении с начальной оценкой.

Что касается размеров тестовой выборки, то ее десятикратное увеличение до 700 случайных образов приведет к почти десятикратному снижению ошибки оценки показателя степени стойкости к атакам подбора, т. е., при наличии знания о виде закона распределения значения контролируемой величины при измерении показателя стойкости защиты погрешность измерения может быть уменьшена до величины 2,8 % и ниже. При этом каких-либо сверхтребований к размерам баз тестовых образов не возникает.

1.5. Оценка погрешности вычислений вероятностей ошибок при тестовой выборке нулевого размера

Весьма интересным является то, что знание закона распределения позволяет вообще отказаться от тестирования нейросетевых решений, косвенно извлекая информацию о вероятностях ошибок первого и второго рода, а так же информацию о погрешностях оценок вероятностей ошибок, при обучении нейронной сети. Технология извлечения этой информации сводится к многократному обучению нейросети на последовательно увеличиваемой выборке примеров образов «Свой».

При монотонном увеличении числа примеров обучения в обучающей выборке возникает так называемый «тренд переобучения» или монотонного снижения качества обучения. При малом числе обучающих примеров прогнозируемое качество обучения высоко, но оно не подтверждается тестированием. С ростом числа примеров обучения качество принимаемых нейронной сетью решений падает, однако при этом оно начинает хорошо подтверждаться на независимых тестах. На рисунке 1.8 приведен типичный «тренд переобучения» нейрона.

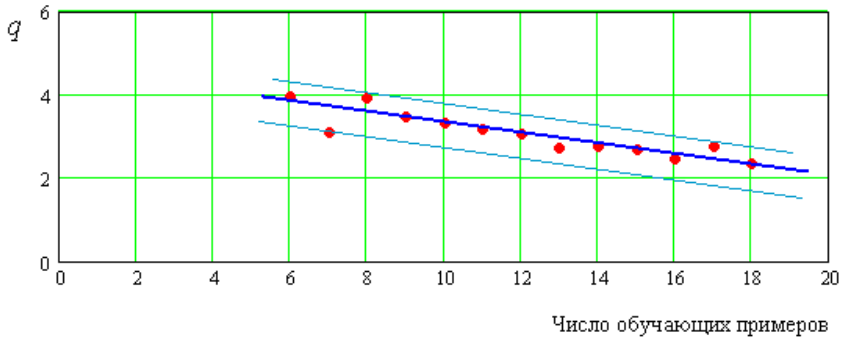


Рисунок 1.8 – Типичный тренд снижения качества обучения при росте числа примеров обучения

Из рисунка 1.8 видно, что при обучении нейрона на 6 примерах $q = 4$, однако если увеличить число примеров обучения до 18, то значение логарифмического показателя качества падает до 2,2. Параллельно «тренду переобучения» построен коридор ожидаемых вариаций показателя качества (тонкие линии на рисунке 1.8). Интуитивно понятно, что коридор вариаций показателя качества обучения будет соответствовать погрешности вычислений этого показателя или точ-

ности прогноза. Получается, что нам удается оценить погрешность прогноза вообще без привлечения дополнительных тестовых примеров (привлекается тестовая выборка нулевого размера).

Естественно, что подобный способ оценки ошибок вычислений является приближенным, однако он крайне интересен. На практике очень часто возникают ситуации, когда примеров для обучения нейросети крайне мало. В этих ситуациях дробить обучающую выборку пополам нерационально. Рационально увеличивать как можно больше обучающую выборку, доводя, если это необходимо, тестовую группу примеров до нулевой размерности.

1.6. Проверка гипотезы нормальности распределения откликов нейросети на образы «Свой» и «Чужие»

Отметим, что все приведенные выше выкладки опираются на гипотезу нормальности законов распределения откликов на линейном выходе последнего нейрона нейросети. В том случае, если мы впервые ищем нейросетевое решение некоторой задачи с использованием некоторого алгоритма обучения, то проверить гипотезу нормальности выходных распределений «Свой» и «Чужой» можно только путем экспериментов. Для этой цели необходимо иметь достаточно много независимых тестовых примеров проверяемых множеств «Свой» и «Чужой». Если мы имеем возможность проверять бесконечно большое число независимых примеров «Свой» и «Чужой», то мы имеем возможность проверить справедливость гипотезы нормальности с как угодно малой погрешностью или с как угодно высокой вероятностью достоверности.

Одним из наиболее эффективных способов проверки гипотезы является вычисление χ^2 – суммарной взвешенной среднеквадратической ошибки тестируемой плотности – p_t и нормальной плотности распределения значений – $p_н$

$$\chi^2 = N \sum_{i=1}^{k} \frac{(p_{t,i} - p_{н,i})^2}{p_{н,i}}, \qquad (1.3)$$

где N – общее число тестов; k – общее число интервалов тестируемой и образцовой нормальной гистограммы; p_t – экспериментально полу-

ченная относительная частота появления отсчетов в i-м интервале гистограммы; $p_н$ – расчетная относительная частота появления отсчетов в i-м интервале идеальной нормальной гистограммы.

В качестве параметров идеальной нормальной плотности – $p_н$ используют экспериментально вычисленные математическое ожидание – m_t

$$\sum_{i=1}^{k} x_i \; p_{t,i} = m_t \; , \tag{1.4}$$

и среднеквадратическую ошибку – σ_t

$$\sum_{i=1}^{k} (x_i - m_t)^2 \; p_{t,i} = \sigma_t^2 \; , \tag{1.5}$$

где x_i – середина i-го интервала гистограммы.

В том случае, если тестируемая плотность распределения значений действительно является нормальной, то $\chi^2 \to 0$ при $n \to \infty$, т. е. для доказательства нормальности тестируемого закона распределения значений достаточно убедиться в наличии асимптотической связи $\chi^2 \to 0$ при $n \to \infty$. Естественно, что проверку следует осуществлять в пределах доступных размеров тестовой выборки.

Крайне важным является то, что распределение значений критерия – χ^2 Пирсона [18, 19] хорошо изучено. По экспериментально вычисленному значению χ^2_t и выбранной при тестировании степени свободы $r = (k - 1)$ через стандартную таблицу можно найти вероятность – $P(\chi^2)$ того, что тестируемый закон распределения значений действительно совпадает с идеальным нормальным законом распределения значений.

Очевидно, что если по гипотезе нормальности выходных законов распределения значений «Свой», «Чужой» получается прогноз вероятности ошибок $P_{EE} = P_1 = P_2 \approx 0{,}05$, то гипотеза нормальности по критерию χ^2 должна выполняться с вероятностью существенно выше 0,95. Желателен, как минимум, трехкратный, а лучше десятикратный запас достоверности

$$P(\chi^2) \geq \left(1 - \frac{P_{EE}}{10}\right). \tag{1.6}$$

Тестирование на нормальность биометрических нейросетевых механизмов требует достаточно высокой квалификации экспериментаторов и значительных размеров баз независимых тестовых биометрических образов. В ряде случаев подобное тестирование непосильно для малых фирм. В связи с этим желательно проводить подобное тестирование специализированными тестовыми лабораториями, обладающими кадрами подтвержденной высокой квалификации и заранее подготовленными большими базами независимых биометрических образов. Чем выше требования к справедливости гипотезы нормальности, тем больше затрат времени и иных ресурсов требует задача проверки этой гипотезы с высоким уровнем достоверности. Чем выше заявляемая производителем достоверность принимаемых его нейросетевым механизмом решений, тем труднее доказать справедливость гипотезы нормальности.

Применительно к средствам высоконадежной биометрической аутентификации [11, 28] объем работ, связанный с проверкой гипотезы нормальности, становится слишком большим для ресурсов малых и средних фирм. Крайне важно объединять усилия по тестированию, например, иметь государственные тестовые лаборатории, осуществляющие независимые тестирования по приемлемым расценкам. Вместо того, чтобы растрачивать усилия на формирование множества больших и сверхбольших частных баз биометрических тестовых образов, желательно иметь одну общую большую базу тестовых биометрических образов. Малые и средние производители должны иметь доступ к такой тестовой базе прямо или через посредника – независимую тестовую лабораторию, владеющую большой общей базой.

1.7. Номограммы вероятностей ошибок вычисления моментов нормального закона распределения значений

Даже в том случае, если мы точно знаем, что законы распределения «Свой», «Чужой» имеют нормальное распределение значений, мы все равно не гарантированы от ошибок, связанных с конечностью тестовой выборки [29]. Очевидно, что чем больше тестовая выборка, тем меньше модуль вычисления погрешности математического ожи-

дания, или при $n \to \infty$ $|\Delta m_t| \to 0$. Пользуясь этим, можно решать и обратную задачу. Если известно, что тестовые воздействия на нейросеть статистически не зависимы и число их известно, то, пользуясь распределением Стьюдента [18, 19], мы можем оценить значение погрешности вычисления математического ожидания, предварительно задавшись доверительной вероятностью.

Аналогом таблиц Стьюдента является приведенная ниже на рисунке 1.9 номограмма.

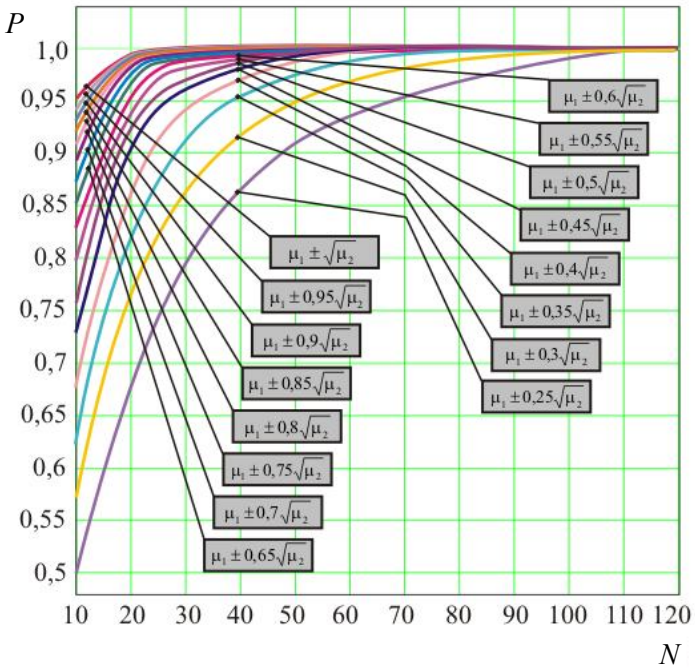


Рисунок 1.9 – Номограмма связи ошибки вычисления первого момента $\mu_1 = m_t$ с расчетным значением второго статистического момента $\mu_2 = \sigma_t^2$ через заданное значение доверительной вероятности – P

Из номограммы рисунка 1.9 видно, что уже при 30 примерах ошибка вычисления математического ожидания попадает в интервал $\pm\sigma_t$ с вероятностью, близкой к единице. Это обусловлено тем, что ошибка вычисления математического ожидания независимых отсчетов пропорциональна σ_t и падает пропорционально $1/\sqrt{N}$. Для

$N = 30\sqrt{30} \approx 5{,}47$. Задавая интервал $\pm\sigma_t$ при $N = 30$, мы с высокой вероятностью оказываемся в интервале $\pm 5{,}47\sigma_m$, где σ_m – собственная дисперсия случайной величины, являющейся математическим ожиданием – m_t, полученной усреднением 30 независимых измерений.

Номограмма, приведенная на рисунке 1.9, интересна тем, что показывает явную статистическую связь ошибки вычисления первого момента $\mu_1 = m_t$ со значением второго момента $\mu_2 = \sigma_t^2$. По индукции следует ожидать наличия подобной связи между ошибкой вычисления второго момента (дисперсии) и значением третьего статистического момента – μ_3 (асимметрии). На рисунке 1.10 приведена номограмма, связывающая ошибку вычисления среднеквадратического отклонения $\sqrt{\mu_2} = \sigma$ со значением $\left|\sqrt[3]{\mu_3}\right|$.

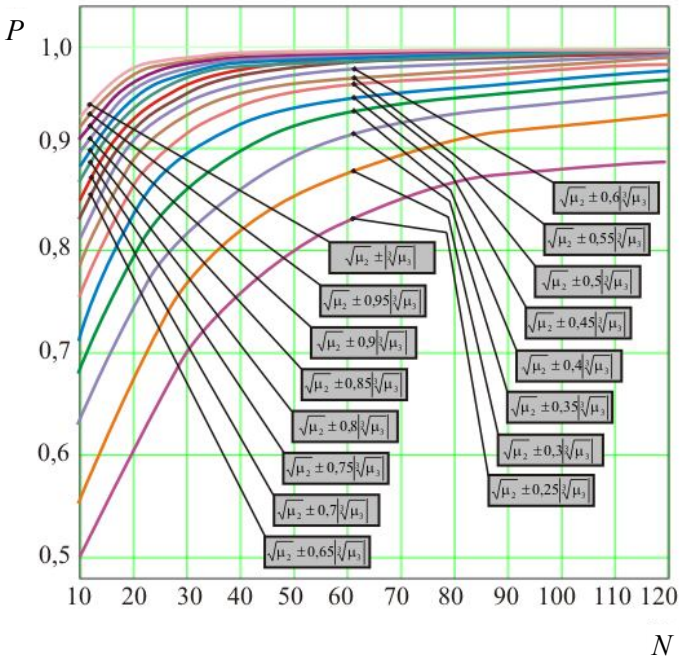


Рисунок 1.10 – Номограмма, связывающая вероятность попадания ошибки вычисления среднеквадратического отклонения $\sqrt{\mu_2} = \sigma$ в интервал, определяемый статистическим моментом третьего порядка – μ_3

Очевидно, что, пользуясь номограммой рисунка 1.10, мы всегда можем оценить ошибку $\Delta\sigma$, задавшись доверительной вероятностью P и зная число независимых тестовых испытаний – N.

В свою очередь, зная модуль ошибки вычисления математического ожидания $|\Delta m|$ и модуль ошибки вычисления среднеквадратического отклонения $|\Delta\sigma|$, мы всегда можем найти ошибку оценки вероятности ΔP в рамках гипотезы нормального закона распределения значений данных на линейном выходе последнего нейрона.

1.8. Идентификация закона распределения как эквивалент формирования измерительного эталона для статистических измерений, осуществляемых при ускоренном тестировании нейросетевых механизмов

Из приведенных выше материалов однозначно следует, что при абсолютно достоверном «знании» закона распределения значений мы всегда можем указать погрешность вычисленных статистических оценок на относительно небольших тестовых выборках. Если мы можем указать погрешность статистических оценок, то мы уже имеем не статистические оценки, а статистические измерения [30, 31]. Проводя аналогию ускоренных статистических испытаний с измерениями можно утверждать, что идентификация закона распределения значений контролируемого параметра – это эквивалент формирования эталонной меры плотности распределения значений контролируемой величины. При обычных измерениях физических величин в измерительном приборе обязательно присутствует эталон измеряемой физической величины (вольта, ампера, секунды,…). При статистических измерениях на неограниченно больших тестовых выборках эталон вероятности в измерительной системе может не быть в явной форме, однако в этом случае высокоточные измерения потребуют огромных затрат времени на формирование достаточно большой тестовой выборки. При этом, как правило, возникают физические, временные, экономические, организационные барьеры, ограничивающие размеры тестовой выборки и тем самым ограничивающие точность статистических измерений.

Попытки обойти эти физические, временные, экономические, организационные барьеры всегда связаны с необходимостью иметь «знание» о контролируемой величине, знание о плотности распределения значений контролируемого параметра. Знание эталона измеряемой величины в форме ее эталона в измерительном приборе и знание закона распределения значений при статистических измерениях – это одинаковые сущности, реализующие экономически оправданные измерения.

Очевидно, что создание измерительного эталона (например, атомного эталона времени) является очень сложной, трудоемкой, но крайне необходимой задачей для всех точных приборов измерения времени. Аналогично высокоточная идентификация закона распределения значений контролируемого параметра в некоторой физической системе – это не что иное, как формирование условий (эталонов) для последующих быстрых измерений. То, насколько точно нам удастся построить эталон, определяет погрешности будущих измерений (обычных и статистических).

В этом плане создание больших и сверхбольших баз независимых биометрических образов и их использование для идентификации закона распределения значений контролируемой величины – это эквивалент формирования статистического эталона. Распространение знаний о законах распределения выходных данных, характерных для биометрико-нейросетевых преобразователей, а также распространение хорошо сбалансированных (представительных) малых баз тестовых образов – это тиражирование эталонов статистических измерений стойкости биометрико-нейросетевых преобразователей.

Так же, как создание эталонов для физических измерений – это государственная функция, создание предпосылок для статистических измерений параметров защиты информации является функцией государства. Государству невыгодно, чтобы его граждан обманывали на рынке через обвешивание. Для противодействия этому создана система сертификации весов, опирающаяся на эталон килограмма. Государству не выгодно, чтобы его граждан обманывали при биометрической аутентификации, и для противодействия этому будет создана система государственной сертификации средств безопасной биометрической аутентификации. В свою очередь, для сертификации высо-

конадежной биометрии именно государству придется создавать статистические эталоны и добывать знания о законах распределения значений контролируемых параметров.

Если подходить к задаче идентификации закона распределения значений контролируемой статистической величины как к элементу формирования высокоточного эталона, то становятся оправданными те огромные затраты, которые могут потребоваться. Более того, затраты на создание высокоточного измерительного эталона не являются разовыми. На протяжении всего жизненного цикла высокоточного измерительного эталона требуются непрерывные усилия на его поддержку в работоспособном состоянии.

Применительно к рассмотренной выше задаче подтверждения гипотезы нормального закона распределения значений формирование статистического эталона есть не что иное, как формирование достаточно большой базы реальных биометрических образов, подтверждающих гипотезу с заданной погрешностью. На каждый конкретный момент времени должно быть известно, что гипотеза нормальности кем-то (например, межведомственной лабораторией тестирования биометрических устройств и технологий при факультете военного обучения Пензенского государственного университета) подтверждена с заданной относительной погрешностью тестирования $\Delta\,\% = 0{,}0000001$.

Это означает, что все малые производители средств биометрической аутентификации, выполнив требования стандарта [11], могут тестировать свои устройства, опираясь на гипотезу нормальности закона распределения значений. Очевидно, что подобное тестирование может быть осуществлено, только если его результат будет хуже, как минимум, на один порядок. Погрешность эталона должна быть на порядок выше погрешности «средства измерения». Однако, как только гипотеза нормальности становится обоснованной, проблемы того, как измерять, как подсчитывать результат, снимаются сами собой в силу того, что все нюансы этого определены соответствующими российскими [32, 33] и международными стандартами.

Проведение предварительных исследований показывает, что подобный подход вполне может быть реализован для относительно

«слабой» биометрии. Можно указать и подтвердить пределы, в которых гипотеза нормальности работает [28].

Много сложнее ситуация складывается для высоконадежных средств биометрической аутентификации. Для них гипотеза нормальности не работает [28], для них необходимо осуществлять идентификацию вида закона распределения и статистически доказывать его адекватность действительности.

Вывод нового закона распределения и экспериментально-расчетное доказательство его адекватности являются весьма и весьма сложной задачей. Для решения этой задачи мало привлечения соответствующих материальных ресурсов, необходимо привлечение не только численных, но и аналитических методов.

Формальное использование аналитики может сводиться к привлечению для целей анализа достаточно большого числа статистических моментов неизвестной плотности распределения значений. Момент k-го порядка аналитически и численно определяется следующим образом:

$$\mu_k = \int_{-\infty}^{+\infty} (x - m_x)^k \; p(x)\, \mathrm{d}x \approx \frac{1}{N-(k-1)} \sum_{i=1}^{N} (x_i - m_x)^k , \qquad (1.7)$$

где m_x – математическое ожидание случайной переменной – x или первый статистический момент – μ_1; $p(x)$ – идентифицируемая плотность распределения значений.

Каждый закон распределения значений $p(x)$ имеет свой уникальный вектор статистических моментов [$\mu_0 = 1$, μ_1, μ_2, μ_3, μ_4, …]. Для того, чтобы уменьшить неопределенность многообразия векторов статистических моментов, стараются исследуемый закон распределения центрировать и нормировать. После центрирования и нормирования вектор контролируемых статистических моментов будет выглядеть следующим образом: [$\mu_0 = 1$, $\mu_1 = 0$, $\mu_2 = 1$, μ_3, μ_4, μ_5,…].

Поясним вышесказанное на примере нормального центрированного и нормированного закона распределения значений. Для нормального центрированного, нормированного закона распределения вектор начальных моментов будет строго определен: [$\mu_0 = 1$, $\mu_1 = 0$, $\mu_2 = 1$, $\mu_3 = 0$, $\mu_4 = 3$, $\mu_5 = 0$, $\mu_6 = 15$, $\mu_7 = 0$, $\mu_8 = 105$,…]. Все нечетные моменты должны быть нулевыми

$$\mu_{(2m+1)} = 0, \quad (1.8)$$

а четные моменты должны геометрически возрастать

$$\mu_{2m} = 1 \cdot 3 \cdot \ldots \cdot (2m - 1), \quad (1.9)$$

где $m = 1, 2, 3, \ldots$.

Очевидно, что аналитико-численное доказательство нормальности идентифицируемого закона распределения значений – $p(x)$ должно сводиться к доказательству асимптотического выполнения соотношений (1.8) и (1.9) при неограниченном росте числа тестовых примеров. При этом, чем выше порядок исследуемого статистического момента, тем медленнее сходятся значения этого момента к своим асимптотам (1.8) и (1.9).

В том случае, если идентифицируемый закон распределения значений – $p(x)$ неизвестен, необходимо найти его аналитическое описание, вычислить аналитически вектор его статистических моментов и численно доказать асимптотическое приближение реально вычисляемых моментов к их аналитическим асимптотам. Очевидно, что подобная работа требует значительных материальных затрат и привлечения коллектива специалистов высокой квалификации.

1.9. Численная оценка ошибки из-за конечного уровня доверия к «знанию» закона распределения значений

Заметим, что абстрактное знание – это абсолютная категория с бесконечным доверием к ней. Абстрактное знание может быть получено через аналитику или путем дедукции из более общих абсолютно достоверных моделей (положений). Абстрактное знание также может быть получено индукцией, но оно становится знанием только после практической проверки. Конкретное численное знание всегда относительно и может быть оценено уровнем доверия к нему (вероятностью ошибки), погрешностью модели, объемом информации, обеспечиваемой моделью конечной точности в конкретных условиях.

Очевидно, что конкретное численное знание закона распределения всегда имеет конечный уровень доверия, конечную погрешность. Например, применительно к ситуации, описанной в параграфе 1.4 (см. рисунок 1.7), абсолютно верное знание нормального закона распределения все же дает погрешность. Эта погрешность может быть оценена в относительных величинах как площадь между двумя нормальными плотностями распределения значений (рисунок 1.11).

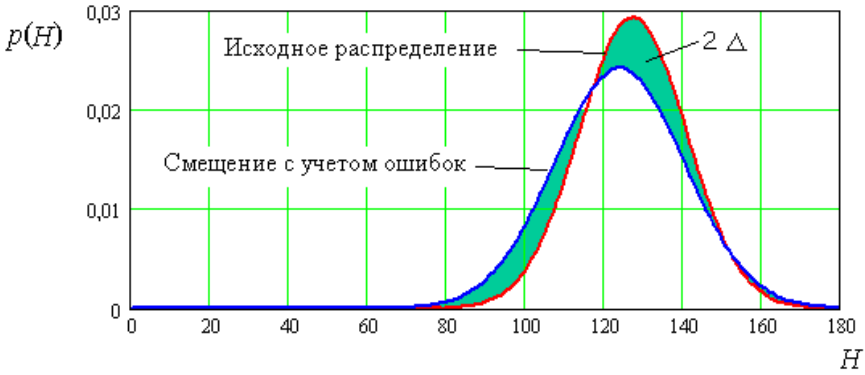


Рисунок 1.11 – Относительная ошибка приближения
двух нормальных законов с ошибочно вычисленными
(статистически измеренными) первым и вторым моментами

Аналитически и численно относительная погрешность знания закона распределения оценивается путем следующих вычислений:

$$\Delta = \frac{1}{2}\int_{-\infty}^{+\infty} |p_t(x) - p_G(x)|\,\mathrm{d}x \approx \frac{1}{2}\sum_{i=1}^{k} |\,p_{i,t}(x_i) - p_{i,G}(x_i)\,|, \qquad (1.10)$$

где $p_t(x)$ – плотность распределения значений, полученная в результате тестирования;

$p_G(x)$ – гипотетическая (проверяемая) плотность распределения значений;

$p_{i,t}(x_i)$, $p_{i,G}(x_i)$ – i-е столбцы тестовой и гипотетической нормированных гистограмм.

При аналитической и численной оценках относительной погрешности знания о законе распределения (1.10) эта оценка может прини-

мать значения в интервале от 0 до 1, что соответствует 0 и 100 % относительной ошибки в ее обычной интерпретации. Относительная ошибка в 100 % соответствует полной несостоятельности гипотезы (две плотности распределения значений $p_t(x)$, $p_G(x)$ вообще не перекрываются). При абсолютно точном знании погрешность составляет 0,0 %, две плотности распределения значений $p_t(x)$, $p_G(x)$ полностью совпадают.

Оценка вида (1.10) универсальна и инвариантна к типам сравниваемых законов распределения значений.

Г л а в а 2

Тестирование идеальных высоконадежных биометрико-нейросетевых механизмов с высокой размерностью выходного вектора принимаемых решений

2.1. Увеличение размерности выходного вектора нейросетевых преобразователей биометрия/код

Практика применения искусственных нейронных сетей для распознавания биометрических образов показывает, что попытки использования достаточно больших нейронных сетей с одним выходом не позволяют получить хороших и очень хороших результатов распознавания. Если оставаться в традиционной парадигме экономии размерности выходного вектора нейросетевого решения, то выходной код длиной в один бит будет всегда давать не более чем удовлетворительные решения [1]. Этот тезис для большинства специалистов по нейросетевой обработке информации на сегодняшний день является неочевидным и нуждается в обосновании.

Для доказательства рассмотрим простейший случай однослойной нейронной сети и возможные топологии нейросетевых решений низкой, средней, высокой и сверхвысокой размерности бинарного вектора выходных решений. Примеры однослойных сетей нейронов приведены на рисунке 2.1.

В левой части рисунка 2.1 изображена вырожденная однослойная нейронная сеть с одним нейроном (одним единственным выходом). Всего однослойная сеть нейронов преобразователя биометрия/код имеет 416 входов (данные продукта «Нейрокриптон 1.0»). В том случае, если бы мы имели идеальную машину обучения, мы могли бы использовать один нейрон с 416 входами [34]. При этом логарифмический показатель качества принимаемых решений должен был быть близок к 6, что соответствует вероятностям ошибок первого и второго рода на уровне 0,000000001. Это хороший результат, однако он практически недостижим из-за отсутствия идеальной машины обучения нейронов. Реальные машины обучения работают много хуже

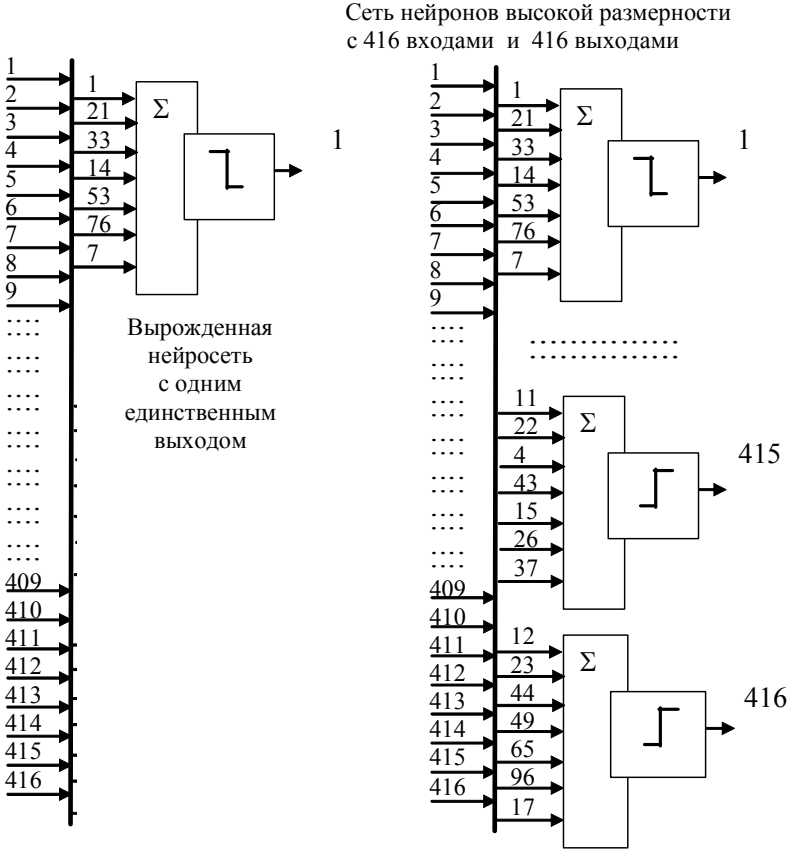


Рисунок 2.1 – Структуры однослойных нейронных сетей с минимальной и средней размерностью бинарного вектора выходных решений

идеальной машины. Как показано на рисунке 2.2, идеальная машина обучения при неограниченном увеличении числа входов у нейрона монотонно увеличивает значение показателя качества обучения, стремясь к некоторой асимптоте обучения, соответствующей предельно возможному качеству обучения одного нейрона на примерах биометрических образов данного типа. К сожалению, все известные на сегодня реальные машины обучения обладают принципиальным дефектом неспособности эффективной борьбы с плохой обусловленностью задачи обучения.

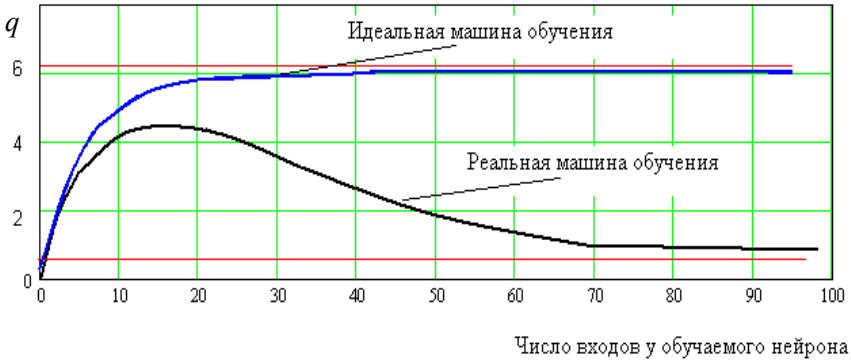


Рисунок 2.2 – Кривая зависимости качества обучения одного нейрона от числа его входов при выделении рукописной буквы «а» из иных символов (среда моделирования «Нейроучитель 1.0», лабораторная работа № 4)

На рисунке 2.2 видно, что реальная машина обучения нейрона достигает наилучшего качества обучения при учете нейроном 16, 17 входных биометрических параметров. При большем числе входов происходит потеря качества обучения, причем эту потерю удается обнаружить только тестированием нейрона после его обучения при монотонно увеличиваемом числе входов, упорядоченных по их качеству.

Из зависимостей, отображенных на рисунке 2.2, следует однозначный вывод о том, что завышение числа входов у нейронов нежелательно. Наоборот, желательно выбирать число входов у нейронов таким образом, чтобы оставаться на участке начального монотонного

подъема функции качества обучения реальной обучающей машиной. В связи с этим обстоятельством выбрано 7 входов у нейронов однослойных нейронных сетей, отображенных на рисунке 2.1.

В соответствии с функцией качества реального обучения рисунка 2.2 при 7 входах среднестатистический нейрон, учитывающий 7 случайно выбранных биометрических параметров, должен принимать решение «Свой»/«Чужой» с показателем качества менее $q \approx 2{,}6$, что соответствует вероятностям ошибок первого и второго рода более 0,005. В итоге мы получаем обычный показатель для относительно слабых биометрических систем – ошибка более чем в 0,5 % случаев. Обычно для относительно слабых систем идентификации человека по рукописному почерку ошибки первого и второго рода составляют от 1 до 5 %.

Заметим, что слабое решение с ошибкой 5 % получается, если мы привлекаем для нейросетевого анализа только 7 биометрических параметров, оставшиеся 409 параметров мы не можем использовать в однослойной сети с одним выходом. Сеть выродилась до одного нейрона, ее структура оказывается практически пустой, что и отображено в левой части рисунка 2.1.

Еще одной неприятной особенностью однослойных сетей является то, что они плохо приспособлены для выделения частного подмножества образов «Свой» из более общего множества «Все чужие». Один нейрон с нечетной функцией возбуждения не способен выделять подмножество, он способен только делить гиперпространство признаков некоторой гиперплоскостью. В случае двухмерного пространства мы получаем ситуацию, отображенную на рисунке 2.3.

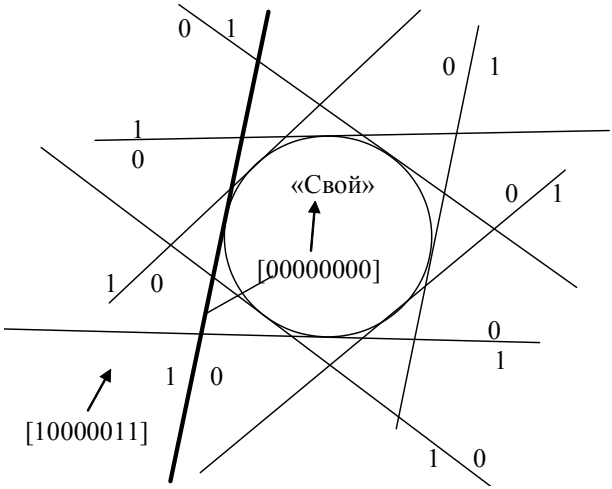


Рисунок 2.3 – Выделение множества «Свой»
однослойной нейронной сетью с несколькими выходами

Из рисунка 2.3 видно, что одновыходовая однослойная сеть из одного нейрона с нечетной пороговой функцией возбуждения способна выделять подмножество «Свой» с вероятностью ошибки, близкой к 0,5 (качество близко к нулевому $q \approx 0{,}0$). Такая вырожденная сеть просто не умеет решать этого типа задачи. Однако если мы используем сеть из множества подобных нейронов (сеть в правой части рисунка 2.1), то положение кардинально изменяется. Достаточно каждый из нейронов настраивать так, чтобы его разделяющая гиперплоскость была касательной к выделяемой гиперсфере «Свой». Получается некоторый бинарный вектор решений на выходе нейросети, дающий очень плохие решения в каждом конкретном разряде выходного кода (вероятность ошибки близка к 0,5), однако совокупное логическое решение на базе этого вектора имеет очень высокое качество. Чем больше размерность выходного вектора, тем выше качество принимаемого далее логического решения.

С помощью одной прямой вообще нельзя описать круг, однако, используя множество касательных, мы можем как угодно точно описать круг (см. рисунок 2.3). Более того, мы можем ввести удобные обозначения разделяемых гиперплоскостью гиперпространств таким образом, что выделяемому подмножеству «Свой» будет соответство-

вать выходной код, состоящий из одних нулей [00000, …, 0]. Появление хотя бы одной из единиц в выходном коде соответствует попаданию в область «Все чужие».

Принципиальным отличием относительно «слабых» нейросетевых решений является крайне низкая размерность вектора выходного нейросетевого решения. Принципиальным отличием высоконадежных нейросетевых решений [1, 11] является искусственное увеличение размерности вектора выходных нейросетевых решений в десятки, сотни, тысячи раз с целью многократно повысить качество конечного логического (криптографического) решения.

2.2. Особые требования к обучению биометрико-нейросетевых преобразователей с большим числом выходов

Реальные нейронные сети преобразователей биометрия–код могут иметь любое число нейронов в слое и любое число слоев в сети. Отказ от прежней парадигмы всемерной экономии размерности выходного вектора нейросетевых решений приводит не только к изменению структуры нейросети, но и к пересмотру отношения к ряду устоявшихся взглядов на обучение нейросети.

Например, классическая догма о том, что при обучении нейронов надо использовать только хорошие (информативные) данные (признаки, параметры) устаревает. В новой парадигме плохих данных практически не существует [1, 20, 22, 23], любая касательная эффективно защищает «Своего» от некоторого множества «Чужих». Для любого параметра можно синтезировать оптимально отсекаемого им «Чужого».

Из рисунка 2.4 видно, что два биометрических параметра v_i и v_j, в которых построена проекция, позволяют хорошо отделять область «Свой» от множеств «Чужой-2», «Чужой-3», «Чужой-4», «Чужой-5», т. е. параметры v_i и v_j высокоинформативны по отношению к перечисленным выше «Чужим». Напротив, по отношению к множеству «Чужой-1» эти параметры v_i и v_j низкоинформативны, однако их нельзя выбрасывать. Они эффективны (информативны) для большого числа «Чужих», т. е. в рамках новой парадигмы нет плохих низкоинфор-

мативных параметров. Любой параметр, даже очень плохой, может оказаться наилучшим для некоторого редко встречающегося случая.

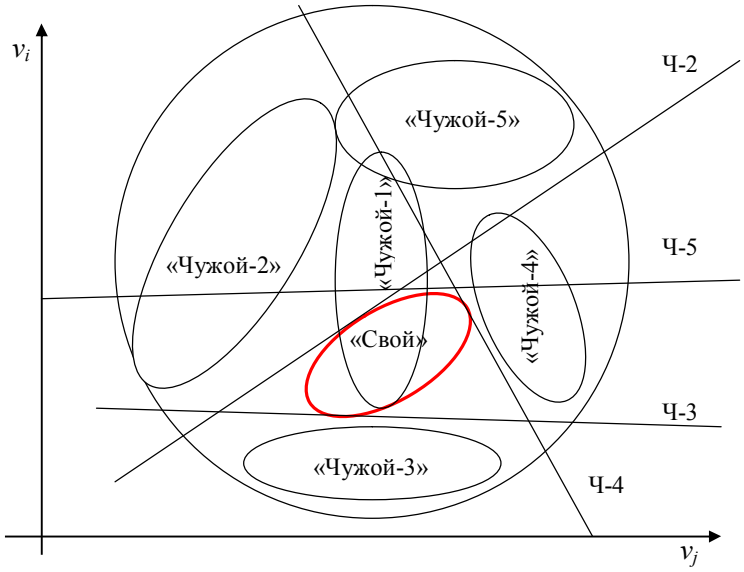


Рисунок 2.4 – Нет плохих данных (любая касательная хорошо защищает от некоторого множества «Чужих-Х»)

В силу всего вышесказанного в новой парадигме искусственного расширения размерности вектора выходных решений нейросети выгодно оказывается не выбрасывать параметры, а использовать их всевозможные случайные сочетания. При формировании нейросети, отображенной в правой части рисунка 2.1, входы нейронов подключены к контролируемым биометрическим параметрам случайно. Так как нейроны этой сети имеют по 7 входов, а контролируется 416 биометрических параметров, можно получить $C^7_{416} \approx 10^{14,6}$ всевозможных сочетаний связей. Это весьма и весьма значительный резерв по возможности увеличения качества принимаемых биометрическими преобразователями решений.

Далее будем различать нейросетевые преобразователи биометрия/код:

1) низкой размерности (число выходов нейросети мало);

2) средней размерности (число выходов нейросети сопоставимо с числом ее входов – сотни выходов);

3) высокой размерности (число выходов превышает число входов нейросети на один или несколько порядков);

4) сверхвысокой размерности, когда число выходов нейросети сопоставимо с теоретически возможным числом сочетаний C^n_N, где n – число входов у нейронов сети; N – число учитываемых биометрических параметров.

По приведенной выше классификации нейросеть в правой части рисунка 2.1 может быть использована преобразователем биометрия/код со средней размерностью вектора выходных нейросетевых решений.

Наряду со взглядами на потенциальную информативность биометрических параметров для многовыходовых сетей нейронов, изменяется и ряд других концептуальных требований к обучению [1, 20, 23, 35]. В проекте национального российского стандарта эти требования сформулированы, опираясь на структурную схему, изображенную на рисунке 2.5.

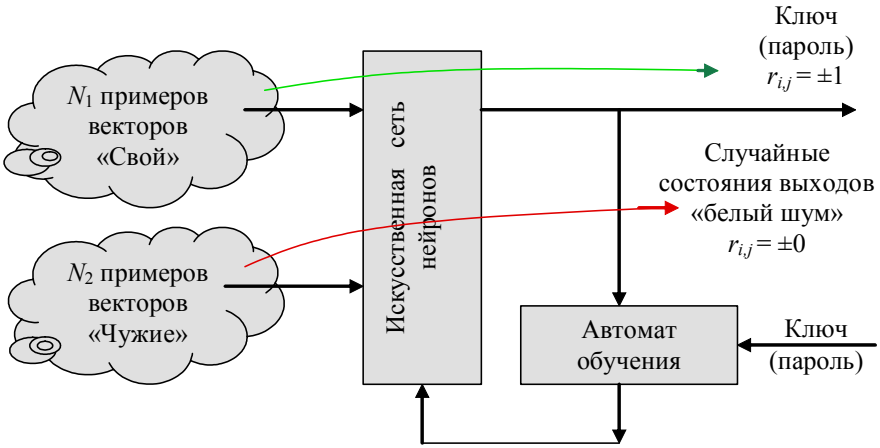


Рисунок 2.5 – Обучение многослойной и многовходовой нейронной сети высоконадежного преобразователя биометрия/код

Так как в стандарте речь идет о безопасном хранении конфиденциальной биометрической и криптографической информации в нейросетевом контейнере (параметрах и структуре нейросети), основным требованием является высоковероятное преобразование обученной нейронной сетью биометрических образов «Свой» в заданный код криптографического ключа. Соответственно при обучении должен быть задан код ключа или длинного пароля.

Обучение нейросети должно вестись автоматически с тем, чтобы с высокой вероятностью, например, с вероятностью 0,95, образы «Свой» преобразовывались в заданный код ключа. Например, если задан код: 001100111, то такой же код должен получаться и при нейросетевой аутентификации. Если коды–отклики на образы «Свой» – будут повторяться с вероятностью, близкой к единице, то корреляция между любой парой разрядов кода будет принимать значения, близкие к «±1». Значение корреляции, близкое к «+1», будет появляться в тех случаях, когда два разряда ключа, попавших в проверяемую пару, будут одинаковыми. Если разряды будут иметь разные значения, то корреляция между ними будет близка к «–1».

Для случайных образов «Чужой» преобразователь биометрия/код должен выдавать случайные выходные коды. Например, коды «Свой» и «Чужие» могут принимать следующие значения:

«Свой» 001100111

«Чужой-1» 010101001

«Чужой-2» 001110010

«Чужой-*N*» 101000111

Каждый разряд выходного кода «Чужой» должен с равной вероятностью принимать значения «0» и «1». Добиться равновероятных значений «0» и «1» в каждом разряде выходного кода несложно. В частности, для однослойных сетей достаточно все разделяющие гиперплоскости проводить через центр множества «Все чужие», как это показано на рисунке 2.6.

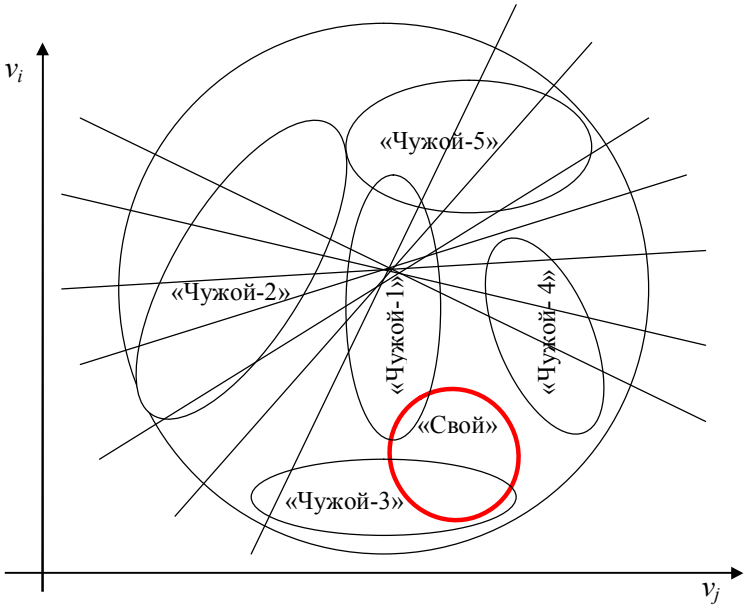


Рисунок 2.6 – Многообразие вариантов гиперплоскостей, обеспечивающих некоррелированность (независимость) выходных кодов «Чужой», не знающих биометрического пароля (биометрического образа «Свой»)

Из рисунка 2.6 видно, что таких гиперплоскостей может быть построено бесконечное множество, и все они располагаются в некотором разрешенном секторе. Рисунок 2.6 сформирован исходя из гипотезы использования при обучении декорреляционных алгоритмов [1, 20–23], которые превращают гиперэллипс «Свой» в гипершар «Свой» и выталкивают его на периферию множества «Все чужие».

В силу того, что каждый из разрядов выходных кодов «Чужие» имеет равновероятные значения «0» и «1», корреляция между любой парой разрядов выходного кода должна быть нулевой, т. е. выходные коды нейросети «Чужой» оказываются случайными, а их разряды попарно независимы (некоррелированы). Далее будем считать преобразователи биометрия/код идеальными, если корреляция между любой парой разрядов i, j кодов «Чужой» отсутствует $r_{i,j} = \pm 0$.

2.3. Связи качества нейросетевых решений с размерностью выходного вектора идеальных преобразователей биометрия/код

То, что у идеального преобразователя биометрия/код выходные коды откликов «Чужой» абсолютно случайны и абсолютно независимы, играет очень важную роль, объясняющую механизм появления нейросетевых решений очень высокого качества, т. е. через существенное увеличение числа выходов нейронной сети при условии их независимости (некоррелированности) можно существенно улучшить вероятностные показатели многомерного нейросетевого решения.

Покажем это, опираясь на меру Хемминга. Мера Хемминга описывает расстояние между сравниваемыми кодами, являясь дискретной величиной. Нам необходима мера Хемминга при сравнении кода «Свой» и кодов «Чужие». Для вычисления меры Хемминга необходимо просуммировать сравниваемые коды по модулю два

$$\text{код}_0 \oplus \text{код}_Y = \text{код}_n, \qquad (2.1)$$

что дает код Хемминга – код_n, в котором единицы соответствуют расхождению разрядов складываемых кодов:

$$\begin{array}{cc} \oplus & \text{код}_0 \\ & \text{код}_Y \\ \hline & \text{код}_n \end{array} \quad = \quad \begin{array}{cc} \oplus & 0000001111111 \\ & 0001001100111 \\ \hline & 0001000011000 \end{array}$$

Суммирование по модулю два – $\oplus$ – дает «1» при одинаковых значениях суммируемых разрядов и «0», если они отличаются:

«0» $\oplus$ «0» = «1»; «0» $\oplus$ «1» = «0»;

«1» $\oplus$ «1» = «1»; «1» $\oplus$ «0» = «0».

Мера Хемминга – это скаляр, суммирующий все единицы в коде Хемминга или подсчитывающий общее число несовпавших бит двух

сравниваемых кодов. Для приведенного выше примера мера Хемминга составляет значение 3 ($H = 3$).

Очевидно, что минимальное значение меры Хемминга $H = 0$ соответствует полному тождеству сравниваемых кодов. Мера Хемминга может принимать максимальное значение, равное длине сравниваемых кодов $H = n$. Это соответствует ситуации, когда сравниваемые коды не совпадают ни в одном разряде.

Еще одним важным моментом является то, что выходные случайные коды нейросетевого преобразователя биометрии описываются биномиальным законом распределения значений. Биномиальное распределение описывается двумя параметрами: длиной кода – n и вероятностью появления в разрядах состояний «0» и «1». На рисунке 2.7 приведены примеры распределений меры Хемминга кодов «Свой» и «Чужие» для разных вероятностей состояний «0» и «1» кодов длиной 256.

Биномиальный закон распределения значений описывает задачу многократного случайного (независимого) бросания монеты. Центральное распределение с математическим ожиданием, равным ровно половине длины кода, соответствует равновероятному выпадению обеих сторон монеты $p = 0{,}5$. Смещение вероятности в пользу одной из ее сторон приводит к смещению закона распределения вправо или влево. Соответствие распределений меры Хемминга выходных кодов биномиальному закону хорошо подтверждается экспериментальными

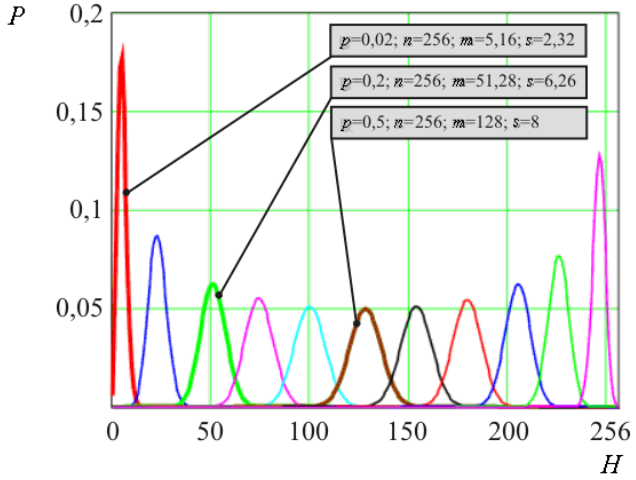


Рисунок 2.7 – Биномиальные распределения расстояний Хемминга между кодом «Свой» и «Чужие» для разных вероятностей состояний разрядов кодов «0» и «1» при длине кода $n = 256$

данными. Точная формула для вычисления вероятности угадывания – H разрядов при n независимых попытках – записывается как

$$P(p,n,H) = C^{H}_{n}(1-p)^{H}\ p^{H-n}\ . \qquad (2.2)$$

Для нас крайне важным является также то, что для идеального биномиального закона распределения известны его моменты. Более того, мы точно знаем значение дисперсии любого из возможных биномиальных распределений

$$\sigma = n\ p\ (1-p). \qquad (2.3)$$

Значение дисперсии зависит от длины выходного кода нейросетевого преобразователя, однако если мы пронормируем меру Хемминга с тем, чтобы она изменялась в интервале от 0 до 1, то получим ситуацию, отображенную на рисунке 2.8. Из рисунка 2.8 видно, что распределение значений образов «Свой» сосредоточено возле нулевого значения меры Хемминга и не зависит от длины выходного кода. Это связано с тем, что коды образов «Свой» очень мало отличаются друг от друга. Практически все они близки к заданному коду и отличаются с вероятностью 0,05 всего в нескольких разрядах. При увеличении длины кода число плохих разрядов пропорционально растет, однако после нормировки этот эффект исчезает.

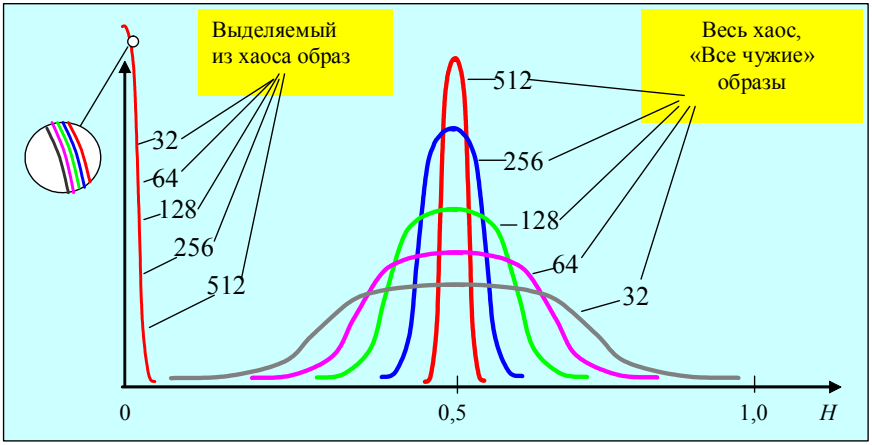


Рисунок 2.8 – Сужение нормального распределения значений нормированной меры Хемминга для выходных кодов идеального преобразователя биометрия/код

При равновероятных состояниях разрядов выходных кодов независимо от их длины биномиальное распределение меры Хемминга практически совпадает с нормальным законом распределения значений. При этом дисперсия центрального нормального закона распределения убывает в $\sqrt{2}$ раз при увеличении в 2 раза длины выходного кода. Для иллюстрации в таблице 2.1 приведены значения дисперсий биномиального дискретного закона распределения для разных длин кодов при значении показателя $p = 0{,}5$.

Таблица 2.1 – Значения дисперсий биномиального дискретного закона распределения для разных длин кодов при значении показателя $p = 0{,}5$

Длина выходного кода	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$	$n = 2048$
Среднеквадратическое отклонение	$\sigma = 4$	$\sigma = 5{,}66$	$\sigma = 8$	$\sigma = 11{,}3$	$\sigma = 16$	$\sigma = 22{,}6$
Среднеквадратическое отклонение нормированной меры Хемминга	$\sigma = 0{,}0625$	$\sigma = 0{,}04414$	$\sigma = 0{,}03125$	$\sigma = 0{,}02207$	$\sigma = 0{,}015625$	$\sigma = 0{,}011035$

Естественно, что данные таблицы 2.1 соответствуют только идеальному преобразователю. Для реальных преобразователей среднеквадратические отклонения оказываются несколько больше, однако эффект сокращения почти в $\sqrt{2}$ раз дисперсии при увеличении в 2 раза длины выходного кода наблюдается. Пользуясь им, мы можем сужать распределение случайных образов «Чужие» до любой наперед заданной величины через необходимое увеличение разрядности выходного кода нейросетевого преобразователя.

Заметим, что все вышесказанное есть не что иное, как «вечный информационный двигатель» или идеальная информационная машина, способная из чего угодно извлекать любой объем информации. Естественно, что это все практически реализовать невозможно, однако первоначальное увеличение числа разрядов биометрико-нейросетевого преобразователя порождает именно такой эффект очень быстрого сужения распределения множества «Чужие». При дальнейшем увеличении числа разрядов выходных кодов эффект ослабевает, при очень большом числе разрядов добиться их независимости (некоррелированности) оказывается технически невозможно, т. е. «перпетум мобиле» невозможен не только в механике, но и в информатике.

2.4. Отсутствие идеального «белого шума» образов «Чужие» у реальных нейросетевых преобразователей, контроль допустимых пределов неидеальности преобразователей

В связи с тем, что одним из основных признаков идеальности нейросетевого преобразователя является отсутствие коррелированности (зависимости) между выходами нейросети, необходимо осуществлять контроль их коррелированности. Для этой цели надо контролировать парную и групповую корреляцию выходов нейросети при воздействии на нее случайными биометрическими образами «Чужие». При измерениях необходимо использовать не менее [11] 300 случайных образов и проводить измерение не менее чем на 100 случайных парах i, j выходов. Расчет корреляционной связи по каждой из пар осуществляется по традиционной формуле

$$r_{i,j} = \frac{1}{N}\sum_{k=1}^{N}\frac{(m(y_i) - y_{i,k})\,(m(y_j) - y_{j,k})}{\sigma_i\,\sigma_j}, \qquad (2.4)$$

где N – число независимых экспериментов; i, j – номера контролируемых выходных разрядов; y_i, y_j – значения контролируемых разрядов, полученные при тестировании; $m(.)$ – символ операции вычисления математического ожидания; σ_i, σ_j – среднеквадратические отклонения состояний контролируемых разрядов.

Проведенные исследования реальных биометрико-нейросетевых преобразователей с большим числом выходов показали, что наиболее вероятным значением коэффициентов корреляции являются нулевые значения. Тем не менее достаточно часто встречаются сравнительно сильно коррелированные выходные данные. На сегодняшний день считается приемлемым иметь среднее значение модулей коэффициентов корреляции менее 0,15.

Проблема неидеальности нейросетевых преобразователей пока не имеет однозначного решения. Пока не ясно, какова причина появления сравнительно сильных корреляционных связей. Вполне возможно, что это следствие присутствия в рукописном почерке некоторых слабых корреляционных связей. Мы все пишем слева направо (кириллическая рукописная традиция), кроме того, некоторые знаки чаще встречаются при воспроизведении слов русского и любого иного языка. Это означает, что примеры кривых колебаний пера по оси Y при воспроизведении случайно выбранных слов, отображенные на рисунке 2.9, нельзя считать «белым шумом».

Второй причиной могут являться случайно появляющиеся корреляционные связи при обучении нейронов одного слоя. Даже если используются декорреляционные алгоритмы обучения [1, 23], они работают только при обучении одного нейрона. Если алгоритм устранения корреляции одинаков и применяется к разным нейронам, то существует вероятность усиления им корреляционных связей нейронов в одном слое. Вероятность такого события растет в случае, если нейроны одного слоя обрабатывают частично перекрывающиеся (частично общие) биометрические параметры. Избежать частичного перекрытия контролируемых параметров у нейронов технически не-

возможно. Так, при случайном выборе 7 входных связей у каждого

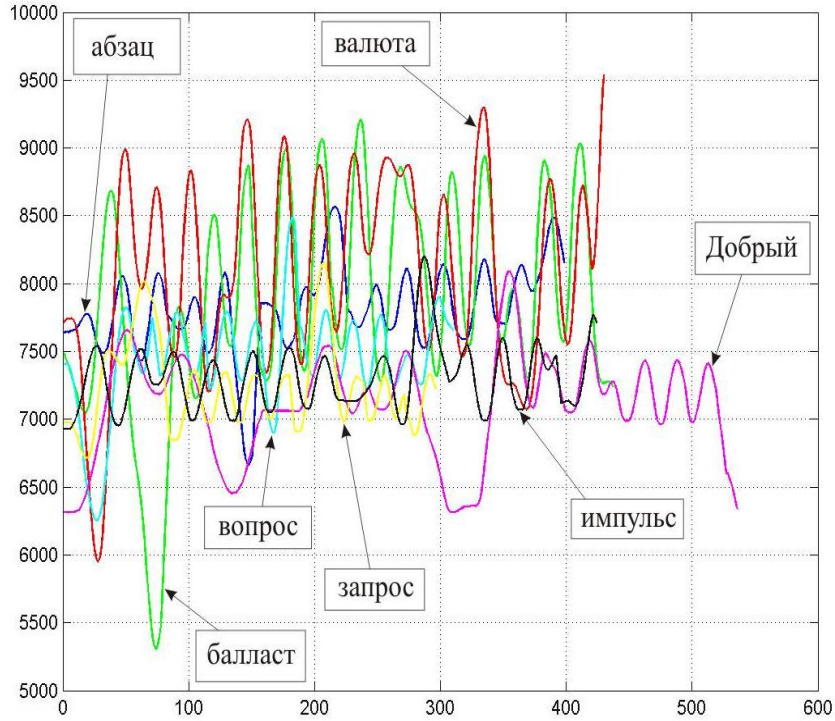


Рисунок 2.9 – Примеры кривых колебаний пера
по оси *Y* при воспроизведении случайно выбранных слов

из 416 нейронов сети нейронов правой части рисунка 2.1 достаточно часто будут встречаться нейроны, имеющие по 1, 2, 3 общим связям. Естественно, что выходы этих нейронов будут иметь повышенный уровень корреляционной связи. При создании хороших биометрико-нейросетевых преобразователей необходимо следить за тем, чтобы ощутимо коррелированных разрядов выходного кода было мало.

Так как мы решаем многомерную задачу преобразования сложных биометрических образов в код контроля корреляции, только пар выходных разрядов кода недостаточно. Теоретически возможна ситуация, когда корреляция будет отсутствовать во всех парах разрядов

кода, но появится в группах из 3, 4, 5, … разрядов. В связи с этим необходимо осуществлять выборочный контроль корреляционных моментов более высокого порядка. Ниже приведена формула для вычисления коэффициента корреляции для группы из 3 разрядов с номерами i, j, e

$$r_{i,j,e} = \frac{1}{N}\sum_{k=1}^{N} \frac{(m(y_i) - y_{i,k})\,(m(y_j) - y_{j,k})\,(m(y_e) - y_{e,k})}{\sigma_i\,\sigma_j\,\sigma_e}. \qquad (2.5)$$

По индукции могут быть получены коэффициенты корреляции для групп, состоящих из большего числа контролируемых параметров.

2.5. Контроль коррелированности кодовых откликов на образы «Свой»

При тестировании нейросетевого преобразователя образами «Свой» обычно стараются оценить вероятность отказа «Своему», при этом в силу дискретного характера тестовых воздействий мы имеем достаточно большую ошибку из-за дискретизации. В частности, при 20 попытках мы получим относительную ошибку из-за дискретизации на уровне 5%. Это достаточно большая ошибка при весьма значительной тестовой выборке, сопоставимой с выборкой обучения.

Очевидно, что для идеального преобразователя корреляция всех разрядов выходного кода «Свой» всегда точно равна ±1. Из-за этого средний модуль корреляции должен быть точно равен 1. С другой стороны, для всех реальных преобразователей биометрия/код усредненный модуль коэффициентов корреляции всегда будет меньше единицы. Можно говорить о топологической эквивалентности вычисления вероятности ошибки первого рода и оценки разницы среднего модуля коэффициентов корреляции и единицы

$$\begin{cases} 1 - P_1 \le 1, \\ 1 - m\left|r_{i,j}\right| \le 1. \end{cases} \qquad (2.6)$$

Очевидно также то, что при прочих равных условиях вычислять средний коэффициент модуля корреляции удается точнее, чем вероятность ошибки первого рода. При вычислении усредненного модуля корреляции получается больше исходной информации. Например,

если нейросетевой преобразователь биометрия/код имеет 256 выходов, то каждый опыт, единичный опыт со своим биометрическим образом, будет эквивалентен 256 зависимым опытам. Оценивая коэффициент корреляции, мы видим, сколько конкретно бит в ключе дали сбой в отвергнутом образе «Свой». Если в одном отвергнутом системой образе оказалось много неисправных бит – это много хуже, чем неисправность только одного бита.

Еще более высокую точность в оценке коэффициентов корреляции можно получить, если перейти от операций с дискретными значениями «0» и «1» к операциям с непрерывными величинами на линейном выходе каждого из нейронов. В качестве примеров на рисунке 2.10 приведены распределения значений множеств «Свой» и «Чужой» на выходах двух разрядов с номерами i, j.

Если учитывать непрерывные распределения значений контролируемых параметров, то их коэффициент корреляции может быть вычислен даже без факта неправильного срабатывания одного из тестовых примеров «Свой».

Вычисления осуществляются исходя из знания прогнозируемого значения вероятности ошибок первого рода в i-м разряде – $P_{1,i}$ и исходя из знания прогнозируемого значения вероятности ошибок первого рода в j-м разряде – $P_{1,j}$. Вычисления осуществляются по следующей формуле:

$$r_{i,j} = (1 - P_{1,i})\,(1 - P_{1,j})\,. \qquad (2.7)$$

В рамках гипотезы нормального распределения значений «Свой» формула (2.7) может быть записана через односторонние интегралы Лапласа и логарифмические показатели качества контролируемых параметров

$$r_{i,j} = (0{,}5 + \Phi_0(2q_i))\,(0{,}5 + \Phi_0(2q_j))\,. \qquad (2.8)$$

Исходя из процедуры вычисления среднего модуля коэффициента корреляции по множеству выбранных пар выходов, мы получим следующую уточненную оценку вероятностей ошибок первого рода:

$$P_1 \approx \sqrt{1 - m\left|r_{i,j}\right|}\,. \qquad (2.9)$$

Это означает, что оценки, входящие в систему неравенств (2.6), связаны между собой квадратичной функцией.

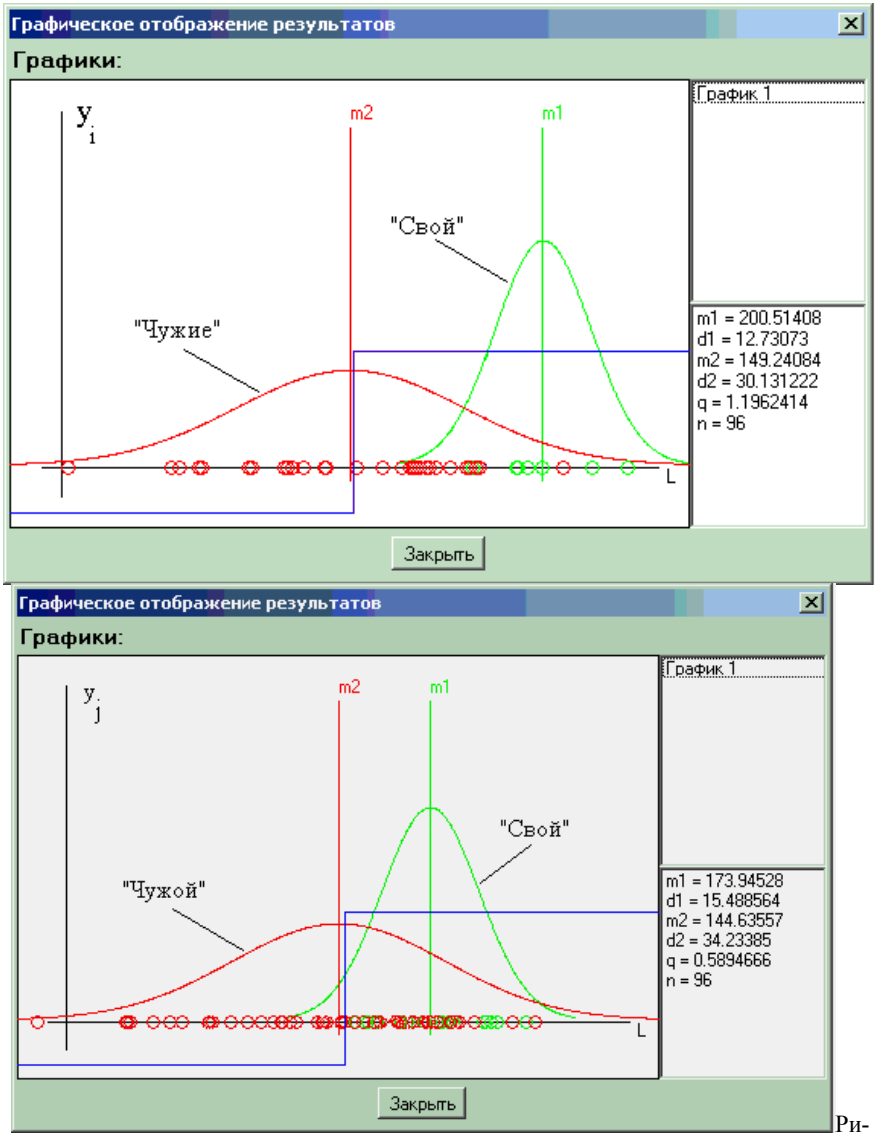


Рисунок 2.10 – Два непрерывных распределения на линейных выходах последних нейронов с номерами *i*, *j*

2.6. Контроль равновероятности состояний разрядов выходных кодов преобразователей

Для преобразователей биометрия/код одним из наиболее важных параметров является равновероятность состояний «0» и «1» каждого из выходов нейросети при воздействии на нее случайными образами «Чужой». Формальным алгоритмом оценки этого параметра является статистический контроль состояний каждого разряда выходных кодов. Для определенности предположим, что испытания дают серию из 60 следующих состояний в первом разряде выходного кода:

001101010101010000101111110101010101010001010110101111111 0,

в этой серии состояний «0» –27, состояний «1» –33. Очевидно, что состояния «1» встречаются чаще, чем состояния «0», однако такое различие вполне допустимо для выборки из 60 примеров. При необходимости увеличить точность оценки следует увеличить число независимых тестовых воздействий.

Еще одним путем повышения точности оценки является контроль параметров распределения «Все чужие» на линейном выходе нейронов (до пороговой функции возбуждения). Примеры таких распределений даны на рисунке 2.10. Быстрые алгоритмы обучения биометрико-нейросетевых преобразователей определяют математическое ожидание множества обучающих образов «Все чужие» и именно в него ставят точку переключения пороговой функции. При тестировании на другой выборке примеров «Чужие» всегда получается другое значение математического ожидания. Требуется оценить, насколько допустимо появившееся расхождение. Эта задача легко решается в рамках гипотезы нормального закона распределения значений множества «Все чужие».

В этом случае вероятность того, что на выходе контролируемого нейрона состояния «0» и «1» будут действительно равновероятны, оценивается как

$$P = 0{,}5 + \Phi_0\left[2\frac{|m - m_t|}{\sigma + \sigma_t}\right], \qquad (2.10)$$

где m, m_t – математические ожидания обучающей и тестовой выборки; σ, σ_t – среднеквадратические отклонения обучающей и тестовой выборки.

2.7. Косвенный контроль равновероятности кодовых состояний через дефекты балансировки преобразователей биометрия/код

Следует подчеркнуть, что прямой контроль равновероятных состояний «0» и «1» дает эффективную дифференциальную оценку сбалансированности каждого из выходных разрядов нейросетевого преобразователя по множеству «Все чужие». Естественно, что такая оценка будет правдива только при условии представительности обучающей выборки «Все чужие» и тестовой выборки примеров «Все чужие» по каждому из контролируемых разрядов выходного кода. Гарантировать действительную сбалансированность тестовых выборок по всем данным и всем разрядам обученной нейросети никто не может. Более того, вполне резонно предположение, что на выборе порядка 100,…, 300 примеров «Все чужие» обязательно должны присутствовать плохо представленные биометрические параметры (плохо сбалансированные выходы нейросети). Очевидно, что много легче сбалансировать в ограниченной выборке данные в среднем, чем по отдельности каждый параметр (каждый выход).

В связи с вышеизложенным интересной является не дифференциальная, а интегральная оценка выполнения условия равновероятности состояний выходных кодов нейросетевого преобразователя. Такую оценку можно осуществить, контролируя математическое ожидание меры Хемминга отклонения кодов тестового множества «Все чужие» от кода ключа «Свой». Если все выходы кода имеют равновероятное состояние «0» и «1», то математическое ожидание меры Хемминга будет точно равно половине ее максимального значения. Верно и обратное утверждение: уменьшение математического ожидания меры Хемминга множества «Все чужие», по сравнению с половиной ее максимального значения, свидетельствует об отсутствии правильной балансировки состояний выходных кодов. На рисунке 2.11 приведены графики, иллюстрирующие это положение.

Тонкие графики, отображенные на рисунке 2.11, получены отключением входного центрирования контролируемых биометрических параметров действующего макета преобразователя биомет-

рия/код. Если входное центрирование биометрических параметров

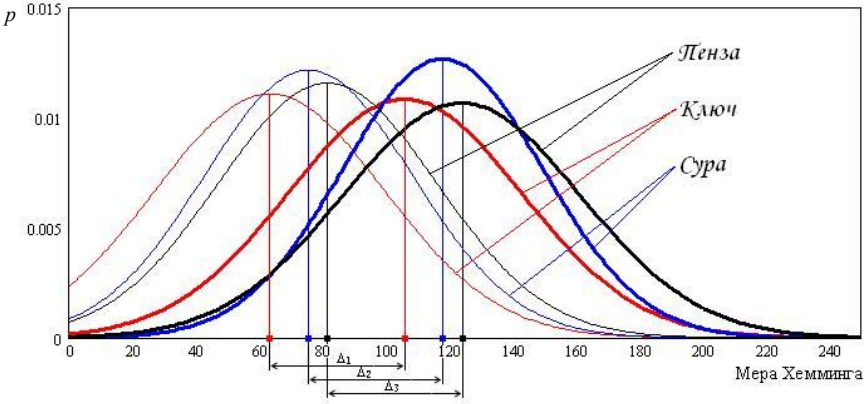


Рисунок 2.11 – Примеры распределений значений меры Хемминга при хорошей балансировке (толстые линии) и плохой балансировке (тонкие линии)

«Все чужие» хорошо работает, то проблема выходной балансировки состояний кодов ослабляется. При идеальном центрировании всех входных биометрических параметров проблема балансировки вообще должна исчезнуть, так как постоянная составляющая на выходах нейронов вообще не может появиться при отсутствии у них смещения. Тогда при любых значениях весовых коэффициентов выходная балансировка состояний кодов сохраняется.

Из рисунка 2.11 видно то, насколько важно входное центрирование биометрических параметров для декорреляционных алгоритмов обучения. Еще одним важным выводом зависимостей, изображенных на рисунке 2.11, является то, что выходная балансировка кодов зависит не только от представительности группы тестовых примеров образов «Все чужие», но и от примеров образов «Свой». Для разных рукописных образов «*Пенза*», «*Ключ*», «*Сура*» получаются разные распределения значений меры Хемминга. Это означает, что при тестировании биометрико-нейросетевого преобразователя мало иметь представительную выборку образов «Все чужие». Тестирование на сбалансированность по состояниям «0» и «1» следует проводить для достаточно большого числа примеров образов «Свой». Прежде чем делать вывод о некорректной балансировке нейросетевого преобра-

зователя по входам и выходам, необходимо убедиться в том, что наблюдаемый небаланс не является некоторой особенностью конкретной обучающей выборки образов «Свой». Возможно, что добавление в обучающую выборку одного, двух дополнительных примеров позволит существенно улучшить балансировку состояний выходов нейросети.

Если улучшить баланс не удается, то по его значению легко вычислить ошибочный небаланс через обращение соответствующего биномиального распределения значений. Пример таблицы такого обращения приведен ниже для преобразователя биометрия/код с 256 выходами (таблица 2.2).

Таблица 2.2 – Связь математического ожидания меры Хемминга для множества образов «Все чужие» с вероятностями появления состояний «0» и «1» в выходных разрядах преобразователя биометрия/код с 256 выходами

$m(H)$	0,0	25,6	51,2	76,8	102,4	128	153,6	179,2	204,8	230,4	256
P	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
$m(H\,\%)$	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0

В последней строке таблицы 2.2 приводится математическое ожидание нормированной меры Хемминга. Нетрудно заметить, что вероятностная мера во второй строке таблицы и математическое ожидание нормированной меры Хемминга полностью совпадают (они тождественны как для идеальных, так и для неидеальных преобразователей биометрия/код).

2.8. Появление структурной корреляции при неоправданном увеличении размерности нейросетевых машин добычи и обогащения данных

Если исходить из тезиса о том, что чем больше выходов у нейронной сети, тем выше качество принимаемого ею решения, то следует стремиться делать как можно больше выходов у нейросетевых преобразователей. В общем случае, когда мы имеем N входов у однослойной сети и нам задано число входов у ее нейронов – k, мы можем получить

$$C_N^k = \frac{N!}{k!(N-k)!} \tag{2.11}$$

возможных несовпадающих сочетаний соединений входов нейронов с входными данными.

От заданного числа входов у нейронов существенно зависит число возможных сочетаний, соответствующий график этой зависимости приведен на рисунке 2.12.

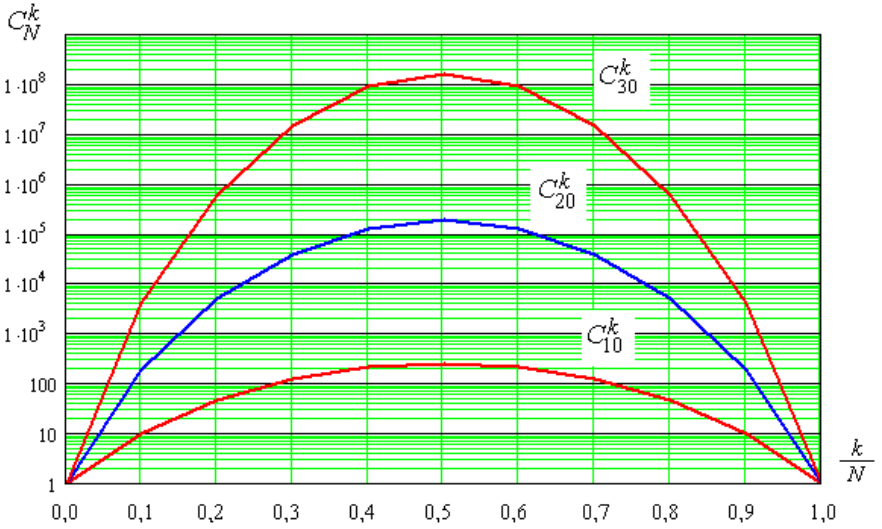


Рисунок 2.12 – Кривая зависимости числа сочетаний C^k_N для $N = 10, 20, 30$ как функция отношения k/N

Из рисунка 2.12 видно, что наибольшее число выходов у нейросети может быть получено, если использовать нейроны с числом входов $k = N/2$. В итоге, ориентируясь на максимальное число выходов нейросетей, мы будем получать при большом числе анализируемых биометрических данных астрономические величины. Примеры таких величин приведены в таблице 2.3.

Таблица 2.3 – Возможное число выходов у нейросетевого преобразователя биометрия/код

Число входов сети	10	50	100	150	200	256	300	350	416
Число входов нейрона	5	25	50	75	100	128	150	175	208

Число выходов сети	$10^{2,4}$	$10^{14,1}$	$10^{29,0}$	$10^{43,96}$	$10^{58,96}$	$10^{75,76}$	$10^{88,97}$	$10^{103,9}$	$10^{123,8}$

Из таблицы 2.3 видно, что число выходов однослойной нейронной сети, обрабатывающей 416 биометрических параметров, может быть очень велико, до $10^{123,8}$, при использовании нейронов с 208 входами. К сожалению, подавляющее большинство из $10^{123,8}$ выходов нейросети будет очень сильно коррелировано, т. е. нарушается требование независимости выходов нейросети, что ставит под сомнение саму идею существенного повышения качества принимаемых решений за счет существенного увеличения числа независимых выходов.

Высокая корреляционная зависимость обусловлена выбором большого числа входов у нейронов. Чем меньшую долю от общего числа входов занимают входы одного нейрона, тем меньше они будут перекрываться с соседним нейроном. Для иллюстрации этого положения рассмотрим три однослойные нейронные сети, нейроны которых соответственно занимают 10, 50 и 90 % от всех входов сети. Фрагменты таких сетей отображены на рисунке 2.13.

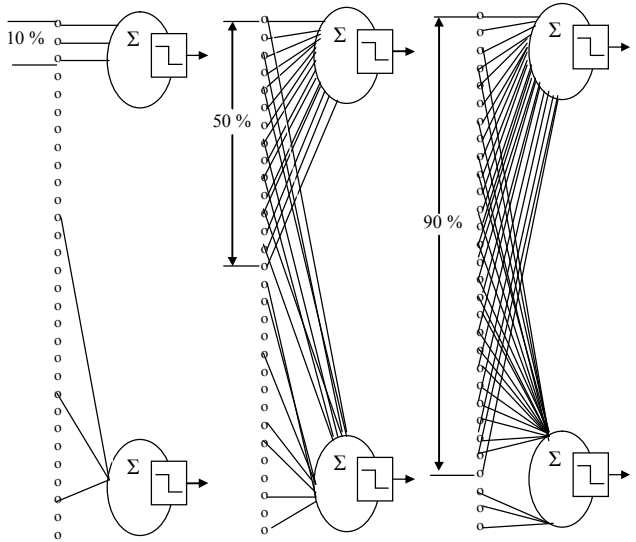


Рисунок 2.13 – Три варианта однослойных сетей с нейронами, имеющими по 10, 50, 90 % входов всей сети

Левая сеть рисунка 2.13 имеет 10 % входов, например, 40 входов из 400. Из-за малого числа входов перекрытие их для двух нейронов маловероятно. Если верхний нейрон занимает 10 % первых входов, а второй нейрон подключается к случайным входам, то вероятность перекрытия их входов составит 0,1. В этой ситуации наиболее вероятное значение модуля коэффициента корреляции выходных данных верхнего и нижнего нейронов также составит 0,1.

Центральная сеть рисунка 2.13 имеет 50 % входов, например, 200 входов из 400. Из-за такого числа входов перекрытие их для двух нейронов (верхнего и нижнего) часто возникает. Если верхний нейрон занимает 50 % первых входов, а второй нейрон подключается к случайным входам, то вероятность перекрытия их входов составит 0,5. В этой ситуации наиболее вероятное значение модуля коэффициента корреляции выходных данных верхнего и нижнего нейронов также составит 0,5.

Правая сеть рисунка 2.13 имеет 90 % входов, например, 360 входов из 400. Из-за такого большого числа входов перекрытие их для двух нейронов (верхнего и нижнего) возникает очень часто. Если верхний нейрон занимает 90 % первых входов, а второй нейрон подключается к случайным входам, то вероятность перекрытия их входов составит 0,9. В этой ситуации наиболее вероятное значение модуля коэффициента корреляции выходных данных верхнего и нижнего нейронов также составит 0,9.

Верхний и нижний нейроны всех трех сетей обучаются по одному и тому же алгоритму преобразовывать входные данные в свой бит выходного кода. Модуль коэффициента корреляции выходов верхнего и нижнего нейронов является случайной величиной, но очень сильно зависит от числа входов у нейронов, общего числа входов у всей сети. Кроме того, важную роль играет их процентное соотношение. На рисунке 2.14 приведены плотности распределения значений модулей коэффициентов корреляции для разного числа входов у сетей и у нейронов.

Из рисунка 2.14 видно, что гипотеза независимости разрядов выходных кодов быстро разрушается по мере увеличения числа входов у нейронов. При неоправданном увеличении числа входов

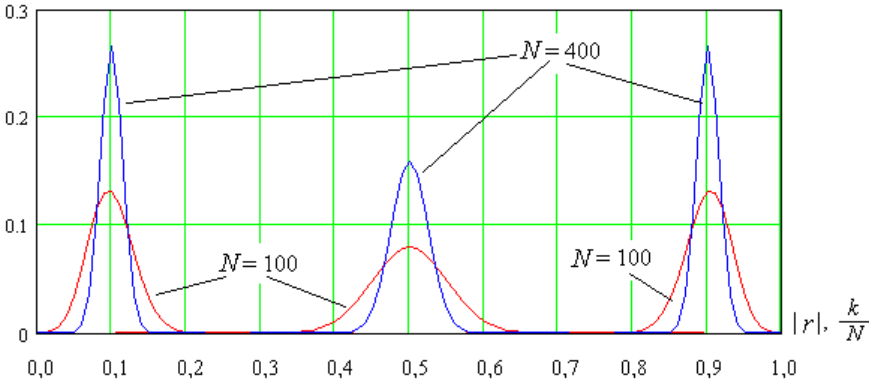


Рисунок 2.14 – Плотности распределения значений модулей коэффициентов корреляции для разного числа входов у сетей и у нейронов

у нейронов возникает так называемая структурная корреляция выходных данных. Чем больше у соседних нейронов общих входных данных, тем выше структурная корреляция. В пределе мы имеем так называемые полносвязные сети, которые являются наихудшим вариантом из всех возможных вариантов. Практика показывает, что желательно иметь нейроны со слабым структурным перекрытием и ограничивать среднее значение модуля коэффициентов корреляции величиной 0,15.

Только в самом начале искусственного увеличения выходной размерности нейронной сети эффект роста качества близок к теоретически возможному. При некотором неоправданном увеличении числа выходов сети (числа нейронов сети), числа входов у нейронов рост выходного качества принимаемых нейронной сетью решений падает. Начиная с некоторого момента, происходит насыщение кривой роста качества, далее увеличивать сложность обработки биометрических образов некоторой информативности становится бессмысленным.

Г л а в а 3

Тестирование биометрико-нейросетевых механизмов с высокой размерностью и зависимыми данными

3.1. Проблема аналитико-численного описания основных законов распределения значений с учетом корреляционной зависимости данных

В предыдущей главе было показано, что у идеальных преобразователей биометрия/код корреляция (парная и многомерная) выходных кодов «Все чужие» должна быть нулевой. Однако этого никогда не удается достичь для реальных преобразователей биометрия/код. Реальные коэффициенты корреляции имеют вполне ощутимые значения. В качестве примера на рисунке 3.1 дана типичная гистограмма значений коэффициентов парной корреляции выходных кодов «Все чужие» преобразователя «Нейрокриптон», предварительно обученного на рукописном образе слова «*Пенза*».

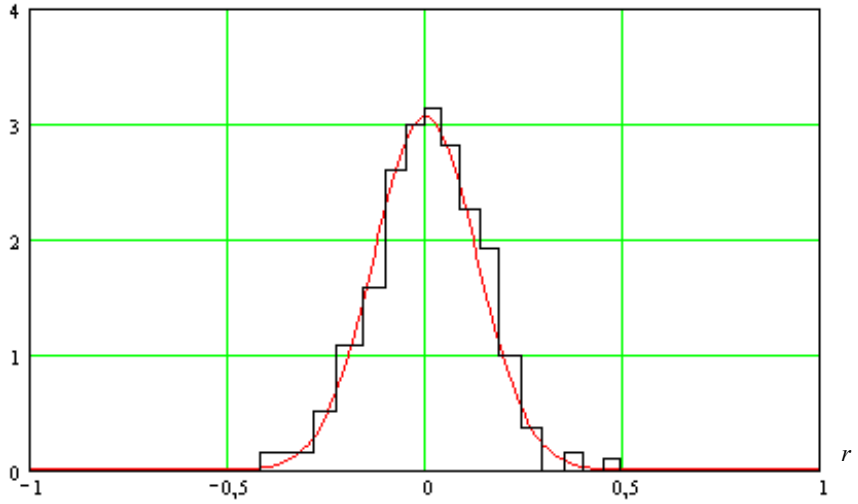


Рисунок 3.1 – Типичная гистограмма значений коэффициентов парной корреляции выходных кодов «Все чужие» преобразователя «Нейрокриптон», предварительно обученного на рукописном образе слова «*Пенза*»

Кроме того, в параграфе 2.7 было показано, что с ростом числа входов у нейронов модули значений коэффициентов парной корреляции должны увеличиваться, т. е. при попытках существенного увеличения числа выходов нейросети (при числе входов нейронов, близком к половине входов всей сети) мы должны с высокой вероятностью получать значение парной корреляции, близкое к $r \approx \pm 0{,}5$ (плотность распределения раздваивается и имеет две дополнительные моды для значения $r \approx -0{,}5$ и для значения $r \approx +0{,}5$).

Для классического биномиального закона распределения мы с ростом числа опытов должны иметь нулевое значение парной корреляции выходных кодов для равновероятных состояний кодов «0» и «1» (параметр $p = 0{,}5$). Мы же на практике имеем существенное отличие от нуля значений коэффициентов корреляции (см. рисунок 3.1). Это означает, что многомерные нейросетевые преобразователи биометрия/код относятся к новому классу статистических объектов, которые должны описываться некоторой модификацией классического биномиального закона распределения значений меры Хемминга. Для терминологического выделения этой модификации далее будем ее называть биномиальным зависимым законом распределения как альтернативу классического биномиального независимого законного распределения (независимого в точке параметра $p = 0{,}5$).

В связи с тем, что нам необходимы высокоуровневые гарантии стойкости биометрико-нейросетевых преобразователей к атакам подбора, мы должны иметь высокоточное (численное или аналитическое) описание нового биномиального зависимого закона распределения значений, т. е. необходимо проделать значительную работу по аналитическому или численному статистическому описанию нового неклассического объекта – биометрико-нейросетевых преобразователей.

Кроме решения задачи идентификации нового неклассического закона распределения, нам необходимо будет доказать его эквивалентность реальному закону распределения выходных кодов преобразователей биометрия/код. При этом возникает еще одна трудность. Дело в том, что классическое распределение χ^2 выведено для независимых опытов (так же, как и классический биномиальный закон).

Задачей второго уровня является корректировка таблиц классического дискретного закона распределения χ^2 для независимых отсчетов под реальную ситуацию проверки существенно зависимых данных.

3.2. Моделирование зависимого биномиального закона распределения значений кодов на выходе многомерных нейросетевых преобразователей

В связи с тем, что мы имеем новый объект статистического исследования – нейросетевые преобразователи биометрия/код с существенно зависимыми данными, мы должны построить (численно или аналитически) модификацию классического биномиального закона распределения значений. Наиболее просто эта задача решается численно. Структурная схема организации соответствующего численного эксперимента приведена на рисунке 3.2.

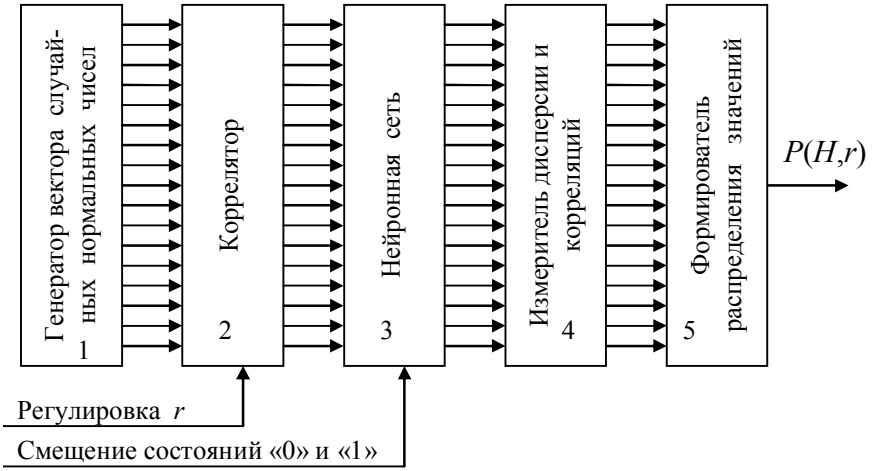


Рисунок 3.2 – Структурная схема, воспроизводящая плотность распределения значений биномиального закона с зависимыми данными

На приведенной выше структурной схеме блок-1 генерирует вектор независимых нормально распределенных данных (эмулируя биометрические данные «Все чужие»). Для реализации блока-1 мо-

гут быть использованы аппаратные или программные генераторы случайных независимых чисел. Блок-2 осуществляет связывание независимых данных с заданным значением парных коэффициентов корреляции. Нужным образом скоррелированные данные поступают на вход нейронной сети, заранее обученной распознавать некоторый биометрический образ. При этом, регулируя параметры нейронов нейросети-3, мы можем смещать вероятности состояний на выходе преобразователя биометрия/код. Данные с выходов преобразователя биометрия/код поступают на блок-4, вычисляющий результирующие коэффициенты корреляции и дисперсии. Следующий блок-5 сглаживает полученные гистограммы данных и формирует кривые биномиального зависимого закона распределения значений.

Общая тенденция изменения закона распределения значений при увеличении связанности (зависимости) исходных данных отражена на рисунках 3.3, 3.4, 3.5.

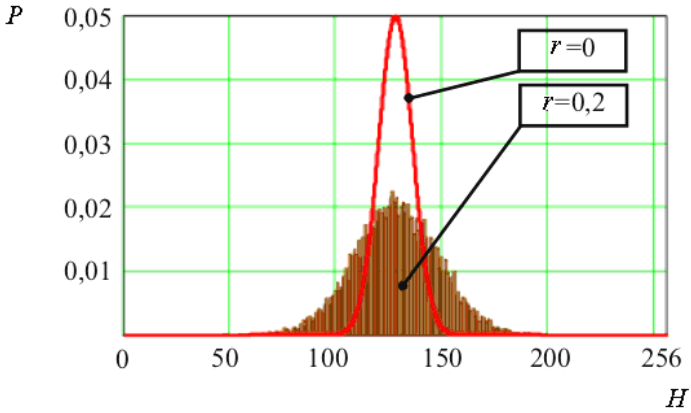


Рисунок 3.3 – Эффект расширения дисперсии нормального закона распределения значения при модулях коэффициентов парной корреляции менее 0,3

Рисунок 3.3 отражает первую фазу метаморфозы биномиального закона распределения значений. В этой фазе вид закона распределения кардинально не меняется. Закон распределения значений продолжает оставаться близким к нормальному, претерпевает изменение

только его второй момент, существенно увеличивается дисперсия наблюдаемого распределения.

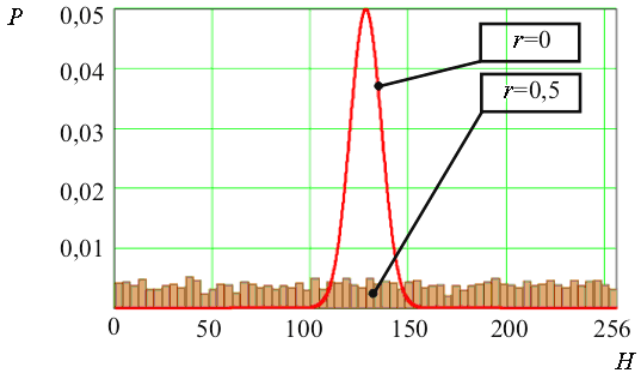


Рисунок 3.4 – Эффект вырождения нормального закона распределения значений в равномерный закон распределения значений в особой точке $|r| = 0{,}5$

Рисунок 3.4 отражает момент, когда нормальный закон полностью выродился в равномерный закон распределения значений. Мы имеем явную, хорошо наблюдаемую численными методами метаморфозу, что позволяет рассматривать идентифицируемую плотность распределения значений в интервале $0 \langle |r| \langle 0{,}5$ как некоторую суперпозицию нормального и равномерного законов распределения значений.

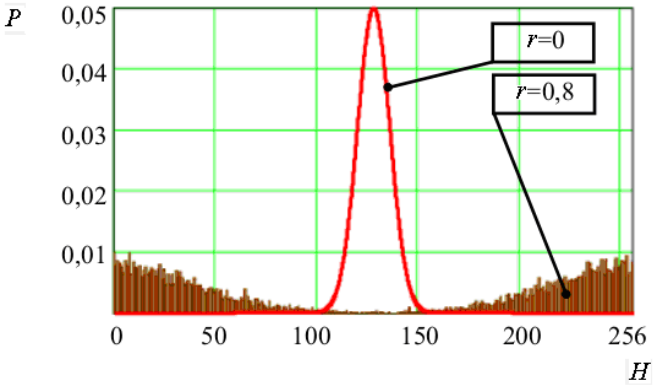


Рисунок 3.5 – Эффект перерождения равномерного закона распределения значений в двумодальный закон типа «arcsin(x)» при $|r| \langle 0{,}5$

Рисунок 3.5 иллюстрирует перерождение равномерного закона распределения значений в некоторый двумодальный закон с минимальной плотностью распределения значений в центре.

В случае синтеза аналитического описания зависимого биномиального закона распределения значений необходимо, чтобы полученное аналитическое выражение описывало без разрывов и искусственных точек переключения все описанные выше метаморфозы. Общее решение для этой задачи пока не найдено. На данный момент построены гистограммы распределений, дающие после сглаживания номограммы, приведенные на рисунках 3.6, 3.7.

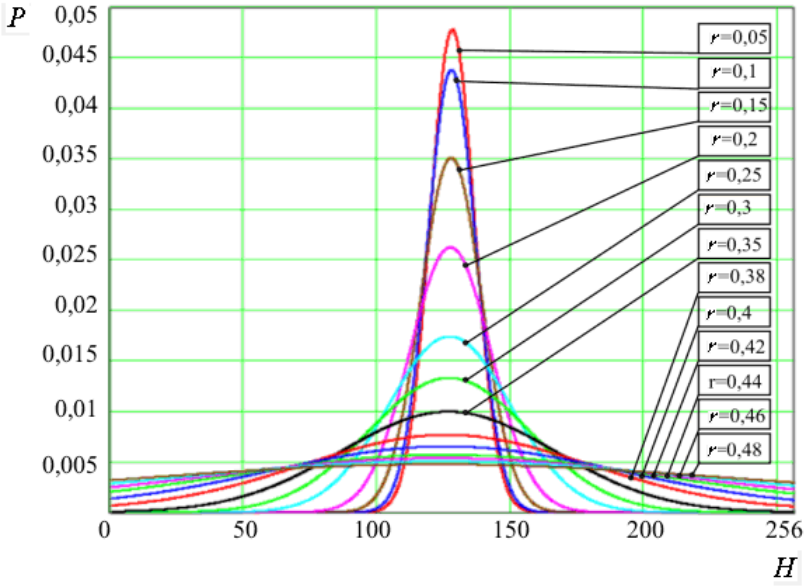


Рисунок 3.6 – Номограмма биномиального зависимого закона распределения значений в диапазоне зависимости $0 \langle |r| \langle 0{,}5$

Следует подчеркнуть, что приведенные выше номограммы получены сглаживанием реальных данных. Реальные данные получены в результате численного эксперимента. Они имеют достаточно большие флуктуации из-за конечности тестовых выборок. Особенно это заметно на редких событиях, имеющих маленькую вероятность.

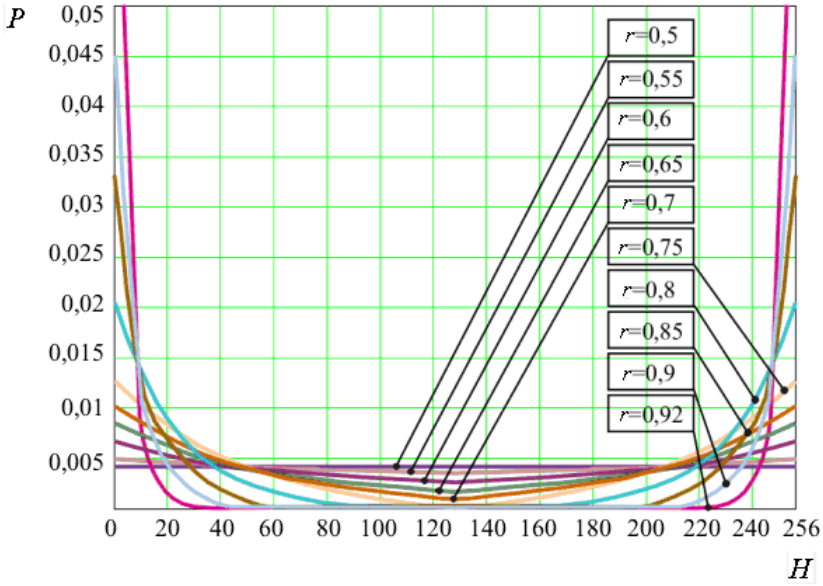


Рисунок 3.7 – Номограмма биномиального зависимого закона распределения значений в диапазоне зависимости $0{,}5 \langle |r| \langle 1{,}0$

3.3. Перечень проблем, связанных с синтезом численно-аналитического описания зависимого биномиального закона распределения значений

В предыдущем параграфе была изложена методика численного моделирования зависимых биномиальных плотностей распределения значений, предполагающая получение исходных гистограмм с достаточно высокой точностью. При этом предполагается по умолчанию, что число экспериментов может быть сделано как угодно большим и точность вычислений (например, разрядность вычислительной машины) может быть сделана достаточно высокой.

Очевидно, что перечисленные выше предположения корректны только для инженерных приложений, а для высокоточных статистических расчетов они нуждаются в проверках. Начнем с самого простого – с разрядной сетки вычислений. Когда нас устраивает погрешность вычислений в на уровне 0,1 % (ошибка 10^{-3}), то нам впол-

не подходят 8-разрядные машины. Для вычислений с ошибкой на уровне 0,0001 % (ошибка 10^{-6}) потребуются уже 16-разрядные машины с соответствующей 16-разрядной математикой. Для статистических расчетов с ошибкой 10^{-12} придется использовать 32-разрядные машины с 32-разрядной математикой. Прямые попытки повысить точность расчетов до ошибок на уровне 10^{-24} приводят к необходимости привлечения 64-разрядной математики. Если учесть, что нам могут понадобиться расчеты еще более точные, то придется признать наличие достаточно серьезной проблемы.

Наиболее серьезная часть этой проблемы не в разрядности современных вычислительных машин, а в наличии соответствующей математики (соответствующего высокоточного программного обеспечения). Стандартное программное обеспечение явно имеет недостаточную точность вычислений и ориентировано на грубые инженерные расчеты. В качестве примера можно привести программные генераторы случайных чисел. Все они псевдослучайные и имеют некоторый достаточно большой период повторения. Внутри этого периода их данные можно считать случайными, а за его пределами данные становятся детерминированными. Особенно это хорошо видно, когда тестовая выборка выбирается длиной, точно кратной периоду псевдослучайности.

Еще одной важной проблемой является проблема сбалансированности (представительности) генераторов псевдослучайных чисел. Как правило, хорошо сбалансированными по статистическим моментам являются только генераторы равномерных законов распределения значений, их балансировать наиболее просто. Генераторы нормального закона распределения значений очень часто имеют дефекты «хвостов», они просто не могут воспроизводить очень редкие события.

Последней существенной проблемой является проблема конечных вычислительных возможностей (по быстродействию и оперативной памяти) вычислительных машин. Работа с большими массивами данных требует не только достаточно высокой разрядности используемых вычислительных машин, но и их высоких возможностей по оперативной памяти и быстродействию. Как только ресурс оперативной памяти вычислительной машины исчерпывается, происходит резкое снижение быстродействия вычислений, так как время обращения к

памяти на магнитном носителе (винчестеру) на несколько порядков ниже времени обращения к оперативной памяти.

Аналогичная ситуация возникает и при использовании вычислительных машин разной мощности (разной тактовой частоты, с разным числом параллельно работающих процессоров). Современные аппаратные средства позволяют снизить время численного эксперимента в десятки и сотни раз, при правильном выборе. Если учесть, что время проведения достаточно точного численного эксперимента может затянуться на несколько лет непрерывной работы нескольких вычислительных машин, выбор их характеристик, оптимизация программного обеспечения и размеров оперативной памяти становятся крайне важными задачами.

3.4. Симметрия функций распределения биномиального зависимого закона распределения значений относительно среднего модуля коэффициентов корреляции

Из приведенных выше рисунков видно, что дисперсия меры Хемминга, как и сам закон распределения значений меры Хемминга множества «Все чужие», сильно зависит от связанности (коррелированности) разрядов случайных кодов. Возникает вопрос о том, какова эта связь. Необходимо путем численных экспериментов идентифицировать эту связь [36].

Для идентификации вида связи, например, $\sigma(r)$, были сделаны проверки на чувствительность зависимости к знаку коэффициентов парной корреляции. Для этой цели сравнивались значения $\sigma(r)$ и вид функций распределения значений $p(H)$ для всех положительных коэффициентов

$$r_{1,2} = +0{,}5,\ r_{1,3} = +0{,}5,\ r_{1,4} = +0{,}5,\ \ldots,\ r_{i,j} = +0{,}5;$$

для всех отрицательных коэффициентов

$$r_{1,2} = -0{,}5,\ r_{1,3} = -0{,}5,\ r_{1,4} = -0{,}5,\ \ldots,\ r_{i,j} = -0{,}5;$$

для коэффициентов, поочередно изменяющих свой знак:

$$r_{1,2} = +0{,}5,\ r_{1,3} = -0{,}5,\ r_{1,4} = +0{,}5,\ \ldots,\ r_{i,j} = \pm 0{,}5.$$

Результаты численного эксперимента показали, что знаки коэффициентов парной корреляции вообще не влияют на вид закона распределения значений и его дисперсию, т. е. с точностью до ошибки, определяемой конечной обучающей выборкой, выполняется тождество

$$\sigma(r = +0{,}5) \equiv \sigma(r = -0{,}5) \equiv \sigma(r = \pm 0{,}5). \qquad (3.1)$$

В точке $r = \pm 0{,}5$ форма закона распределения значений $p(H, p = 0{,}5)$ равномерная, и она продолжает оставаться равномерной (см. рисунок 3.4) при любом распределении знаков коэффициентов парной корреляции. Неизменность формы плотности распределения значений – $p(n, H, p = 0{,}5, |r| = 0{,}5)$ – проще всего констатировать именно при наблюдении равномерного закона распределения значений как самого простого.

При других значениях модулей коэффициентов корреляции соотношение (3.1) и неизменность формы $p(n, H, p = 0{,}5, |r| = 0{,}5)$ продолжают выполняться. Это означает, что тождество (3.1) может быть записано в более общем виде

$$\sigma(+r) \equiv \sigma(-r) \equiv \sigma(\pm r) \equiv \sigma(|r|). \qquad (3.2)$$

Аналогичное соотношение может быть записано и для функций распределения значений в центральной точке симметрии

$$p(n, H, p = 0{,}5, +r) \equiv p(n, H, p = 0{,}5, -r) \equiv$$
$$\equiv p(n, H, p = 0{,}5, \pm r) \equiv p(n, H, p = 0{,}5, |r|), \qquad (3.3)$$

а также в любой иной точке

$$p(n, H, p, +r) \equiv p(n, H, p, -r) \equiv p(n, H, p, \pm r) \equiv p(n, H, p, |r|). \quad (3.4)$$

Подчеркнем, что все вышесказанное выполняется только для одинаковых по модулю коэффициентов корреляции, имеющих только неопределенный знак.

В случае, если модуль коэффициентов парной корреляции может принимать несколько значений, например $|r| = 0{,}0,\ 0{,}25,\ 0{,}5$, ситуация усложняется и мы будем получать разные значения дисперсий и разные функции плотности распределения значений, в зависимости от вероятности состояний $|r| = 0{,}0,\ 0{,}25,\ 0{,}5$, т. е. можно говорить только о том, что дисперсии и функции плотности распределения

значений будут совпадать при совпадении плотностей распределения значений модулей коэффициентов парной корреляции.

Для учета множества возможных значений состояний необходимо осуществлять весовое суммирование результата каждой компоненты. Наиболее просто это осуществляется через синтез необходимого числа примеров с заданными значениями коэффициентов парной корреляции, объединением всех примеров в одну тестовую группу и последующим ее статистическим исследованием.

Более сложным путем является синтез специальных генераторов, которые дают выходные коды с заранее заданным распределением значений коэффициентов парной корреляции.

3.5. Проверка гипотезы биномиального зависимого закона распределений значений выходных кодов многомерного нейросетевого преобразователя

Проблема доказательства соответствия распределений кодов «Все чужие» зависимому биномиальному закону связана с тем, что реальные выходные коды многомерных нейросетевых преобразователей не могут иметь однозначных коэффициентов парной корреляции. Они всегда имеют некоторое практически непрерывное распределение значений коэффициентов парной корреляции $p(r)$ (см. рисунок 3.1). Естественно, что это распределение будет иметь моду и математическое ожидание. Для симметричных функций распределения значений они совпадают, однако в более общем случае несимметричных функций распределения значений мода и математическое ожидание перестают совпадать.

Корень проблемы состоит в том, что при численном синтезе зависимого биномиального закона распределения значений (см. рисунок 3.2) мы имели возможность задать строго фиксированные значения параметров распределения – p, r. Если же мы пытаемся сделать обратное преобразование и проверить то, насколько соответствуют реальные данные зависимому биномиальному закону, мы сталкиваемся с размытым, нечетким характером данных. Вместо конкретных, фиксированных параметров p, r мы имеем их размытые, нечеткие непрерывные распределения, и нам нужно им поставить в соответствие четкую многомерную функцию $p(n, H, p, r)$. Для выхода из этой

ситуации разобьем конечную шкалу параметра – r на i-поддиапазонов и в процессе эксперимента будем осуществлять по этим поддиапазонам сортировку данных. При сортировке будем убирать неопределенность (размытость) многообразия коэффициентов корреляции через вычисление математического ожидания модуля всех возможных значений – $m\left|r_{i,j}\right|$. Все возможные значения $m\left|r_{i,j}\right|$ находятся в диапазоне от 0,0 до 1,0. Необходимо разбить этот диапазон на несколько поддиапазонов и при тестировании сортировать результаты, относя их к выделенным поддиапазонам.

На рисунке 3.8 приведена структурная схема реализации численного эксперимента, позволяющего оценить соответствие распределения меры Хемминга реальных кодов и синтезированного ранее зависимого биномиального закона распределения значений.

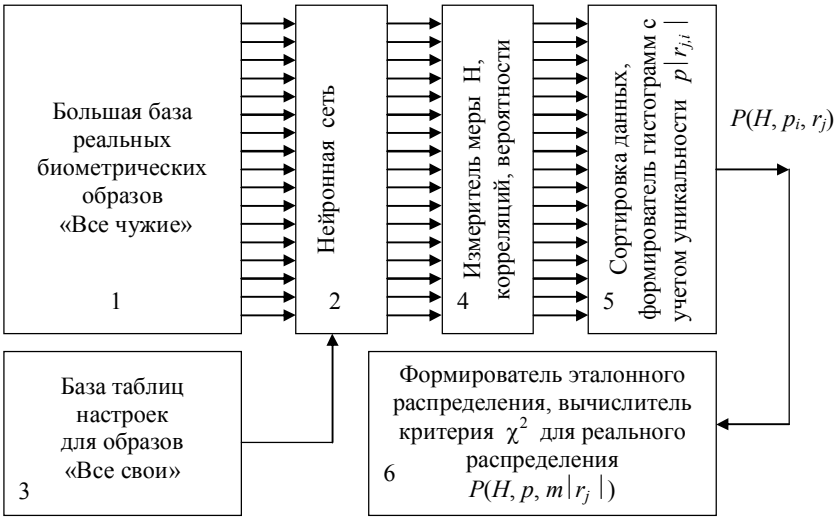


Рисунок 3.8 – Структурная схема проведения численного эксперимента по проверке гипотезы биномиального зависимого распределения

Блок-1 на рисунке 3.8 соответствует большой базе реальных биометрических образов «Все чужие», имеющей порядка 10^7, …, 10^9 рукописных или голосовых образов. Блок-2 соответствует тестируемому нейросетевому преобразователю. Блок-3 является базой таблиц настроек для множества образов «Все свои», имеющей порядка 10^4

настроек. Блок-4 осуществляет измерение меры Хемминга и множества коэффициентов парной корреляции выходных разрядов кодов. Блок-5 осуществляет сортировку данных по распределениям коэффициентов корреляции. Блок-6 осуществляет формирование эталонного распределения значений для полученного на практике распределения значений коэффициентов корреляции и вычисляет критерий согласия – χ^2.

Численный эксперимент по проверке гипотезы может занимать очень большое время, если требуется иметь результаты высокой достоверности [37]. Приходится проверять множество настроек образов «Все свои», причем для каждого из десятка тысяч этих образов необходимо пропустить через эмулятор нейросети (блок-2) всю базу образов «Все чужие». На обычной ПЭВМ удается эмулировать порядка 3000 больших нейронных сетей в секунду, т. е. на полную проверку 10^9 реальных биометрических образов уйдет порядка 10 лет непрерывных вычислений. Несколько лет необходимо для тестирования только одного из 10 000 образов «Свой». Задача прямого сравнения теоретического распределения значений зависимого биномиального закона оказывается очень сложной, необходимо ее упрощение за счет использования некоторых хорошо проверяемых аналитических преобразований. Например, в ряде случаев может быть использована нормализация биномиального зависимого закона распределения значений меры Хемминга.

3.6. Оценка границ применимости гипотезы нормальности распределения меры Хемминга для выходных кодов многомерных преобразователей

Синтез нового биномиально зависимого закона распределения значений и его использование для сокращения ресурсов на тестирование высоконадежных средств биометрической аутентификации несомненно являются очень сложной задачей. Более простой является задача использования классического нормального закона распределения значений, аппроксимирующего биномиальный зависимый закон «Все чужие». Естественно, что такая упрощающая задачу аппроксимация может быть применима в ограниченном диапазоне кор-

релированности данных. Необходимо оценить этот диапазон для множества биометрических образов «Все чужие».

На рисунке 3.9 приведена относительная ошибка (в процентах) от аппроксимации биномиального центрированного зависимого закона нормальным законом как функция от коэффициента корреляции.

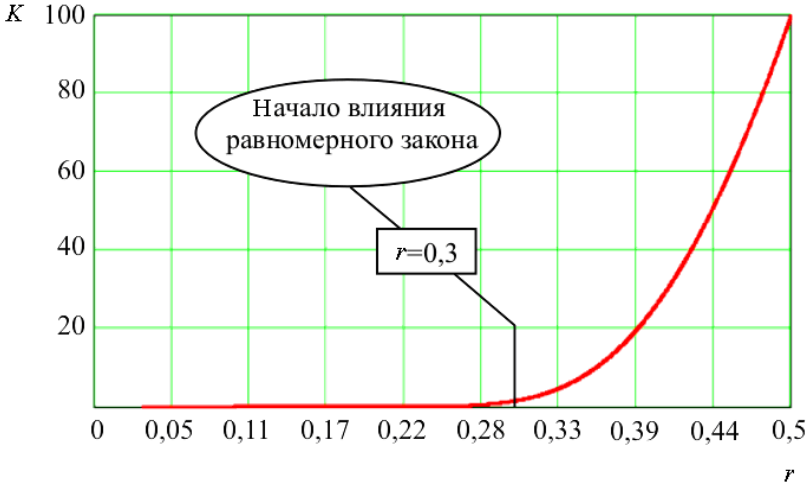


Рисунок 3.9 – Относительная ошибка от аппроксимации биномиального зависимого закона нормальным как функция коэффициента корреляции

Из рисунка 3.9 видно, что при связанности данных $r = \pm 0{,}3$ относительная ошибка приближения составляет порядка 3 %, т. е. относительная ошибка замещения биномиального зависимого закона нормальным будет составлять менее 3 %, если модули коэффициентов парной корреляции будут меньше 0,3. Мы наблюдаем существенно нелинейную связь значений коэффициентов парной корреляции и со значением относительной ошибки аппроксимации. Из рисунка 3.9 видно, что со снижением взаимной корреляции ошибка очень быстро падает и становится неразличимой.

Для того, чтобы увидеть малые значения ошибки аппроксимации, необходимо перейти к логарифмическому масштабу. Кривая логарифма относительной ошибки отображена на рисунке 3.10.

Из этого рисунка видно, что, снизив значение коэффициентов парной корреляции до величины менее 0,08, удается уменьшить ошибку приближения на 12 порядков. Для идеальных биометрико-

нейросетевых преобразователей с нулевыми коэффициентами парной корреляции нормальный закон распределения значений меры Хемминга аппроксимирует биномиальный закон распределения значений с очень высокой точностью. При уменьшении корреляции до нуля точность замены одного закона на другой может быть как угодно мала. График, отображенный на рисунке 3.10, ограничивается ошибкой 10^{-12} только из-за использования 32-разрядной арифметики при вычислениях.

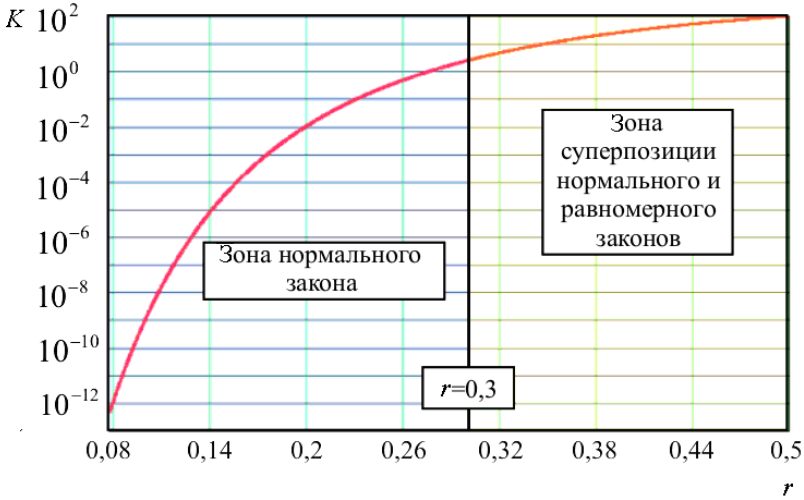


Рисунок 3.10 – Относительная ошибка от аппроксимации биномиального зависимого закона нормальным как функция коэффициента корреляции при логарифмической шкале

Заметим, что порог приемлемости аппроксимации нормальным законом распределения значений является относительной величиной. Для низконадежных биометрических систем с вероятностями ошибок второго рода на уровне 10^{-3}, …, 10^{-4} и парной коррелированностью на уровне $|r| \langle 0{,}14$ ошибка замещения оказывается вполне приемлема и составляет менее 10^{-5}. Однако если мы будем тестировать высоконадежные системы биометрической аутентификации с ошибкой менее 10^{-12}, то даже при корреляции $|r| \langle 0{,}14$ мы уже не можем использовать замещение биномиального зависимого закона нормальным законом с математическим ожиданием в центре шкалы меры Хемминга.

Необходимо обратить внимание на то, что классический биномиальный закон является действительно независимым только в одной точке $p = 0{,}5$. Только в этом случае корреляция между тестами отсутствует полностью. Как только $p \neq 0{,}5$, появляется ненулевая корреляция между опытами, и чем больше вероятность исходов отличается от 0,5, тем больше будет корреляционная связь. При $p = 0$ и $p = 1$ $r = 1$. Зависимость корреляции исходов последовательно совершаемых бросаний монеты от значения параметра – p описывается параболой, обращающейся в «0» в точке $p = 0{,}5$:

$$r = (1 - 2p)^2 . \qquad (3.5)$$

Кривая зависимости $r(p)$ приведена на рисунке 3.11.

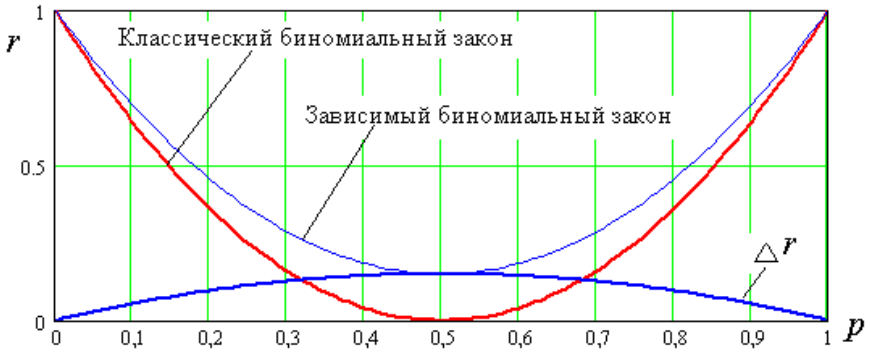


Рисунок 3.11 – Зависимость коэффициентов корреляции от вероятности состояний «0» и «1» для классического «независимого» биномиального закона и зависимого биномиального закона

В случае зависимого биномиального закона распределения значений коэффициент корреляции никогда не обращается в ноль. Он может опуститься только до некоторого остаточного коэффициента корреляции – $r_{ост}$. В этом случае зависимость будет описываться следующим соотношением:

$$r = (1 - r_{ост})(1 - 2p)^2 + r_{ост} . \qquad (3.6)$$

График $r(p)$ для зависимого биномиального закона воспроизведен на рисунке 3.11 тонкой линией.

Разница между выражениями (3.6) и (3.5) дает относительную ошибку корреляции – Δr между классическим и зависимым биномиальными законами. Очевидно, что эта ошибка будет определять и

ошибку аппроксимации зависимого биномиального закона нормальным законом, т. е. при смещении от центра, где самая большая ошибка приближения, к краям погрешность замещения падает. В крайних точках «0» и «1» она оказывается нулевой.

3.7. Оценка стойкости к атакам подбора открытого (скомпрометированного) биометрического образа

Одной из важнейших характеристик биометрико-нейросетевого преобразователя является оценка его стойкости к атаке подбора при открытом (скомпрометированном) биометрическом образе. Предполагается, что злоумышленник знает парольное слово (парольную фразу) в ее письменном рукописном (голосовом варианте), а также имеет ее формулированную запись (например, печатными буквами).

Очевидно, что стойкость системы биометрической защиты будет тем ниже, чем сильнее будет скомпрометирован биометрический образ. Различают три уровня компрометации:

1) злоумышленнику известен только сам пароль (образцов почерка, голоса «Своего» у злоумышленника нет);

2) злоумышленнику известны сам пароль и почерк, голос «Своего» без детальных подробностей;

3) злоумышленнику известен биометрический парольный образ (осуществлен перехват тайного биометрического образа и изготовлен его электронный или физический муляж).

В ситуации 3 при полной компрометации биометрического образа спасти положение могут только организационно-технические меры, препятствующие злоумышленнику на защищенной территории (наблюдение и контроль за аутентифицируемым, препятствующее применение им физического или электронного муляжа, пломбирование средств аутентификации, контроль целостности аппаратной и программной составляющих,…).

В ситуации 2 стойкость биометрической защиты определяется только умением злоумышленника изменять свой голос, почерк в рамках своей естественной группы почерков (голосов) и иной биометрии.

В ситуации 1 стойкость биометрической защиты определяется только уникальностью конкретного биометрического образа пользователя или числом групп пользователей. Внутри одной группы система не способна различать пользователей.

Оценены и измерены могут быть только вероятности удачи атаки подбора в первой и второй ситуациях. В третьей ситуации возможны только экспертные оценки.

Для оценки вероятности удачи атаки подбора в первой ситуации необходимо иметь достаточно большую базу одинаковых по смысловому содержанию, но разных по информативности биометрических образов. Например, для оценки биометрической защиты по анализу рукописных почерков необходимо иметь множество образов одного и того же слова «*Пенза*», написанного разными почерками. Далее необходимо обучить систему под одного из пользователей. Затем, подавая на вход системы имеющиеся биометрические образы «*Пенза*», мы должны оценить параметры распределения меры Хемминга «Чужого», знающего слово-пароль. Найдя математическое ожидание распределения «Чужих» и их дисперсию, мы можем оценить вероятность удачи атаки подбора, опираясь на гипотезу нормальности.

Как правило, вероятность удачи атаки подбора известного слова-пароля существенно зависит от числа букв в слове-пароле. При использовании коротких слов вероятность удачи оказывается высока из-за их малой информативности. С ростом длины парольного слова вероятность удачи его подбора снижается, однако некоторые системы не приспособлены для работы со слишком длинными (слишком сложными) биометрическими образами. При использовании слишком сложных биометрических образов вероятность атаки может оставаться на одном уровне или даже уменьшаться. Тщательные испытания должны давать зависимости вероятностных характеристик системы защиты как функцию от длины скомпрометированного биометрического образа.

Тестирование в ситуации 2, когда злоумышленник имеет возможность учиться почерку (голосу) «Своего», должно давать вероятность удачи атаки выше, чем тестирование в ситуации 1. Для организации этого тестирования «Чужому» следует предоставить обратную связь

(сообщать то, насколько его вариант далек от «Своего»). Должны использоваться заинтересованные в удачной атаке добровольцы. Статистические данные собираются при их самообучении и попытках выполнить как можно более близкий вариант подделываемого биометрического образа. Оценки вероятности удачи ведутся в рамках гипотезы нормального закона распределения значений множеств «Свой» и «Чужой». Автоматизировать оценку вероятности удачи при атаках подбора во второй ситуации не удается.

3.8. Упрощение тестирования защиты за счет частичной (побуквенной) компрометации биометрического образа

Заметим, что проверка вероятностей ошибок первого рода (отказ «Своему») и ошибок второго рода (пропуск «Чужого») при скомпрометированном пароле является не очень сложной задачей, так как эти вероятности достаточно большие и не требуют больших объемов тестовых образов. Положение меняется, когда необходимо оценить вероятность ошибок второго рода нескомпрометированного пароля или почти нескомпрометированного пароля. В этом случае вероятность ошибочного пропуска чужого может быть крайне мала и для ее оценки требуются много времени и большая база тестовых образов. Заниматься полноценным тестированием из-за его дороговизны пользователю нет смысла. Привлекать кого-либо со стороны опасно, так как можно скомпрометировать свой тайный биометрический образ.

Решить эти проблемы пользователь может самостоятельно, искусственно ослабив систему защиты. Кроме того, по требованиям национального стандарта [11] потребителю должна быть предоставлена возможность самостоятельной оценки стойкости средств, заявленных производителем как высоконадежные.

Одним из способов ускоренного тестирования является применение минимально возможных модификаций рукописного пароля, на котором система биометрической защиты обучена. В качества примера на рисунке 3.12 приведена экранная форма с рукописным написанием слова-пароля *«Пенза»* почерком «Своего», а также два близких варианта написания пароля *«Пензу»* и *«Пензе»*, отличающихся только одной последней буквой.

Естественно, что при написании другого рукописного пароля, отличающегося даже в одном знаке, на выходе преобразователя биометрия код будут появляться другие ключи. Расхождение получаемых

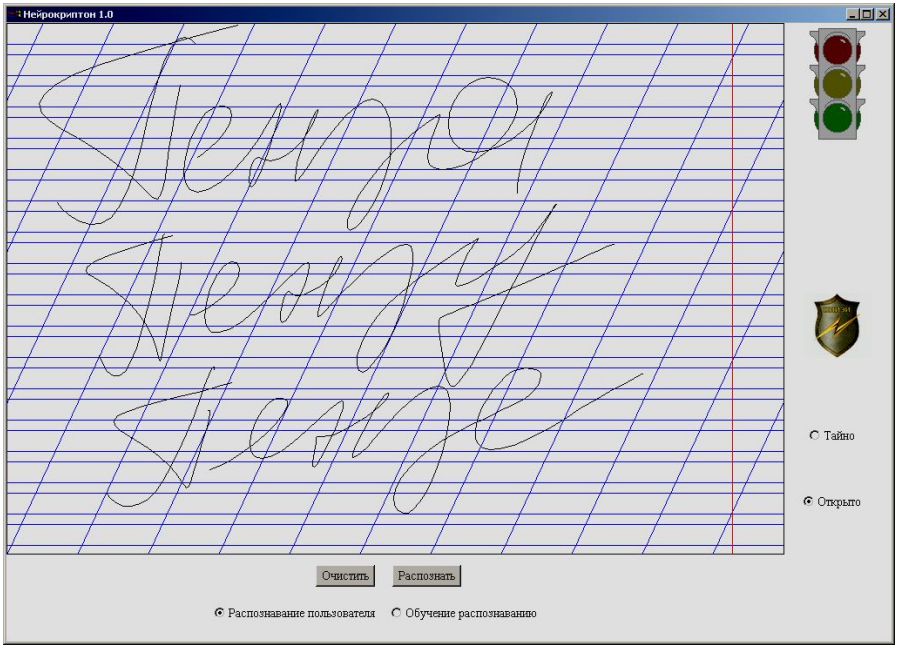


Рисунок 3.12 – Варианты минимального изменения рукописного слова-пароля в написании одной последней буквы рукою «Своего» пользователя

кодов с действительным ключом будет частичным. Рукописный пароль в нашем случае состоит из 5 букв соответственно, изменив одну букву, мы вправе ожидать изменения примерно 10, …, 20 % бит в близких кодах. Эту величину пользователь вполне может контролировать, вычисляя число не совпавших бит у правильного ключа и близких к нему неправильных ключей. В частности, для ключей длиной 256 бит приведенные на рисунке 3.12 близкие пароли дают расстояния Хемминга 13 и 70 бит. В таблице 3.1 приведены расстояния Хемминга, полученные для 14 вариантов последней буквы рукописного пароля. В этой же таблице даются значение математического

ожидания меры Хемминга (m = 42,0) и значение среднеквадратического отклонения меры Хемминга (σ = 27,1).

Таблица 3.1 – Расстояния Хемминга для близких кодов, полученных модификацией последней буквы рукописного пароля «*Пенза*»

Последняя буква	а	у	е	ы	о	и	ю	к	н	б	в	г	д	з
Мера Хемминга	0	13	70	6	11	4	18	3	90	41	35	53	82	30
	0	26	53	9	44	20	93	16	83	57	45	27	74	39
	0	18	49	24	30	9	48	33	129	81	12	48	49	47
	0	40	61	18	16	51	79	17	64	34	73	39	55	19
Статистические моменты		m = 42,0 (математическое ожидание); σ = 27,1 (среднеквадратическое отклонение)												

Заметим, что мера Хемминга является по определению дискретной положительной величиной, распределение значений меры Хемминга также является дискретным и хорошо описывается биномиальным законом. На рисунке 3.13 дан пример дискретного биномиального распределения и его непрерывной огибающей, которые построены по данным таблицы 3.1. В силу того, что переход от дискретного биномиального распределения к непрерывному биномиальному распределению сопряжен с рядом вычислительных проблем, при расчетах рекомендуется использовать аппроксимацию биномиального распределения нормальным законом распределения. Из рисунка 3.13 видно, что непрерывный биномиальный закон распределения и эквивалентный ему нормальный закон распределения достаточно близки, интегральная ошибка в расчетах от замещения одного закона другим не превышает 20 %.

В рамках аппроксимации данных таблицы 3.1 нормальным законом распределения вычисление вероятности пропуска «Чужого» составляет величину 0,061 (площадь под кривой нормального закона распределения значений, соответствующая отрицательным значениям «псевдомеры Хемминга»).

Располагая полученной оценкой вероятности пропуска «Чужого», мы можем оценить стойкость биометрического пароля. Если изменение пароля в одной пятой букве дает оценку $P_{2,5} \cong 0{,}061$, то оценить стойкость всего пароля можно, возведя эту величину в 5-ю степень:

$$P_2 \cong (P_{2,5})^5 = (0{,}061)^5 = 10^{-6,1} . \qquad (3.7)$$

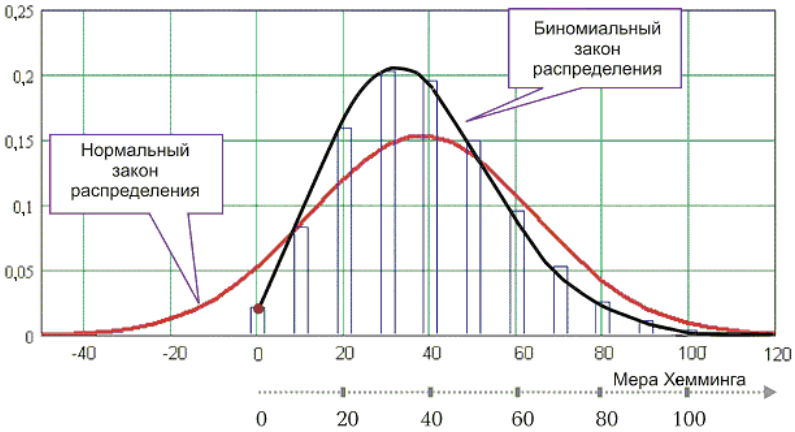


Рисунок 3.13 – Аппроксимация дискретного биномиального закона распределения значений меры Хемминга эквивалентным непрерывным нормальным законом распределения

Корректность оценки обусловлена тем, что для паролей мы имеем право перемножать вероятности удачи подбора каждой из 5 букв, так как злоумышленник не знает ни букв, ни особенностей их рукописного воспроизведения «Своим», т. е.:

$$P_2 \cong P_{2,1}\, P_{2,2}\, P_{2,3}\, P_{2,4}\, P_{2,5} . \qquad (3.8)$$

При предложенном выше способе оценки стойкости системы к атакам подбора пользователь никого не привлекает для экспериментов. Он сам пытается взломать свою защиту, используя свой личный почерк. Это обстоятельство существенно занижает оценку стойкости системы, так как злоумышленник не имеет возможности атаковать систему защиты, используя уникальный рукописный почерк владельца системы. Скорректировать заниженную оценку можно через ее умножение на вероятность подбора полностью скомпрометированного пароля «Чужим». На сегодняшний день производители заявляют вероятность ошибки второго рода для известных (скомпрометированных) рукописных и голосовых образов на уровне 10^{-2}. Как следствие, оценка вероятности пропуска «Чужого» в нашем случае должна составить $10^{-9,1}$.

3.9. Оценка снизу стойкости преобразователей к атакам подбора случайными рукописными фразами

По аналогии с описанным выше способом оценки стойкости можно применять и другой похожий метод тестирования. Он состоит в том, что используется база «Чужих» рукописных образов, имеющая порядка 100, …, 300 примеров разных случайных (слабо коррелированных) биометрических образов. Применительно к тестированию голосовых и рукописных средств высоконадежной аутентификации примеры случайных парольных слов (парольных фраз) могут быть выбраны из книги, открытой на случайной странице.

После этого необходимо обучить средство защиты на образах «Свой» и, подавая на вход обученного средства базу примеров «Чужие», найти параметры соответствующего распределения меры Хемминга. На рисунке 3.14 приведены результаты тестирования одного из макетов биометрико-нейросетевого преобразователя при испытании в разных режимах и аппроксимации результатов испытаний нормальным законом распределения значений.

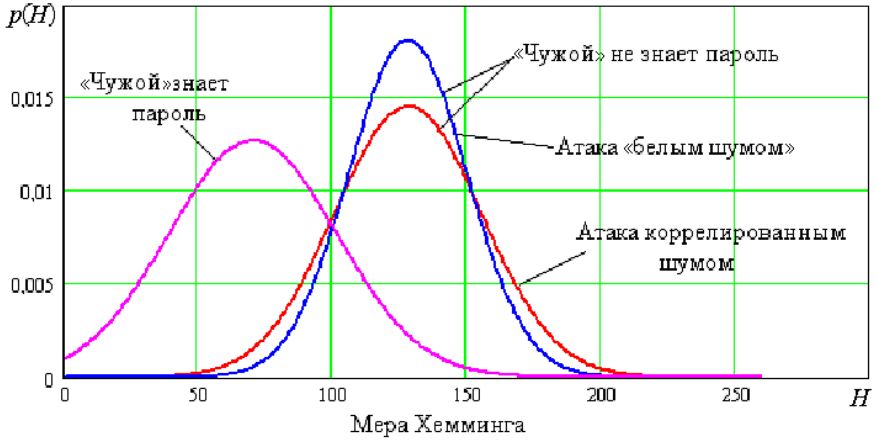


Рисунок 3.14 – Расстояния Хемминга между заданным кодом ключа «Свой» откликами выходных кодов при воздействии «Чужого»

Описанное выше тестирование может быть условно названо «атакой коррелированным шумом». Если злоумышленник при атаке будет писать слова из словаря или извлекать образы из заранее состав-

ленного словаря, где слова расположены как в обычном словаре, то соседние биометрические образы, написанные одной рукой (воспроизведенные одним голосом), будут сильно коррелированы.

Практика подобного тестирования показывает, что оно дает существенно заниженные оценки стойкости средств защиты к атакам коррелированным шумом. Это, видимо, обусловлено тем, что аппроксимация распределения меры Хемминга для высоконадежных средств биометрии оказывается слишком грубой. Кроме того, мера Хемминга не учитывает то обстоятельство, что злоумышленнику нужно угадать все до одного бита ключа в нужном порядке. Мера Хемминга учитывает только число угаданных бит, оставляя без внимания порядок расположения угаданных и не угаданных разрядов кода ключа «Свой».

3.10. Оценка снизу стойкости преобразователей к атакам подбора случайными некоррелированными данными

В соответствии с идеологией проекта национального стандарта [11] биометрический вектор входных биометрических параметров нейросети, как правило, оказывается слабее, чем выходной ключ и использующая его криптография. Как следствие, злоумышленнику выгоднее атаковать защиту со стороны ее биометрических входов [38].

Для организации эффективной атаки подбора биометрических данных злоумышленнику необходимо иметь статистику распределения значений контролируемых механизмом защиты биометрических параметров всех возможных биометрических образов. Для сбора статистики достаточно подать на вход взламываемой биометрической системы несколько сот однотипных биометрических образов и, контролируя каждый из входов нейросетевого преобразователя, рассчитать необходимое число статистических моментов. Подобные статистические исследования упрощаются тем, что, как правило, контролируемые биометрические параметры имеют нормальный закон распределения, а при предварительной обработке биометрических данных зачастую осуществляется их центрирование. Тогда злоумышленнику необходимо определить только вектор дисперсий контролируемых биометрических параметров. Блок-схема предварительных

статистических исследований биометрического механизма защиты приведена на рисунке 3.15.

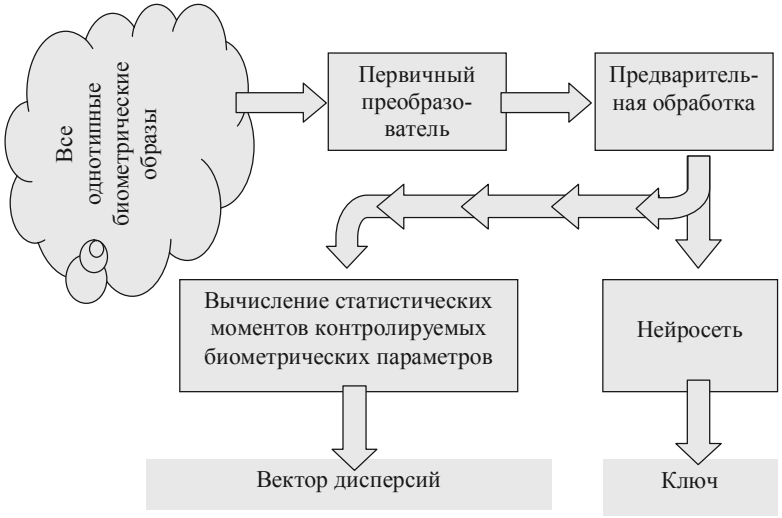


Рисунок 3.15 – Блок-схема проведения статистического исследования биометрического механизма защиты

Зная статистическое распределение биометрических параметров, злоумышленник может попытаться подобрать входной биометрический образ. Для организации атаки подбора он должен подавать на входы нейронной сети случайные числа, соответствующие обнаруженным им распределениям входных случайных величин. Так как злоумышленник ничего, кроме статистики распределения входных биометрических данных, не знает, разумно предположить, что он будет использовать наиболее простые для реализации генераторы независимых случайных величин.

Организация подобной атаки вполне реальна, и необходимо оценивать стойкость защиты от нее обученного преобразователя биометрия/код. Оценка осуществляется путем программной эмуляции атаки случайного подбора. Как и в предыдущем тесте, оценивается распределение значений меры Хемминга и в рамках гипотезы нормального закона ее распределения, оценивается вероятность удачи.

Этот тест целесообразно называть именем атаки, на базе которой он построен. Результаты тестирования стойкости защиты к атакам «белого шума»» отражены на рисунке 3.14. Из этого рисунка видно, что распределение «атаки “белого шума”» и распределение «атаки коррелированным шумом» близки. Это свидетельствует о хорошей балансировке тестируемого преобразователя к обоим типам атак. При неудачно выбранной структуре нейросети и плохом ее обучении возможны весьма и весьма серьезные расхождения стойкости преобразователя к разнотипным атакам. Именно по этой причине проект стандарта [11] требует от производителя контроля стойкости преобразователей к разнотипным атакам подбора.

3.11. Оценка стойкости ослабленных преобразователей биометрия/код с использованием тестовых машин случайного подбора

Одной из самых надежных тестовых процедур является использование машин, реализующих атаку подбора входных биометрических данных пользователя «Свой» до первой удачи. Так как подобрать все входные биометрические параметры не удается, задачу подбора упрощают, считая часть биометрических параметров известными. Например, можно выбрать случайным образом 4 входных биометрических параметра, считать все другие параметры известными и начать подбор только неизвестной четверки. Подбор осуществляется подстановкой независимых случайных чисел, соответствующих закону распределения значений «Все чужие» для подбираемых параметров.

Так как задача сильно упрощена, удача при подборе (появления на выходе нейросети кода ключа «Свой») наступает достаточно быстро. Тест повторяют многократно (по 100 раз) для разных входных параметров, среднее число попыток до первой удачи характеризует стойкость биометрической защиты, ослабленной в 100 раз (подбираются 4 параметра из 416).

Для повышения точности оценки стойкости необходимо постепенно увеличивать число подбираемых биометрических параметров до 2, 3, 4, …, 30 %. При этом время подбора и число попыток растут экспоненциально. Эта ситуация отображена на рисунке 3.16.

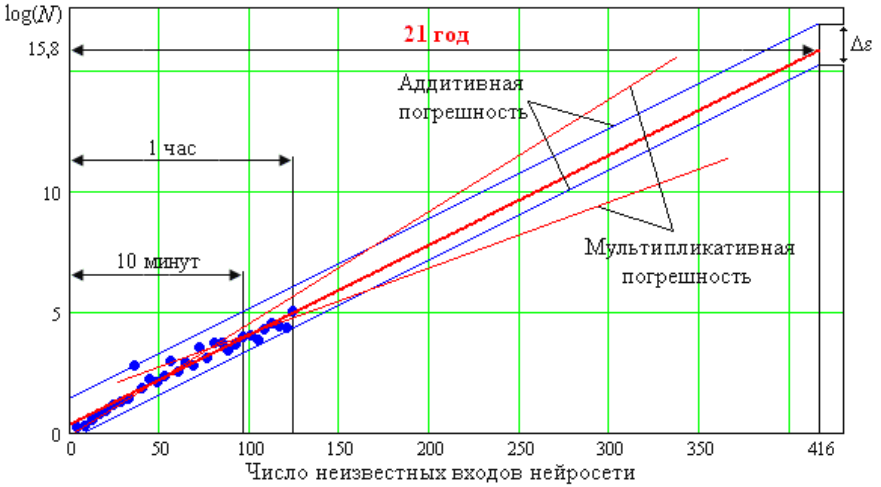


Рисунок 3.16 – Результаты работы тестовой машины,
способной подобрать за 1 ч 120 биометрических параметров
из общего числа – 416 биометрических параметров

Из рисунка 3.16 видно, что на подбор 96 биометрических параметров по 100 раз у тестовой машины, реализованной на Pentium 4 (3 ГГц, ОЗУ 512 Мб), уходит порядка 10 мин. На 100-кратный подбор 120 входных биометрических параметров уходит примерно 1 ч машинного времени. Нетрудно показать, что на 100-кратный подбор всех 416 параметров подобной тестовой машине потребуется 21000 лет.

Следует подчеркнуть, что тестовые машины работают много медленнее атакующих машин подбора из-за многократного повторения подборов. Реальная атакующая машина смогла бы осуществить подбор 416 входных биометрических параметров за время порядка 21 года при ее реализации на ПЭВМ Pentium 4.

Многократное повторение подбора множества сочетаний входных биометрических параметров тестовой машиной нужно, чтобы снизить разброс получаемых данных. Из рисунка 3.16 видно, что 100-кратное усреднение тестовых результатов приводит к хорошей группировке данных возле некоторого линейного тренда. Параметры этого тренда дают оценку стойкости биометрической защиты к атакам подбора $10^{15,8}$ (порядок 15,8).

Аддитивная погрешность подобной оценки строится путем проведения параллельных линии тренду, проходящих через наиболее удаленные от линии тренда отсчеты. Мультипликативная погрешность прогноза может быть оценена разбиением всей тестовой выборки пополам и построением двух других прогнозов по каждой из половин выборки (см. рисунок 3.16).

Следует подчеркнуть, что описанная выше тактика подстановки несвязанных и неупорядоченных случайных данных на входы биометрического преобразователя только на первый взгляд кажется крайне неэффективной. Такая тактика тестирования и реализации атаки намного эффективнее прямого перебора входных биометрических данных. Так, если мы динамический диапазон входных биометрических данных разобьем хотя бы на поддиапазоны и начнем перебирать их состояния, мы получим 10^{126} возможных комбинаций. Перебор такого большого числа комбинаций нецелесообразен. Тактика использования подстановки несвязанных случайных чисел много эффективнее, так как она приводит к удаче примерно через каждые 10^{16} комбинаций. Такая тактика эффективнее прямого перебора на 110 порядков.

Тактика прямого перебора множества размытых, нечетких биометрических секретов оказывается неэффективной из-за того, что число эквивалентных входных дискрет оказывается много меньше числа входов. Дискреты оказываются размыты между входами, одна дискрета приходится примерно на 8 биометрических нечетких входов. Классические приемы перебора состояний криптографических ключей оказываются абсолютно не пригодны при реализации атак подбора биометрических параметров.

Глава 4

Проблемы формирования больших и сверхбольших статистически представительных баз биометрических образов

4.1. Оценка затрат времени и людских ресурсов на формирование больших баз естественных биометрических образов

В связи с тем, что средства высоконадежной биометрической аутентификации могут иметь достаточно высокую стойкость к атакам подбора, желательно иметь большие базы биометрических образов, сопоставимые по своим размерам со стойкостью тестируемых средств [12, 37]. В частности, тестовая машина дает стойкость биометрической защиты к атакам подбора порядка 10^{16}, что вполне достаточно для ряда практически значимых приложений. Как следствие, для тестирования такого средства прямой подстановкой потребуется база биометрических образов, содержащая на два–три порядка больше образов, чем заявленная производителем стойкость биометрической защиты.

Попытаемся оценить затраты времени и людских ресурсов на формирование подобных баз случайных биометрических образов. В случае, если формируется база рукописных образов, то испытуемый, формируя данные, должен выполнить ряд операций:

1) осознать то, что он должен написать;
2) написать рукописное слово (фразу);
3) проконтролировать корректность рукописного ввода;
4) ввести биометрический образ.

Естественно, что все эти операции занимают определенное время. Хронометраж затрат времени показывает, что специально подготовленный человек (время подготовки занимает порядка 40 мин) вводит рукописные образы длиной по 5–6 символов за время порядка 10 с. Затраты времени на похожие операции ввода рукописных символов могут существенно различаться для людей разного темперамента и

разного рода профессий. Естественно, что люди, сталкивающиеся в своей профессиональной деятельности с необходимостью рукописных записей, вводят в ПЭВМ биометрические рукописные образы быстрее.

Для формирования базы из 10^{19} рукописных образов при затратах 10 с на один образ потребуется 10^{20} с. Один год составляет всего миллион секунд (10^6 с). Это означает, что на формирование базы нужных размеров потребуется 10^{14} лет работы без отдыха одного человека. Если к работе привлечь все население миллионного города, то базу нужного размера удастся сформировать через 10^8 лет (через десять миллиардов лет). Естественно, что такие затраты времени и людских ресурсов не могут быть осуществлены, необходимо оценить технически и экономически реализуемые базы биометрических образов.

В Пензенском государственном университете обучается 16 000 студентов (средний срок обучения 5 лет). Привлекая всех студентов для формирования соответствующих биометрических образов (по 8 ч в течение 3 дней за все 5 лет учебы), мы можем получить всего 10^8 (сто миллионов рукописных биометрических образов).

Для организации этой работы потребуется 10, …, 20 ПЭВМ, обслуживающихся одним человеком, которые должны работать без серьезных поломок в течение 5 лет при условии, что каждый студент прошел предварительную 40-минутную подготовку и ПЭВМ оснащены программным обеспечением, автоматически задающим испытуемым графически или голосом то, что им нужно записать. В случае оплаты испытуемым по минимальным нормам их работы стоимость аппаратных затрат на формирование базы биометрических образов оказывается много меньше, чем затраты на привлечение людей.

Отметим, что при формировании биометрических голосовых образов затраты времени и ресурсов оказываются примерно такими же. Голосовая информация вводится быстрее, однако ее приходится вводить больше. Голос обладает несколько меньшей информативностью, если сравнивать между собой одинаковые слова-пароли (парольные фразы).

4.2. Требования к качеству преобразователей биометрических образов физического уровня в биометрические электронные образы

Разнотипные высоконадежные аутентификаторы (распознаватели) во многом сходны. Обобщенная схема устройств высоконадежной биометрической аутентификации личности человека приведена на рисунке 4.1.

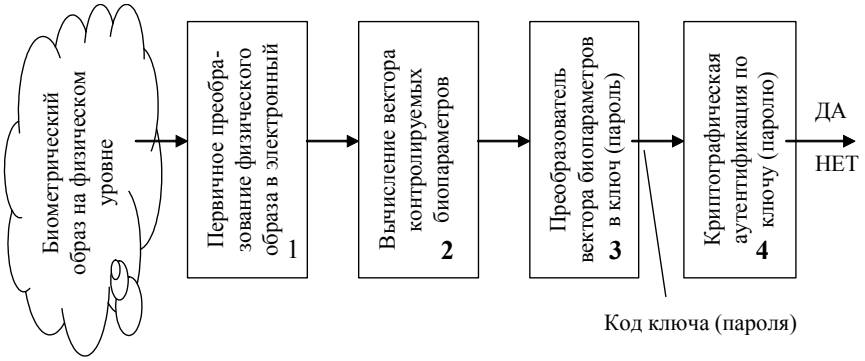


Рисунок 4.1 – Структурная схема типового высоконадежного аутентификатора личности человека

В структурной схеме рисунка 4.1 блок-1 осуществляет преобразование физического нечеткого биометрического образа человека в электронный биометрический нечеткий образ через первичные преобразователи физических величин в электронные цифровые данные. Блок-2 осуществляет нормировку электронных образов и вычисление вектора биометрических параметров, например, в виде коэффициентов Фурье, в средствах аутентификации по динамике воспроизведения рукописного пароля. Блок-3 осуществляет преобразование вектора биометрических параметров в код ключа (пароля) для последующей криптографической аутентификации. Блок-4 осуществляет криптографическую аутентификацию пользователя по его ключу или паролю, выдавая на выход решение «ДА/НЕТ».

Очевидно, что при формировании баз биометрических образов (рукописных, голосовых, отпечатков пальцев, …) желательно использовать именно те датчики, которые использует конкретная система биометрической защиты. Однако такой подход может быть ис-

пользован только при тестировании «слабой» биометрической защиты. Для «слабой» биометрии достаточно формировать тестовые базы, состоящие из малого числа биометрических образов. Как следствие, всегда можно повторить работу по формированию малых баз биометрических образов заново для любого датчика.

При формировании больших баз биометрических образов такой подход неприемлем. Из-за большой трудоемкости этой работы базы должны быть универсальны, т. е. при их формировании необходимо использовать наиболее точные на текущий момент средства ввода рукописной графики и звука. Это делается из-за того, что может возникнуть необходимость тестирования очень точных систем аутентификации, тогда собранная информация будет использована в том виде, в каком хранится. Если потребуется снизить точность преобразований физических биометрических образов в электронные биометрические образы, то это всегда можно сделать с помощью специально написанного программного конвертора. Из хороших данных всегда можно сделать плохие данные. Из плохих данных сделать хорошие данные невозможно.

Таким образом, при формировании больших и универсальных баз биометрических образов требуется использовать как можно более точные и как можно более информативные преобразователи. Для формирования рукописных образов на сегодня подходят для этой цели практически все графические планшеты с размером поля 5×4 дюйма или выше, способные воспринимать 512 уровней давления и имеющие разрешение 40064 линий на дюйм. Требований к шумам и линейности преобразователей не предъявляется. Частота съема информации о положении пера и его нажиме должна быть не менее 20 Гц.

При формировании голосовых баз биометрических образов должна использоваться звуковая карта, способная оцифровывать звук с частотой 44 кГц, в режиме стереозаписи при дискретизации АЦП по уровню не менее 10 разрядов. Должен использоваться широкополосный микрофон, закрепленный на гарнитуре, и второй широкополосный микрофон, закрепленный перед пользователем. При записи голосовых парольных фраз должен быть обеспечен низкий уровень посторонних шумов. Очевидно, что из-за требования к низкому уровню сторонних шумов формировать голосовые базы данных сложнее. Требуются специально оборудованные помещения.

4.3. Требования к программному обеспечению автоматизированного формирования базы биометрических тестовых образов

В связи с высокой трудоемкостью формирования больших баз биометрических образов необходимо стремиться к всемерной экономии затрачиваемых ресурсов. При формировании малых баз возможно привлечение квалифицированного инструктора, под руководством которого каждый пользователь будет выполнять свою работу. Такой подход при формировании больших баз биометрических образов неприемлем, так как существенно увеличивает их стоимость.

При формировании больших баз биометрических образов допустимо привлечение квалифицированного инструктора на этапе обучения тестируемого (40-минутные инструктаж и пробная работа). Далее программное обеспечение ввода биометрических образов должно заменить тестируемого инструктора. Оно должно быть максимально доступно для понимания тестируемым, иметь минимум режимов управления и контролировать все действия тестируемого. При отклонениях тестируемого от заданного оптимального режима программное обеспечение должно автоматически сообщать тестируемому о его ошибке. Например, говорите чуть громче или пишите крупнее и быстрее. Сообщения тестируемому должны выдаваться в голосовой и графической формах.

Особые требования предъявляются к программным средствам формирования парольных слов (фраз). Тестируемый должен воспроизводить только заданные ему образы (рукописные и голосовые). Генераторы парольных слов и фраз программного обеспечения должны иметь возможность взаимной синхронизации из центра. Эти генераторы должны быть способны воспроизводить заданную из центра последовательность слов (фраз) или формировать независимую последовательность случайных слов (фраз).

При формировании рукописных и голосовых парольных образов допускается использование словарей, допускается усиление слов словарей их вариациями, принятыми в языке тестируемого, допускается их усиление числами в форме, принятой при написании или голосовом воспроизведении.

Все данные о поведении тестируемого и формировании его биометрического образа документируются. Программное обеспечение должно вести автоматический журнал регистрации полученной биометрической информации и времени ее получения.

При формировании баз биометрических образов должна быть обеспечена анонимность тестируемых с тем, чтобы их биометрическая информация не могла быть использована кем-либо против них в настоящем или будущем времени.

4.4. Представительность баз биометрических образов

Ценность баз биометрических образов состоит в том, что они хорошо отражают естественное распределение биометрических признаков среди людей. Обычно предполагают, что чем больше база биометрических образов, тем она точнее отражает действительность. Это не всегда так. Вернее, это может быть так, если представители различных возрастных групп людей, различных профессий и различных темпераментов представлены в тестовых группах в тех же пропорциях, что и в обществе. Добиться подобной представительности крайне сложно. В связи с этим необходимо проводить специальные исследования, позволяющие оценить представительность полученной тестовой выборки. Под каждый тип биометрических образов должны быть сформированы свои критерии представительности тестовой выборки.

Если критерии представительности тестовой выборки построены, то представительность самой выборки уже не зависит от ее размеров. В частности, может быть искусственно сформирована малая по размерам тестовая выборка, хорошо удовлетворяющая по ее представительности вектору критериев представительности. Эта выборка должна строиться таким образом, чтобы, с одной стороны, она состояла из реальных или правдоподобно синтезированных биометрических образов, а с другой стороны, она должна верно отражать (отображать с заданной погрешностью) вектор выбранных критериев представительности.

В качестве критериев представительности могут выступать:

- статистические характеристики среднестатистического пользователя по некоторому биометрическому параметру (математическое ожидание, среднеквадратическое отклонение, коэффициенты корреляции и другие статистические моменты);
- статистические характеристики среднего по некоторой группе пользователя, выделенной по некоторому биометрическому параметру (математическое ожидание, среднеквадратическое отклонение, коэффициенты корреляции и другие статистические моменты);
- представительность по численности групп пользователей, классифицированных по некоторому биометрическому параметру.

В качестве биометрических параметров могут быть использованы любые параметры, например: стабильность, уникальность, стойкость к атакам подбора.

4.5. Классификация пользователей по стабильности их биометрических образов

Все реальные биометрические образы обладают некоторой нестабильностью. Именно неоднозначность, нестабильность, размытость биометрических образов является основной проблемой при их преобразовании в однозначный, четкий криптографический ключ [7, 11]. Как следствие, необходимо контролировать нестабильность воспроизведения пользователем его биометрического образа и классифицировать пользователей по их нестабильности. На рисунке 4.2 приведены два распределения одного и того же параметра для множеств «Свой-1» и «Свой-2», обладающих разной стабильностью.

Из рисунка 4.2 видно, что множество «Свой-1» существенно у́же множества «Свой-2». Соответственно стабильность контролируемого параметра для множества «Свой-1» выше, чем стабильность того же параметра для множества возможных значений «Свой-2».

Для того, чтобы оценить стабильность биометрического образа в целом, необходимо найти математическое ожидание всех дисперсий для каждого из контролируемых биометрических параметров. Сравнивая между собой математические ожидания дисперсий разных

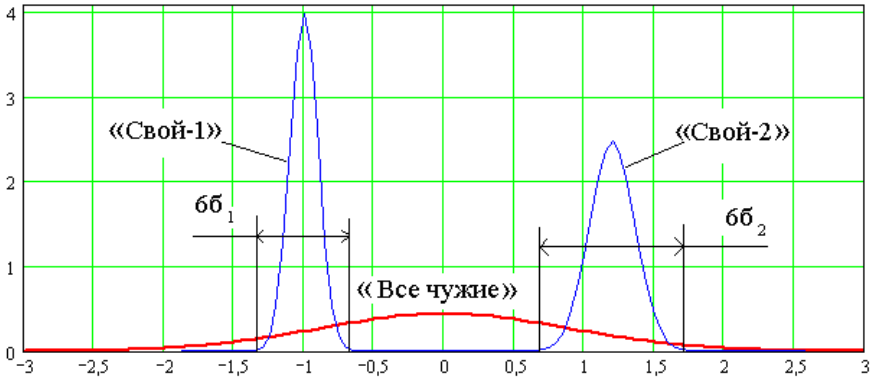


Рисунок 4.2 – Два распределения значений одинаковых биометрических параметров «Свой-1» и «Свой-2» с разными уровнями стабильности их воспроизведения

биометрических образов $m(б_1)$ и $m(б_2)$, мы можем сравнивать стабильность воспроизведения разных биометрических образов.

Следует подчеркнуть, что стабильность воспроизведения биометрических образов является относительной величиной. Она определяется, в первую очередь, вектором контролируемых биометрических параметров. Во вторую очередь, она зависит от умения пользователя стабильно воспроизводить свой биометрический образ. Если это касается воспроизведения рукописных образов, то, прежде чем тренироваться стабильно воспроизводить слово-пароль, желательно попытаться изменить это слово, отыскивая наиболее стабильные для Вашего конкретного почерка сочетания рукописных букв.

Все пользователи для каждой конкретной биометрической системы могут быть классифицированы по стабильности воспроизведения ими их биометрических образов. Для формирования классификации необходимо оценить стабильность нескольких сотен пользователей и построить их нормированное распределение показателя стабильности. Пример такого распределения приведен на рисунке 4.3. Каждый столбец гистограммы соответствует одному из классов стабильности пользователей. Из рисунка 4.3 видно, что реальная гистограмма распределения пользователей по классам смещена в сторону низкостабильных пользователей с повышенными значениями дисперсий. Аппроксимация гистограммы нормальным законом распределения зна-

чений дает достаточно хорошую точность, однако не учитывает естественную асимметрию реальных гистограмм.

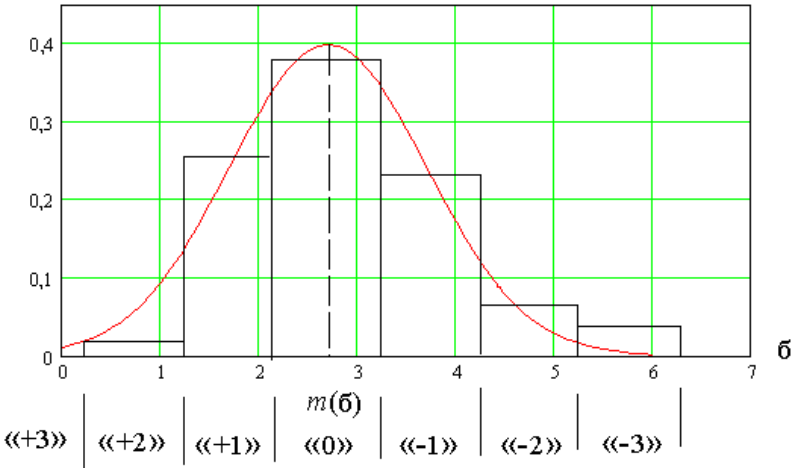


Рисунок 4.3 – Экспериментально полученное распределение пользователей по стабильности их биометрических образов

Классификация пользователей по стабильности осуществляется делением распределения «Все свои» на интервалы, равные среднеквадратическому отклонению так, чтобы в центре интервала оказывалось математическое ожидание. Соответственно мы получаем классификацию стабильности «+3», «+2», «+1» «0», «–1», «–2», «–3».

Очевидно, что стабильность воспроизведения биометрических образов является крайне важным статистическим показателем. Естественно требовать от формируемых баз биометрических образов того, чтобы они хорошо отражали реальную действительность по процентному содержанию биометрических образов, принадлежащих к разным классам стабильности.

При экспериментальном синтезе классифицирующего распределения пользователей по их стабильности (см. рисунок 4.3) следует обратить внимание на то, что оно должно строиться только для дееспособных людей (людей, способных пользоваться биометрической защитой). Недееспособные люди должны выпадать за пределы клас-

сификации (класс «–4»). Как правило, малые дети, старики, люди с нервными расстройствами имеют очень нестабильный рукописный почерк. Система должна предупреждать таких пользователей, что она не в состоянии обеспечить надежную биометрическую защиту.

Еще одним важным применением классификации по стабильности является выявление фактов саботажа (сговора), когда часть пользователей намеренно ослабляет свою биометрическую защиту. Такая ситуация редка при защите пользователями своих интересов, но достаточно часто встречается при защите корпоративных данных. Для прекращения саботажа служащих, как правило, достаточно самого факта его выявления и объяснения служащему негативных для него последствий его же действий (несоответствие занимаемому служебному положению, по квалификации, психологическому состоянию, нелояльность к корпоративной биометрии).

4.6. Классификация пользователей по уникальности их биометрических образов

Не менее важна для систем биометрической защиты степень уникальности биометрического образа «Свой». Очевидно, что злоумышленник всегда будет стараться выбирать при атаках подбора наиболее вероятные состояния входов биометрической защиты, т. е. злоумышленник имеет модель среднестатистического «Своего» (или модель «Все чужие») и будет стараться использовать ее при организации атак. Естественно, что чем больше образ «Свой» будет отличаться от среднестатистического образа, тем выше степень биометрической защиты.

Интуитивно понятно, что уникальность одного биометрического параметра можно оценить через вероятность случайного попадания в интервал «Свой» при эмуляции данных «Все чужие». Эта вероятность в рамках гипотезы нормальности законов распределения значений (см. рисунок 4.2) будет описываться следующим выражением:

$$P_{\text{с}} = \frac{1}{\sqrt{2\pi}\sigma_{\text{В}}} \int_{m_{\text{с}}-3\sigma_{\text{с}}}^{m_{\text{с}}+3\sigma_{\text{с}}} \exp\left(\frac{(m_{\text{В}} - x)^2}{2\sigma_{\text{В}}^2}\right) \text{d}x\,, \qquad (4.1)$$

где $m_{\text{с}}$, $\sigma_{\text{с}}$ – математическое ожидание и дисперсия распределения параметра «Свой»;

$m_В$, $\sigma_В$ – математическое ожидание и дисперсия распределения параметра «Все чужие».

Очевидно, что вероятность (4.1) будет уменьшаться при снижении дисперсии «Своего», а также при вытеснении центра множества «Свой» на периферию распределения «Все чужие».

Уникальность конкретного биометрического параметра будет описываться обратной величиной вероятности (4.1). Так как анализируемых биометрических параметров много, необходимо оценивать среднюю уникальность этих параметров или меру уникальности всего биометрического образа

$$U = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{P_{C,i}}. \qquad (4.2)$$

Уникальность, как и любой иной значимый статистический параметр, может быть использована для классификации биометрических образов. Классификация осуществляется через статистические исследования нескольких сотен биометрических образов по аналогии с процедурами, описанными в предыдущем параграфе.

4.7. Классификация биометрических образов по их относительной информативности

Необходимо отметить, что различные биометрические образы обладают разной информативностью (разной сложностью). Очень простые биометрические образы легко подбираются и не могут обладать сколько-нибудь существенной стойкостью. Выделяют статические – неизменные биометрические образы, данные человеку от рождения [25]. К ним относятся рисунки отпечатков пальцев, дерево сосудов глазного дна, радужная оболочка глаз, геометрия ладони, геометрия лица. Как правило, человек не может по своему желанию изменять (усложнять) свой статический биометрический образ. Как следствие, статический образ легко компрометируется и обладает ограниченной информативностью.

Свои динамические биометрические образы, напротив, человек легко может изменить. Например, рукописный образ слова-пароля легко может быть изменен сменой самого пароля [24]. Динамические биометрические образы могут быть изменены (усложнены) при не-

обходимости. Их можно сохранять в тайне и за счет усложнения повышать их стойкость к атакам подбора. В таблице 4.1 приведены характеристики стойкости наиболее распространенных на сегодня биометрических технологий [11]. Для динамических биометрических образов (рукописный и голосовой почерк, клавиатурный почерк) в таблице 4.1 приводятся только ограничения снизу на длину эквивалентного ключа. Ограничений сверху на этот параметр для этих технологий нет, однако сложность самого биометрического образа и эффективная длина ключа связаны. Эта связь дается в таблице 4.2.

Таблица 4.1 – Рекомендуемые интервалы выбора длины ключей (паролей) при совместном использовании разнотипных биометрических образов

Название биометрической технологии	Стойкость к атакам подбора	Минимальная длина ключа или пароля (бит)	Максимальная длина ключа или пароля (бит)
Анализ кровеносных сосудов глазного дна	От 10^8 до 10^{12}	27	40
Анализ радужной оболочки глаза	От 10^6 до 10^9	20	30
Двухмерный и трехмерный анализ геометрических особенностей лица в видимом и инфракрасном спектрах света	От 10^2 до 10^4	7	14
Анализ особенностей геометрии ушных раковин	От 10^2 до 10^3	7	10
Анализ особенностей голоса	От 10^2 до …	7	Нет ограничений
Анализ особенностей папиллярного рисунка одного пальца	От 10^4 до 10^{13}	12	39
Анализ геометрии ладони, включая рисунки складок кожи ладони и папиллярные рисунки различных фрагментов кожи ладони	От 10^2 до 10^5	7	17
Анализ рисунка кровеносных сосудов, складок кожи тыльной стороны ладони	От 10^2 до 10^3	7	10
Анализ рукописного почерка	От 10^2 до …	7	Нет ограничений
Анализ клавиатурного почерка	От 10^2 до …	7	Нет ограничений
Анализ геометрических соотношений частей тела	От 10^3 до 10^6	10	20

Анализ особенностей походки	От 10^1 до 10^3	4	10

Таблица 4.2 – Рекомендуемые длины ключей (паролей) для среднестатистического пользователя в зависимости от числа букв биометрического пароля или от информативности тайного биометрического образа (данные ФГУП «ПНИЭИ» 2006 г.)

Число букв (цифр) в пароле, образующем биометрический образ без учета пробелов между словами	Длина ключа (пароля), получаемого из рукописного пароля (бит)	Длина ключа (пароля), полученного из голосового пароля (бит)	Длина ключа (пароля), полученного из динамических параметров клавиатурного почерка (бит)
4	32	10	-------
5	40	13	------
6	48	16	------
7	56	18	------
8	64	21	------
9	72	23	------
10	80	26	------
12	96	31	------
14	112	36	-----
16	128	42	7
18	144	47	8
20	160	52	10
24	192	64	11
26	224	76	14
32	256	88	17
36	288	100	20
40	320	112	23

П р и м е ч а н и е 1. В зависимости от стабильности и уникальности биометрического образа конкретного человека длина его ключа может сокращаться в три раза или увеличиваться до трех раз. Рекомендуется уточнять приведенные цифры под каждый конкретный биометрический образ через использование встроенных в биометрическое приложение механизмов тестирования и прогнозирование ожидаемой стойкости.

П р и м е ч а н и е 2. Для преобразователей биометрия/код эффективная длина ключа может составлять 10, …, 30 % от реальной длины выходного биометрического ключа на выходах нейронной сети.

Чем сложнее биометрический образ, тем сложнее его подбор. Это непреложное правило, справедливое для любых (статических или динамических) биометрических образов. В таблице 4.3 дана связь числа особых точек в рисунке отпечатка пальца и его стойкости к атакам подбора.

Таблица 4.3 – Рекомендуемые длины ключей (паролей) для среднестатистического анонимного пользователя в зависимости от числа особых точек в учитываемом фрагменте рисунка отпечатка пальца (данные ФГУП «ПНИЭИ» 2005 г.)

Число особенностей в учитываемом фрагменте рисунка отпечатка	Вероятность удачи при подборе с первой попытки	Рекомендуемая длина бинарного кода ключа
16	$10^{-5,6}$	17
18	$10^{-6,3}$	19
20	10^{-7}	21
22	$10^{-7,7}$	23
24	$10^{-8,4}$	25
26	$10^{-9,1}$	27
28	$10^{-9,8}$	29
30	$10^{-10,5}$	31
32	$10^{-11,2}$	33
34	$10^{-11,9}$	35
36	$10^{-12,6}$	37
38	$10^{-13,3}$	39

Примечание 1. Компрометация рисунка отпечатка пальца, например, путем снятия его с поверхности датчика после прохода «Своего», снижает стойкость к атакам подбора практически до нуля.

В силу того, что стойкость биометрических образов прямо зависит от их сложности и эта сложность может быть достаточно просто оценена, необходимо при формировании больших баз биометрических образов их балансировать по сложности биометрических образов, т. е. базы рукописных и голосовых биометрических образов должны содержать число слов из 5 букв (число отпечатков пальцев с 22 особенностями) в процентном отношении столько же, сколько их содержится в естественном языке (в естественном распределении рисунков отпечатков пальцев).

4.8. Классификация биометрических образов по их стойкости к атакам подбора

Так как мы имеем достаточно быстрые процедуры прогнозирования стойкости биометрической защиты к атакам подбора (см. параграф 3.10, рисунок 3.16), мы можем осуществлять классификацию биометрических образов по их стойкости к атакам подбора. Технология классификация прежняя. Необходимо использовать несколько сотен биометрических образов одинаковой сложности, оценить их стойкость к атакам подбора и вычислить статистические моменты распределения прогнозируемой стойкости.

Первый момент (математическое ожидание), видимо, будет являться одной из важнейших характеристик средства биометрической защиты, объявляемой его производителем. Второй статистический момент (дисперсия) является необходимым инструментом классификации пользователей и прогноза ожидаемого интервала разброса этого контролируемого параметра.

По требованиям национального стандарта [11] средства высоконадежной биометрической аутентификации должны сообщать пользователю о прогнозируемой стойкости его биометрического образа, кроме того, крайне желательно дополнительно доводить до пользователя место его биометрического образа в классификации. Для пользователя крайне важно знать ответы на следующие вопросы:

1. Хуже или лучше стойкость его биометрического образа по отношению к заявленной производителем среднестатистической стойкости?

2. Насколько классов его биометрический образ хуже или лучше среднестатистической стойкости?

Проблема состоит в том, что для неспециалистов по защите информации показатели степени стойкости являются трудно воспринимаемой величиной. По этой причине основные показатели системы защиты (и в том числе стойкость к атакам подбора) должны доводиться до потребителя в понятной ему форме, в частности, в форме классификации хуже/лучше среднего, насколько хуже/лучше, что такое средняя стойкость (сколько лет на обычной машине злоумышленник должен подбирать тайный биометрический образ).

4.9. Корректное снижение размеров баз реальных биометрических образов при сохранении их высокой представительности

Предположим, что удалось создать достаточно надежные механизмы классификации биометрических образов по их стойкости к атакам подбора. Легко показать, что на это должны быть затрачены достаточно большие материальные (вычислительные) ресурсы. При проведении этой работы, скорее всего, просто прогнозов вычислительной стойкости (см. параграф 3.10, рисунок 3.16) будет недостаточно. Переход от прогнозов к реальным проверкам стойкости сразу же многократно увеличивает материальные затраты.

Естественно, что затраты больших материальных ресурсов на проверки для малых фирм непосильны. Получается, что малые фирмы не могут участвовать в конкурентной борьбе за рынок высоконадежных средств биометрической аутентификации. Это может крайне негативно сказаться на уровне защищенности больших и сверхбольших систем.

Для того, чтобы предоставить равные возможности в конкурентной борьбе, государство или иная авторитетная организация должны взять на себя ответственность за формирования относительно небольших, но достоверно отражающих действительность «эталонных» баз биометрических примеров.

Например, эталонная база рукописных биометрических образов, содержащих по 5 букв (символов), может состоять всего из 7 образов (каждый образ представлен 20 примерами). При этом организация – держатель «эталона» – должна гарантировать, что:

- 1-й образ с погрешностью ±5 % находится в центре самого стойкого класса «+3»;
- 2-й образ с погрешностью ±5 % находится в центре стойкого класса «+2»;
- 3-й образ с погрешностью ±5 % находится в центре стойкого класса «+1»;
- 4-й образ с погрешностью ±5 % находится в центре среднего класса «0»;

- 5-й образ с погрешностью ±5 % находится в центре нестойкого класса «–1»;
- 6-й образ с погрешностью ±5 % находится в центре нестойкого класса «–2»;
- 7-й образ с погрешностью ±5 % находится в центре самого нестойкого класса «–3».

Имея такую малую эталонную базу биометрических образов, малые производители смогут существенно сэкономить свои ресурсы, идущие на биометрические проверки. При наличии такой базы нет необходимости проверять, сравнивать, классифицировать тысячи (миллионы) биометрических образов. Эта большая работа уже проделана, и малая «эталонная» база является представительной по заданному критерию представительности. Может потребоваться создание своей малой «эталонной» базы биометрических образов под каждую из биометрических технологий и по каждому из критериев представительности.

Пока вопрос о числе малых баз биометрических образов и способе их распространения открыт. Однако совершенно ясно, что организация честной и эффективной конкуренции на рынке высоконадежной биометрии невозможна без наличия общедоступных малых «эталонных» баз биометрических образов.

Г л а в а 5

Умножение размеров баз биометрических образов через формирование дополнительных синтетических образов

5.1. Синтез искусственных биометрических образов «Свой» и «Чужой»

В связи с тем, что реально воспроизведенных на физическом уровне биометрических образов «Чужой» не может быть собрано достаточно для полного тестирования средств высоконадежной биометрии, необходимо реальные базы тестовых образов дополнять синтетическими биометрическими образами.

Для построения генератора синтетических биометрических образов «Чужие» необходимо иметь достаточно большую базу естественных биометрических образов и путем ее статистических исследований найти статистические законы распределения биометрических параметров ее образов. Подобные исследования, проводимые «Испытательной лабораторией биометрических устройств и технологий» при факультете военного обучения Пензенского государственного университета показали, что биометрические параметры рукописных образов «Чужие» имеют:

- законы распределения значений, близкие к нормальным;
- близкие к нулю математические ожидания;
- существенно отличающиеся дисперсии для разных биометрических параметров;
- случайные коэффициенты малой корреляции с нулевым математическим ожиданием и малым значением дисперсии.

На рисунке 5.1 изображена обобщенная структура генератора синтетических биометрических образов.

В качестве генератора случайных чисел-1 может быть использован любой достаточно качественный программный [39] или аппарат-

ный генератор [40]. Коррелятор-2 может быть выполнен по разным схемам [41, 42], более подробно вопросы синтеза корреляторов будут обсуждаться в следующих параграфах. Масштабирующий преобразователь-3 необходим для приведения выходной дисперсии каждого биометрического параметра к заданным значениям. Блок-4 осуществляет заданное заранее смещение плотности распределений значений. Блок-5 осуществляет генерирование случайных векторов коэффициентов корреляций, математических ожиданий и дисперсий.

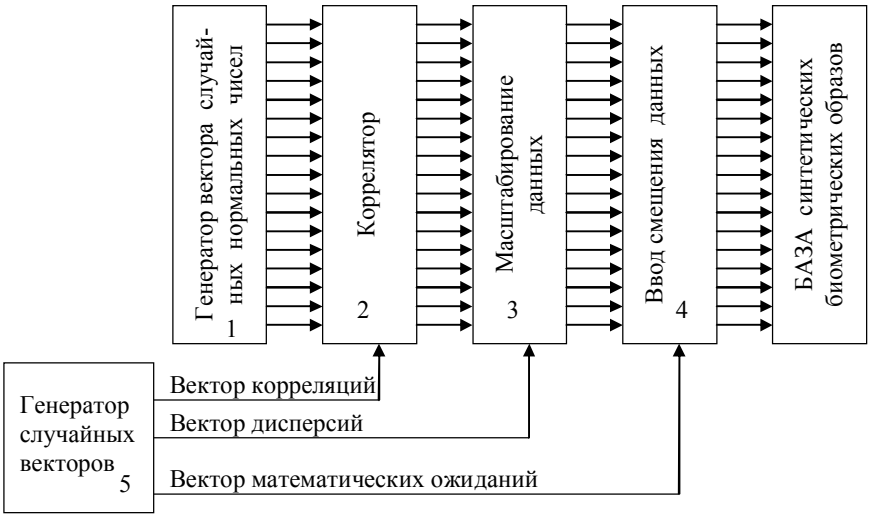


Рисунок 5.1 – Структурная схема генератора синтетических биометрических образов

Псевдослучайный генератор векторов-5 должен быть сбалансирован так, чтобы в итоге статистические параметры множества синтетических образов «Все чужие» совпали со статистическими параметрами естественных биометрических образов «Все чужие». Балансировка этого генератора-5 будет обсуждаться позднее.

Следует подчеркнуть, что введение в структурную схему блока-2 и блока-4 необязательно для упрощенных синтезаторов относительно малого числа образов «Чужой». Появление этих блоков необходимо только в достаточно точных синтезаторах образов «Чужой-*N*», воспроизводящих заданное слово, или в синтезаторах образов «Свой».

Для образов «Свой» характерны значительные величины математических ожиданий (см. рисунок 4.3) и значительные корреляционные связи между параметрами.

5.2. Простейшее размножение образов «Свой», «Чужой» размыванием одного образа

В связи с тем, что ручной способ формирования биометрических образов дорог и обеспечивает слишком низкую производительность (скорость формирования больших баз крайне низка), необходимо попытаться увеличить число формируемых образов в единице времени на несколько порядков. Один из способов искусственного размножения близких биометрических образов отображен на рисунке 5.2.

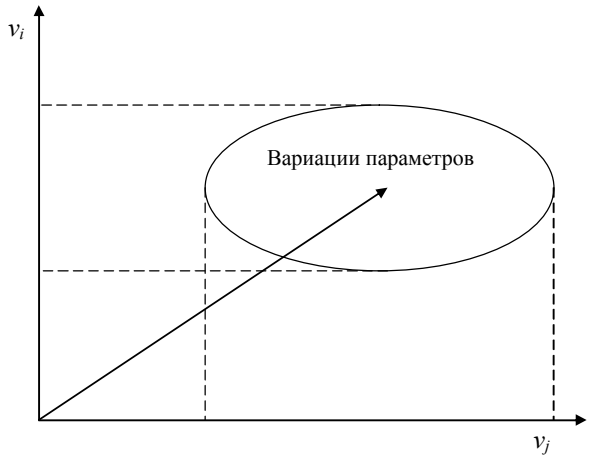


Рисунок 5.2 – Синтез множества близких биометрических образов «Свой», «Чужой-К» добавлением случайных данных с малой дисперсией

По этому способу число биометрических образов многократно увеличивается путем добавления к полученным биометрическим параметрам некоторого незначительного шума. Амплитуду шума (его дисперсию) следует выбирать сопоставимую с естественной дисперсией группы одинаковых рукописных образов, воспроизводимых одним человеком.

Размножая таким образом образы «Свой», мы несколько снижаем трудозатраты пользователя на обучение биометрической системы. Формально можно обучать систему всего на одном примере биометрического образа «Свой», однако подобная экономия не оправдана. Используя только один образ «Свой», можно сильно ошибиться, искусственно размножая примеры. Особенно велики ошибки при случайном попадании первого и единственного образа на периферию множества образов, тогда искусственное размножение ошибочно делает периферию центром множества. Еще одна проблема состоит в выборе размеров искусственного размножения. Корректное размножение образов «Свой» может быть только в случае, когда заранее известен центр этого множества и дисперсии по каждому из параметров.

Перечисленные выше проблемы возникают только при размножении образов «Свой». При размножении образов «Чужой» таких проблем не возникает, так как требований к их корректной группировке возле некоторого центра не возникает. Это означает, что применительно к увеличению размеров тестовой базы «Чужие» описанный выше прием вполне пригоден. Число допустимого увеличения размеров числа примеров образов (число размножения) определяется практически. Интуитивно понятно, что чем выше размерность (информативность) биометрического образа, тем больше требуется увеличивать число примеров. Из-за роста размерности пространства это, видимо, можно осуществить, однако должны существовать и ограничения на допустимое значение «числа размножения». На данный момент эти ограничения не установлены. В этом направлении в ближайшее время должны быть осуществлены соответствующие аналитические и численные исследования.

5.3. Синтетическое размножение образов «Свой» и «Чужой» через равномерное и псевдослучайное заполнение промежутков между соседями

Предположим, что мы получили два разных естественных биометрических образа. Очевидно, что многомерное пространство между ними может быть равномерно заполнено линейными комбинациями этих образов с некоторым шагом (шаг или число промежуточ-

ных образов в общем случае выбирается произвольно). Ситуация равномерного заполнения промежуточными образами гиперпространства отображена на рисунке 5.3.

Для синтеза равномерной гиперсетки сетки из (k – 1) промежуточных образов необходимо разницу между каждыми значениями № 1 и 2 по каждому параметру разбить на k частей и, задавшись направлением движения (например, от образа № 1 к образу № 2), восстановить все промежуточные значения:

$$v_{i,\xi} = \frac{k-\xi}{k} v_{i,1} + \frac{\xi}{k} v_{i,2} \quad \text{при} \quad \xi = 1,2,3,...,(k-1) . \qquad (5.1)$$

Очевидно, что промежуточные образы могут быть построены не только с равномерным направленным шагом. Могут быть и другие способы построения гиперсеток. Например, можно произвольно выбирать направление движения от одного образа к другому. При этом мы получим (k – 1) (N!) вариантов разных промежуточных биометрических образов (где N – размерность анализируемого вектора биометрических параметров).

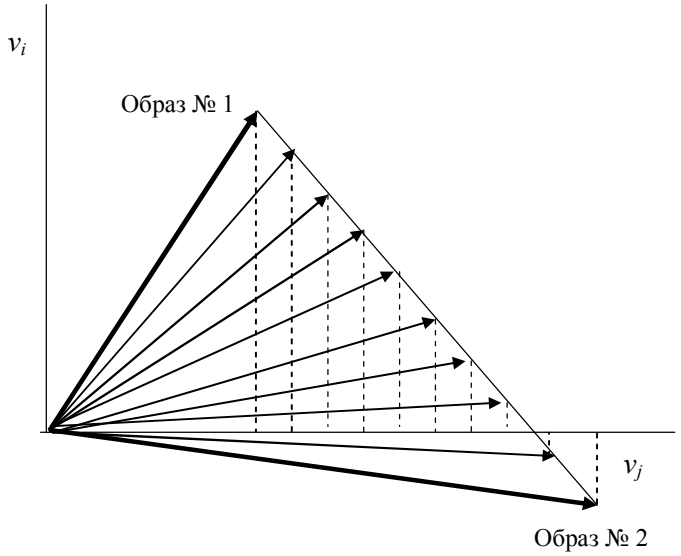


Рисунок 5.3 – Равномерное преобразование одного биометрического образа в другой

Еще одним интересным способом формирования промежуточных биометрических образов является случайный выбор значения параметра ξ из ряда допустимых значений 1, 2, 3,…, (k – 1). В этом случае мы будем иметь заполнение случайными образами гиперпространства между опорными реальными образами.

Очевидно, что при синтезе промежуточных образов мы можем отказаться от равномерного шага гиперсетки, сделав шаг по каждому из параметров случайным. В этом случае ξ будет являться случайной величиной, например, с равномерным законом распределения значений в интервале от 0 до 1, тогда промежуточные значения параметров будут описываться следующей формулой:

$$v_{i,\xi} = (1 - \xi_i)\, v_{i,1} + \xi_i\, v_{i,2}\,. \qquad (5.2)$$

Вариант размножения сигналов в соответствии с (5.2) имеет существенно больший потенциал, по сравнению с вариантами с равномерным шагом. Пользуясь случайным шагом, мы можем получить существенно большее число промежуточных биометрических образов.

5.4. Синтетическое размножение биометрических образов через перестановки

5.4.1. Синтетическое размножение биометрических образов через перестановки фрагментов «сырых», необработанных биометрических данных

Главной проблемой синтеза больших и сверхбольших баз естественных биометрических образов является низкая производительность людей и соответственно высокая стоимость технологии сбора биометрических образов. Обычный человек под диктовку способен воспроизводить рукописно порядка 20, …, 40 букв за время 20 с, т. е. скорость ввода составляет не более 1, 2 символов в секунду. На рукописное воспроизведение словаря паролей из 100 000 слов средней длиной по 8 букв один пользователь потратит 160 000 с, или 444 ч, или около 60 рабочих дней по 8 ч.

Постоянное писание под диктовку с небольшими перерывами в течение нескольких часов – это тяжелый ручной труд, который дол-

жен достаточно хорошо оплачиваться. Кроме того, этим трудом могут заниматься только профессионально подготовленные люди (сегодня это студенты, вынужденные писать конспекты).

Для того, чтобы снизить трудозатраты на формирование больших баз биометрических образов, можно пойти по пути формирования рукописных фрагментов, соответствующих каждому из рукописных вариантов букв алфавита и цифр. Испытуемый вводит образцы своего рукописного почерка на примерах написания конкретных букв и их сочетаний. В качестве примера такой операции на рисунке 5.4 приведены образцы рукописного написания нескольких первых букв кириллического алфавита.

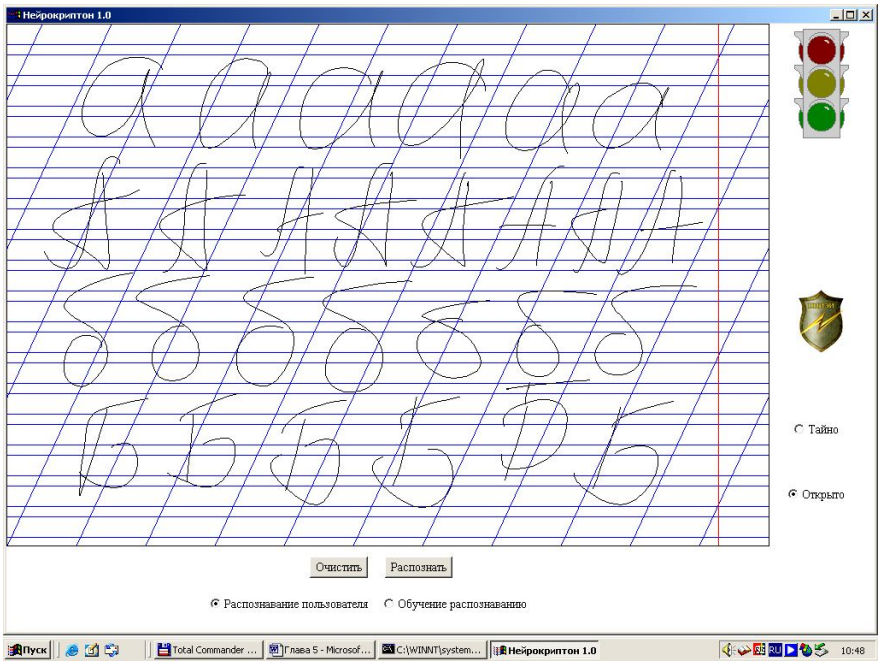


Рисунок 5.4 – Пример формирования биометрических рукописных образов отдельных букв кириллического алфавита

Если придерживаться подобной тактики, то через сравнительно короткий интервал времени мы получим достаточно представительную базу простейших образцов почерка человека. Имея такую базу,

мы далее можем создать автомат, формирующий из нее слова-пароли по заданному словарю. При этом на формирование рукописных образов букв и цифр – примерно 30 мин, остальная работа по формированию больших и сверхбольших словарей парольных слов (фраз) проделывается автоматически. Это позволяет существенно сократить трудозатраты на формирование практически естественных баз биометрических образов больших и сверхбольших размеров.

Следует отметить, что в автоматах синтеза баз вопросы стыковки отдельных рукописных букв должны решаться через использование аппроксимации точек их соединения конец/начало сплайнами. В этом отношении синтез рукописных баз большого и сверхбольшого размера осуществляется много проще синтеза баз голосовых парольных фраз. К сожалению, сформировать голосом эталоны фонем не удается. Фонемы необходимо вырезать из слитно произнесенных слов (фраз). Пока эта процедура не доступна искусственному интеллекту, однако технологии быстро развиваются, и через некоторое время описанный выше подход может оказаться применим и для голосовых технологий.

5.4.2. Синтетическое размножение биометрических образов через перестановки групп векторов контролируемых параметров

В том случае, когда не удается осуществлять корректную стыковку элементарных биометрических образов на физическом уровне (например, при формировании голосовых баз), эта стыковка может быть осуществлена на уровне биометрических параметров фрагментов этих образов. На рисунке 5.5 отображена голосовая фраза с ее кадровым и фонемным разбиением.

Из рисунка 5.5 видно, что на каждую фонему попадает несколько кадров дискретной обработки звука. Если биометрическая речевая информация кодируется вектором из 20 параметров, то каждая из фонем парольной фразы будет описываться 3, …, 6 векторами по 20 параметров. Возможен синтез автомата, выделяющего фонемоподобные фрагменты речи, который каждому выделенному фрагменту будет ставить в соответствие несколько векторов контролируемых биометрических параметров.

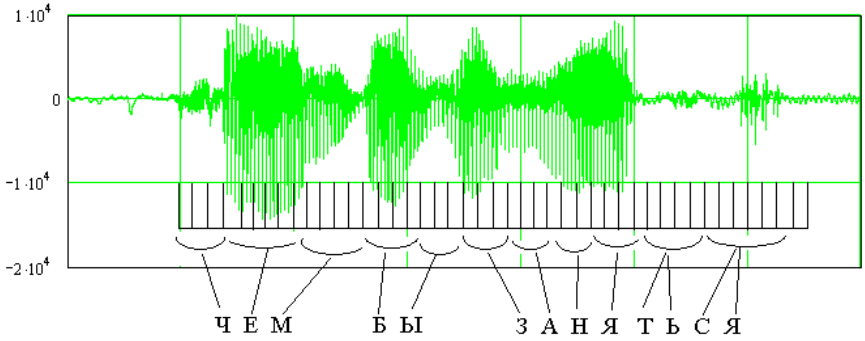


Рисунок 5.5 – Парольная фраза с ее кадровым и фонемным разбиением

Мы не умеем «сшивать» между собой фонемы, произнесенные разными людьми или одним человеком, однако мы легко можем «сшивать» их векторные описания. Для этого достаточно осуществлять стыковку групп векторов биометрических параметров по значениям. Состыковываться и приводиться друг к другу должны амплитуда, период основного тона и другие параметры. После грубой стыковки осуществляется сплайн-аппроксимация одноименных параметров вектора, исключающая дефекты производной первого порядка.

В отличие от предыдущего случая стыковки фрагментов биометрических образов на физическом уровне, осуществляется стыковка групп векторов биометрических параметров. Очевидно, что иметь дело с биометрическими векторами конечной длины много проще, чем с гораздо более сложным естественным звуковым сигналом.

5.5. Генераторы векторов зависимых случайных данных

5.5.1. Генераторы с равнокоррелированными выходными данными

Получить вектор независимых случайных параметров с нормальным законом распределения несложно. Для этого можно воспользоваться любым из известных вариантов программных или аппаратных генераторов случайных чисел [39, 40]. Для того, чтобы сделать данные зависимыми, необходимо их вектор умножить на некоторую

матрицу связанности [41]. Например, может использоваться матрица, состоящая из единиц с одинаковыми числами на диагонали:

$$\begin{bmatrix} a & 1 & :: & 1 \\ 1 & a & :: & 1 \\ :: & :: & :: & :: \\ 1 & 1 & :: & a \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ :: \\ x_n \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ :: \\ v_n \end{bmatrix} \quad \Rightarrow \quad \left[r_{V_i V_j} \right] = \begin{bmatrix} 1 & r & :: & r \\ r & 1 & :: & r \\ :: & :: & :: & :: \\ r & r & :: & 1 \end{bmatrix}, \qquad (5.3)$$

где x_i – случайные числа, полученные от генератора с нормальным законом распределения значений, нулевым математическим ожиданием и единичной дисперсией; a – регулируемый параметр, влияющий на значение коэффициентов корреляции – r.

Зависимость коэффициента корреляции от параметра «a» при разной размерности генерируемого вектора приведена на рисунке 5.6.

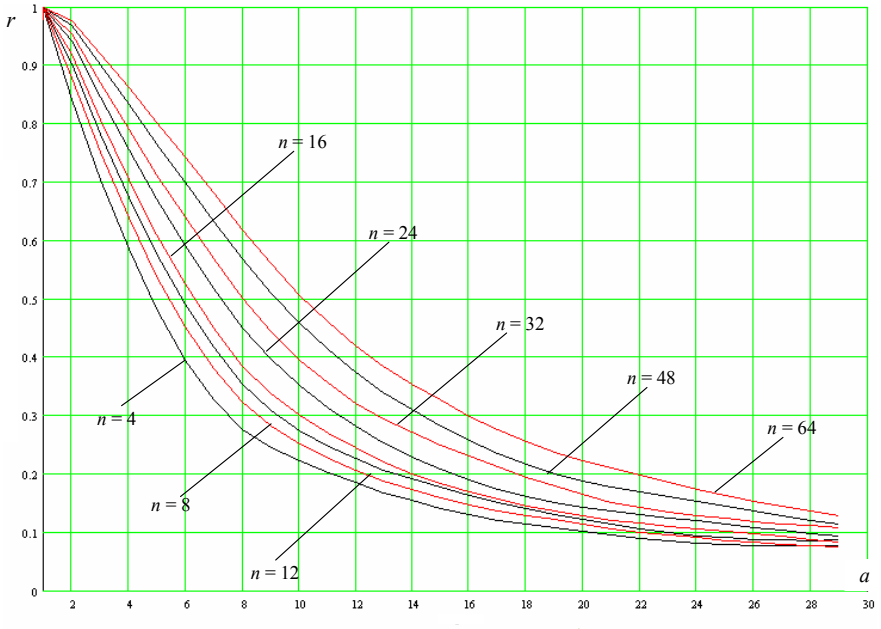


Рисунок 5.6 – Номограмма связи коэффициента – r с параметром «a»

Можно показать, что описанная выше процедура генерации зависимых данных дает вектор с одинаковыми среднеквадратическими отклонениями

$$\sigma(v_i) = \sqrt{n - 1 + a^2} \tag{5.4}$$

и нулевыми математическими ожиданиями для каждой компоненты v_i.

Кажется, что корреляционная матрица с одинаковыми коэффициентами корреляции является слишком простой, однако она оказывается очень удобной при моделировании. Более сложные корреляционные матрицы получаются незначительными вариациями процедуры (5.3).

5.5.2. Генераторы с равнокоррелированными по модулю выходными данными

Необходимо отметить, что биометрические данные с одинаковыми по знаку корреляционными функциями не встречаются на практике. В связи с этим необходимо усовершенствовать механизм синтеза зависимых данных с тем, чтобы он позволял получать более сложные данные с корреляционными матрицами, в которых знаки коэффициентов корреляции меняются.

Для этой цели необходимо использовать матрицу преобразования исходных независимых случайных данных со строками, имеющими разный знак:

$$\begin{bmatrix} a & 1 & :: & 1 \\ -1 & -a & :: & -1 \\ :: & :: & :: & :: \\ 1 & 1 & :: & a \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ :: \\ x_n \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ :: \\ v_n \end{bmatrix} \quad \Rightarrow \quad \left[r_{v_i v_j}\right] = \begin{bmatrix} 1 & -r & :: & r \\ -r & 1 & :: & r \\ :: & :: & :: & :: \\ r & r & :: & 1 \end{bmatrix}. \tag{5.5}$$

Заметим, что такого же эффекта можно добиться, если сменить знак перед синтезируемыми компонентами вектора:

$$\begin{bmatrix} a & 1 & :: & 1 \\ 1 & a & :: & 1 \\ :: & :: & :: & :: \\ 1 & 1 & :: & a \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ :: \\ x_n \end{bmatrix} = \begin{bmatrix} v_1 \\ -v_2 \\ :: \\ v_n \end{bmatrix} \quad \Rightarrow \quad \left[r_{v_i v_j}\right] = \begin{bmatrix} 1 & -r & :: & r \\ -r & 1 & :: & r \\ :: & :: & :: & :: \\ r & r & :: & 1 \end{bmatrix}. \tag{5.5a}$$

Выражения (5.5) и (5.5а) тождественны. Если требуется получить некоторый рисунок распределения знаков по корреляционной матрице, то его можно добиться, перебирая все возможные сочетания знаков в свободной части уравнения (5.5а). В частности, периодическое изменение знака каждого следующего элемента синтезируемого вектора даст корреляционную матрицу с чередующимися по знаку коэффициентами корреляции.

При реализации процедуры со случайным распределением знаков коэффициентов корреляции необходимо правую часть матричного уравнения (5.5а) домножить на $\varepsilon_i = \pm 1$ или

$$\begin{bmatrix} a & 1 & :: & 1 \\ 1 & a & :: & 1 \\ :: & :: & :: & :: \\ 1 & 1 & :: & a \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ :: \\ x_n \end{bmatrix} = \begin{bmatrix} \varepsilon_1 \cdot v_1 \\ \varepsilon_2 \cdot v_2 \\ :: \\ \varepsilon_n \cdot v_n \end{bmatrix}, \qquad (5.6)$$

где ε_i – числа ± 1 со случайным знаком.

5.5.3. Генераторы с положительно коррелированными, но случайно коррелированными по значению выходными данными

Для изменения коэффициента корреляции в одном столбце или строке корреляционной матрицы (при сохранении одинаковыми всех других коэффициентов корреляции) достаточно к одному из диагональных элементов матрицы преобразования прибавить положительное приращение Δa. Например, может быть изменен второй элемент матрицы преобразования

$$\begin{bmatrix} a & 1 & 1 & 1 & 1 \\ 1 & a+\Delta a & 1 & 1 & 1 \\ 1 & 1 & a & 1 & 1 \\ 1 & 1 & 1 & a & 1 \\ 1 & 1 & 1 & 1 & a \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & b & r & r & r \\ b & 1 & b & b & b \\ r & b & 1 & r & r \\ r & b & r & 1 & r \\ r & b & r & r & 1 \end{bmatrix}, \qquad (5.7)$$

где $b < r$.

Связь сниженных значений «*b*» в строках и столбцах корреляционной матрицы [8×8] с регулируемым параметром «Δ*a*» для разных значений «*a*» отражена на рисунке 5.7.

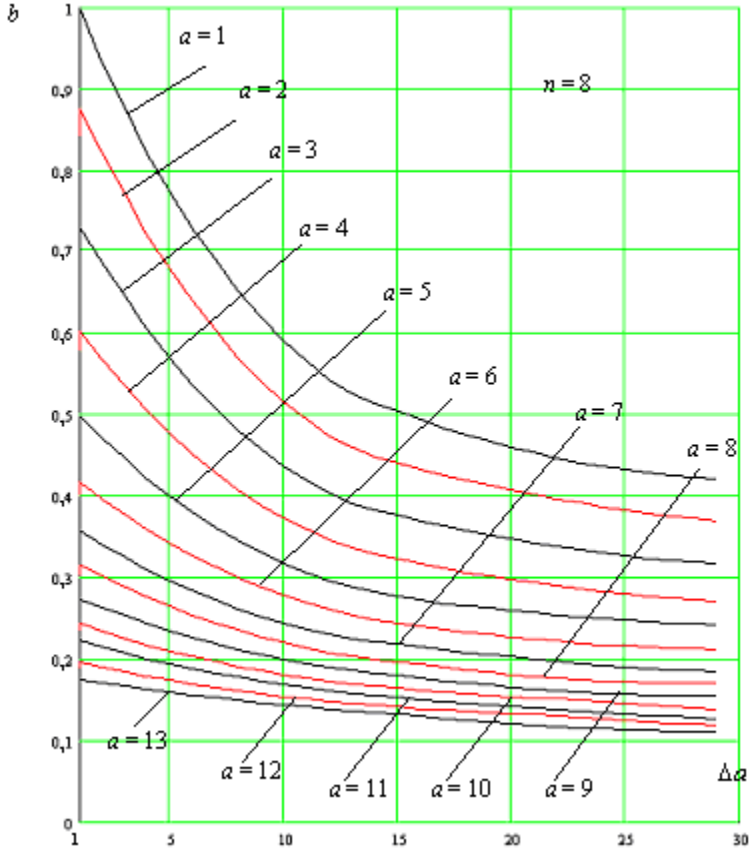


Рисунок 5.7 – Связь заниженного коэффициента корреляции в строке и столбце корреляционной матрицы с регулируемым параметром «Δ*a*» для матрицы размерности 8×8

Связь сниженных значений «*b*» в строках и столбцах корреляционной матрицы с размерами самой корреляционной матрицы дана на рисунке 5.8 для фиксированного значения $a = 5$.

Исходя из изложенного выше, можно показать, что имитация случайного распределения значений коэффициентов корреляции био-

метрических данных может быть достигнута случайным выбором значений диагональных элементов матрицы преобразований с единицами вне диагонали:

$$\begin{bmatrix} a_1 & 1 & 1 & 1 & 1 \\ 1 & a_2 & 1 & 1 & 1 \\ 1 & 1 & a_3 & 1 & 1 \\ 1 & 1 & 1 & a_4 & 1 \\ 1 & 1 & 1 & 1 & a_5 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{12} & 1 & r_{23} & r_{24} & r_{25} \\ r_{13} & r_{23} & 1 & r_{34} & r_{35} \\ r_{14} & r_{24} & r_{34} & 1 & r_{45} \\ r_{15} & r_{25} & r_{35} & r_{45} & 1 \end{bmatrix}, \quad (5.8)$$

где a_i – случайные значения коэффициентов, которые всегда больше единицы, но меньше некоторого максимального числа.

Минимальное значение a_i находится по номограмме рисунка 5.7 и должно соответствовать максимальному значению отмеченного в имитируемой выборке коэффициента корреляции. Максимальное значение a_i находится по номограммам рисунков 5.7, 5.8 и должно

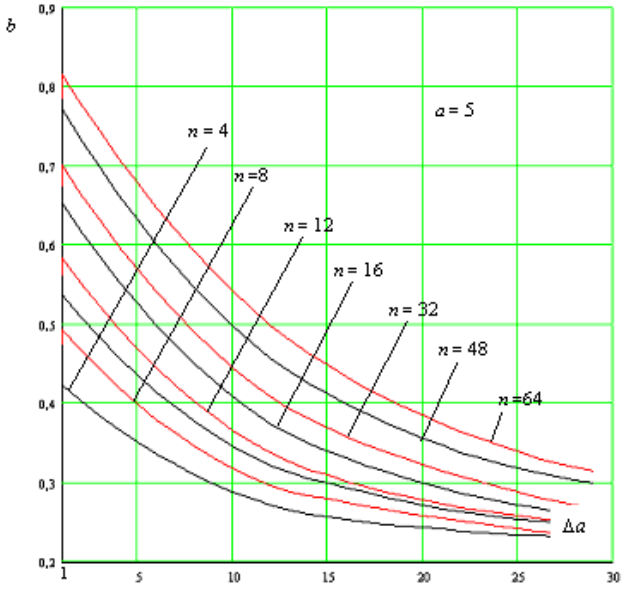


Рисунок 5.8 – График связи значений заниженных коэффициентов корреляции в одном из столбцов и в одной из строк с размерностью самой корреляционной матрицы

соответствовать минимальному значению отмеченного в имитируемой выборке коэффициента корреляции.

5.5.4. Формирование зависимых данных со случайными дисперсиями и случайной знакопеременной матрицей коэффициентов корреляции

Одним из недостатков описанного выше способа формирования зависимых данных является то, что параметры вектора имеют близкие значения дисперсий. Вторым недостатком способа является постоянство знаков коэффициентов корреляции. Оба этих недостатка исчезают, если умножить вектор зависимых данных на случайные числа χ_i

$$\begin{bmatrix} a_1 & 1 & 1 & 1 & 1 \\ 1 & a_2 & 1 & 1 & 1 \\ 1 & 1 & a_3 & 1 & 1 \\ 1 & 1 & 1 & a_4 & 1 \\ 1 & 1 & 1 & 1 & a_5 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} v_1 \cdot \chi_1 \\ v_2 \cdot \chi_2 \\ v_3 \cdot \chi_3 \\ v_4 \cdot \chi_4 \\ v_5 \cdot \chi_5 \end{bmatrix}, \tag{5.9}$$

где χ_i – однократно полученные случайные числа от генератора нормальных чисел с нулевым математическим ожиданием и единичной дисперсией.

Необходимость использования в (5.9) именно случайных чисел с нормальным законом распределения значений обусловлена тем, что в этом случае хороших данных с малыми значениями среднеквадратических отклонений оказывается меньше, чем плохих данных с большими значениями среднеквадратических отклонений. Именно эта ситуация характерна для биометрических данных.

Таким образом, процедура (5.9) позволяет достаточно просто имитировать зависимые биометрические данные. При имитации образов «Свой» необходимо однократно получить случайные значения векторов:

– $(m_1, m_2, \ldots, m_n)^{\mathrm{T}}$ –математических ожиданий;

– $(a_1, a_2, \ldots, a_n)^{\mathrm{T}}$ – задания значений матрицы коэффициентов корреляции;

– $(\chi_1, \chi_2, \ldots, \chi_n)^{\text{т}}$ – задания значений дисперсий данных и распределения знаков коэффициентов корреляций.

Далее для формирования N примеров образов «Свой» необходимо получить от генератора случайных чисел N векторов $(x_1, x_2, \ldots, x_n)^{\text{т}}$ и воспользоваться преобразованием по (5.9) в вектор $(v_1, v_2, \ldots, v_n)^{\text{т}}$. Далее вектор $(v_1, v_2, \ldots, v_n)^{\text{т}}$ необходимо сложить с вектором математических ожиданий $(m_1, m_2, \ldots, m_n)^{\text{т}}$.

При формировании образов «Чужой», знающий пароль, перечисленные выше операции повторяются с той лишь разницей, что вектор математических ожиданий $({}^{\text{ч}}m_1, {}^{\text{ч}}m_2, \ldots, {}^{\text{ч}}m_n)^{\text{т}}$ должен получаться из аналогичного вектора множества «Свой». Остальные векторы $({}^{\text{ч}}a_1, {}^{\text{ч}}a_2, \ldots, {}^{\text{ч}}a_n)^{\text{т}}$, $({}^{\text{ч}}\chi_1, {}^{\text{ч}}\chi_2, \ldots, {}^{\text{ч}}\chi_n)^{\text{т}}$ формируются независимо от аналогичных векторов, задающих множество образов «Свой».

При формировании образов «Чужой», не знающих пароля, все векторы $({}^{\text{ч}}m_1, {}^{\text{ч}}m_2, \ldots, {}^{\text{ч}}m_n)^{\text{т}}$, $({}^{\text{ч}}a_1, {}^{\text{ч}}a_2, \ldots, {}^{\text{ч}}a_n)^{\text{т}}$, $({}^{\text{ч}}\chi_1, {}^{\text{ч}}\chi_2, \ldots, {}^{\text{ч}}\chi_n)^{\text{т}}$, $({}^{\text{ч}}x_1, {}^{\text{ч}}x_2, \ldots, {}^{\text{ч}}x_n)^{\text{т}}$ формируются независимо для каждого конкретного примера. Такая процедура имитации оказывается наиболее экономичной по потребляемым вычислительным ресурсам, что облегчает тестирование стойкости нейросетевых систем хранения конфиденциальной биометрической информации к атаке прямого перебора возможных биометрических образов.

5.5.5. Синтез зависимых данных с ленточными матрицами коэффициентов корреляции

Отметим, что процедуры, описанные в предыдущем параграфе, позволяют имитировать данные со сложными случайными матрицами коэффициентов корреляции и воспроизводить ситуации, возникающие в биометрических системах. Как следует из предыдущего параграфа, эти ситуации достаточно просто воспроизвести и численно смоделировать на обычных персональных компьютерах.

Параллельно с численным описанием искусственных нейронных сетей огромный интерес представляют попытки их аналитического описания, построенные на некоторых общепринятых упрощениях. Одним из таких упрощений является использование ленточных матриц. Соответственно, необходимо указать деформации процедур,

описанных в предыдущих параграфах, приводящие к появлению ленточных матриц коэффициентов корреляции.

Переход к ленточным корреляционным матрицам достаточно просто осуществим. Для этого перехода достаточно в уравнениях (5.3), …, (5.9) заменить полные коррелирующие матрицы на ленточные матрицы. Например, для матриц простейшей коррелирующей матрицы (5.3) подобная замена будет выглядеть следующим образом:

$$\begin{bmatrix} a & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & a & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & a & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & a & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & a & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & a & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & a \end{bmatrix} \Rightarrow \begin{bmatrix} ka & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & a & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & a & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & a & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & a & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & a & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & ka \end{bmatrix}. \quad (5.10)$$

В правой матрице k – это коэффициент пересчета, учитывающий то, что первая и последняя строки этой матрицы имеют по два ненулевых элемента, все другие матрицы имеют по 3 ненулевых элемента. Коэффициент пересчета k выбирается по номограмме, подобной номограмме рисунка 5.7, исходя из заданного значения a, ширины ленты и требуемого значения r в конечной корреляционной матрице. В частности, для $a = 4$ при ленте шириной из трех элементов и $r = 0{,}52$ коэффициент $k = 0{,}92$. В конечном итоге замена левой коррелирующей матрицы (5.10) на правую коррелирующую матрицу этого же выражения приводит к изменениям конечных корреляционных матриц следующего вида:

$$\begin{bmatrix} 1 & r & r & r & r & r & r \\ r & 1 & r & r & r & r & r \\ r & r & 1 & r & r & r & r \\ r & r & r & 1 & r & r & r \\ r & r & r & r & 1 & r & r \\ r & r & r & r & r & 1 & r \\ r & r & r & r & r & r & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & r & 0 & 0 & 0 & 0 & 0 \\ r & 1 & r & 0 & 0 & 0 & 0 \\ 0 & r & 1 & r & 0 & 0 & 0 \\ 0 & 0 & r & 1 & r & 0 & 0 \\ 0 & 0 & 0 & r & 1 & r & 0 \\ 0 & 0 & 0 & 0 & r & 1 & r \\ 0 & 0 & 0 & 0 & 0 & r & 1 \end{bmatrix}. \quad (5.11)$$

Подчеркнем, что при таком подходе формирования связанных данных ширина ленты конечной корреляционной матрицы может быть любой от 1 до n, где n – порядок корреляционной матрицы. Кроме того, все преобразования с полными коррелирующими матрицами, описанные в предыдущих параграфах, могут быть распространены и на ленточные матрицы.

Соответственно, ленточные корреляционные матрицы могут иметь одинаковые и случайные элементы лент, элементы с одинаковыми и разными знаками.

5.6. Генераторы с зависимыми выходными данными, имеющими степенную корреляционную матрицу

Одним из приемов аналитического упрощения является сведение реальных связей и процессов к зависимым выходным данным. В связи с этим рассмотрим синтез векторов зависимых данных, имеющих степенную корреляционную матрицу.

Подобные связи получаются, если формировать данные следующим образом:

$$\begin{cases} v_1 = x_1, \\ v_2 = (x_2 + a\, v_1)/\sqrt{1+a^2}, \\ v_3 = (x_3 + a\, v_2)/\sqrt{1+a^2}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ v_k = (x_k + a\, v_{k-1})/\sqrt{1+a^2}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{cases} \qquad (5.12)$$

где x_k – это случайные числа, полученные от генератора независимых случайных чисел с нормальным законом распределения значений, нулевым математическим ожиданием и единичной дисперсией.

Преобразования вида (5.12) дают корреляционную матрицу с убыванием значений коэффициентов корреляции пропорционально степени этих коэффициентов, причем степень определяется расстоянием коэффициента корреляции от диагонали

$$
\left[r_{v_i v_j} \right] = \begin{bmatrix} 1 & r & r^2 & ::: & r^n \\ r & 1 & r & ::: & r^{n-1} \\ r^2 & r & 1 & ::: & r^{n-2} \\ ::: & ::: & ::: & ::: & ::: \\ r^n & r^{n-1} & r^{n-2} & ::: & 1 \end{bmatrix}. \qquad (5.13)
$$

Связь коэффициента корреляции – r и задаваемого параметра – a приведена на графике рисунка 5.9.

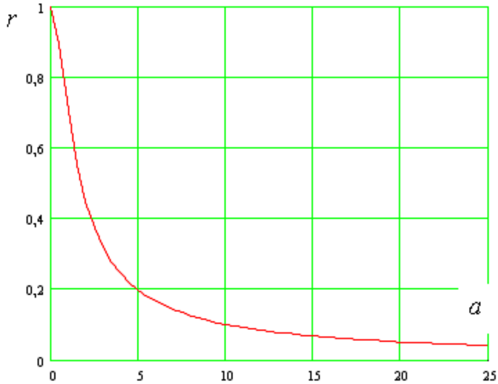


Рисунок 5.9 – График связи задаваемого параметра «a» и значений коэффициентов зависимых выходных данных, имеющих степенную корреляционную матрицу

Следует отметить, что связь задаваемого коэффициента «a» и коэффициентов корреляции достаточно проста. Порядок корреляционной матрицы или длина генерируемого вектора не влияет на функцию связи $r(a)$.

В том случае, если регулируемые параметры в корреляторе (5.12) сделать разными

$$
\begin{cases} v_1 = x_1, \\ v_2 = (x_2 + a_1 v_1) / \sqrt{1 + a_1^2}, \\ v_3 = (x_3 + a_2 v_2) / \sqrt{1 + a_2^2}, \\ \cdots\cdots\cdots\cdots \\ v_k = (x_k + a_{k-1} v_{k-1}) / \sqrt{1 + a_{k-1}^2}, \\ \cdots\cdots\cdots\cdots \end{cases} \qquad (5.14)
$$

мы получим гораздо более сложную корреляционную матрицу связей между генерируемыми биометрическими параметрами

$$\left[r_{v_i v_j}\right]=\begin{bmatrix} 1 & r_1 & r_{1,2}{}^2 & ::: & (r_{1,n-1})^n \\ r_1 & 1 & r_2 & ::: & (r_{2,n-2})^{n-1} \\ r_{1,2}{}^2 & r_2 & 1 & ::: & (r_{3,n-3})^{n-2} \\ ::: & ::: & ::: & ::: & ::: \\ (r_{1,n-1})^n & (r_{2,n-2})^{n-1} & (r_{3,n-3})^{n-2} & ::: & 1 \end{bmatrix}. \qquad (5.15)$$

В выражении (5.15) одиночный нижний индекс при r соответствует тому, что имеющий его элемент корреляционной матрицы зависит только от одного параметра a_i. Двойной нижний индекс при r соответствует тому, что имеющий эти 2 индекса элемент корреляционной матрицы зависит от 2 параметров a_i с соответствующими индексами.

Нетрудно показать, что для получения матрицы (5.15) с разными знаками при ее элементах, необходимо домножить левую часть уравнений (5.14) на ±1 со случайным знаком.

Для получения коррелированных данных со случайными значениями дисперсий и случайным распределением знаков коэффициентов корреляции необходимо ввести в уравнения (5.14) случайные числа χ_i, однократно полученные от генератора нормальных чисел с нулевым математическим ожиданием и единичной дисперсией:

$$\begin{cases} v_1 = \chi_1 x_1, \\ v_2 = \chi_2 (x_2 + \dfrac{a_1 v_1}{|\chi_1|}) / \sqrt{1 + a_1{}^2}, \\ v_3 = \chi_3 (x_3 + \dfrac{a_2 v_2}{|\chi_2|}) / \sqrt{1 + a_2{}^2}, \\ \\ v_k = \chi_k (x_k + \dfrac{a_{k-1} v_{k-1}}{|\chi_{k-1}|}) / \sqrt{1 + a_{k-1}{}^2}. \end{cases} \qquad (5.14)$$

Таким образом, синтезировать матрицы делающие зависимыми данные, достаточно просто. Создать программную реализацию корреляторов структурной схемы, изображенной на рисунке 5.1, оказывается несложно. Намного труднее создать и сбалансировать качественный генератор случайных чисел.

Г л а в а 6

Теория информации в приложении к высоконадежной биометрической защите

6.1. Криптографическая защитная информация (оценка защитной информации)

Высоконадежная биометрическая аутентификация пользователей возможна только тогда, когда биометрические механизмы надежно сопряжены с криптографическими механизмами аутентификации. При совместном описании биометрических и криптографических механизмов возникает проблема стыковки их терминов. На данный момент термины биометрии и термины защиты информации не состыкованы и даже при полном языковом тождестве имеют разное смысловое наполнение.

Одним из путей решения этой задачи является объединение терминов биометрии и «защиты информации», а также введение недостающих терминов в эти две предметные области, через использование более общего терминологического аппарата «теории информации». «Теория информации» активно развивалась в 50-е, 60-е гг. Прошлого века силами таких ученых, как Шеннон [43], Колмогоров [44], Фишер и Кульбак [45]. «Теория информации» развивалась в основном для приложений кодирования дискретных данных каналов связи, измерительной техники, автоматического управления и контроля. В 90-е гг. и начале этого века наступило определенное снижение активности публикаций по «теории информации», однако ее потенциал по-прежнему оказывается востребованным. Покажем возможности «теории информации» на примере объединения и дополнения терминов и понятий биометрии и «защиты информации».

Одним их основных понятий теории защиты информации является уровень защищенности, обеспечиваемый тем или иным механизмом защиты. Для примера рассмотрим криптографический механизм защиты информации, построенный на алгоритме симметричного шифрования с длиной ключа 256 бит. Для измерения уровня защи-

щенности по теории информации следует построить некоторый функционал вероятности преодоления этой защиты. Определим этот функционал следующим образом:

$$J_2 = -\log_2(P_2), \qquad (6.1)$$

где J_2 – уровень защищенности, измеряемый в битах (длина эквивалентного симметричного двоичного ключа или логарифмическая мера числа возможных состояний эквивалентного ключевого поля); P_2 – вероятность преодоления защиты с первой попытки или вероятность удачи атаки случайного подбора ключа с первой попытки.[1]

Криптографические механизмы защиты информации, построенные на базе симметричного шифрования, следует рассматривать как эталонные. Для них длина ключа и уровень защищенности практически совпадают, т. е. для симметричного алгоритма шифрования по ГОСТ 28147–89 уровень защищенности составит $J_2 \cong 256$ бит. Полное тождество $J_2 \equiv 256$ будет наблюдаться только в системах, где допускается использование не только сильных, но и всех слабых ключей.

Функционал (6.1) мы можем построить для совершенно разных криптографических механизмов защиты информации. В таблице 6.1 приведены логарифмические показатели уровня защищенности для парольной аутентификации и ряда иностранных криптографических механизмов по данным [46].

Таблица 6.1 наглядно показывает, что длина криптографического ключа и уровень защищенности соответствующего криптографического механизма далеко не всегда совпадают. Более того, при точных оценках уровня защищенности в общем случае должны получаться дробные (не целые) длины эквивалентного ключа, что является следствием непрерывности (не дискретности) функционала (6.1).

---

[1] Для всех систем защиты информации основной (первой) задачей является предоставление доступа к информации аутентифицированным пользователям, соответственно, вероятность отказа в доступе является вероятностью ошибок первого рода – P_1. Вторая задача – не дать доступ «Чужому», соответственно, вероятности ошибок второго рода – P_2 будет тождественна вероятности преодоления защиты.

Таблица 6.1 – Сравнительная оценка длины криптографического ключа и уровня защищенности

Наименование механизма криптографической защиты и его характеристики	Значение логарифмического показателя уровня защищенности
Парольная аутентификация, классический восьмисимвольный пароль, выбираемый лично пользователем	22,7 бит
Алгоритм DES с 56-битным ключом (ANSI X9.9 и другие DES механизмы аутентификации, шифрования и т. д.	54 бита
Аутентификация на основе одноразовых паролей SecurID	63 бита
Аутентификационный механизм с 512-битовым открытым ключом для цифровых подписей	63 бита
Криптографический механизм с 768-битовым ключом – минимальный размер RSA-ключа согласно рекомендации RSA Security 1999 г.	76 бит
Механизмы аутентификации с 1024-битовыми открытыми ключами	86 бит
Механизмы аутентификации с 2048-битовыми открытыми ключами с высокой защищенностью	116 бит
Криптографический механизм с алгоритмом ASE со 128-битным ключом	127 бит

6.2. Относительность оценки меры защитной информации, содержащейся в биометрическом образе

По аналогии с криптографическими механизмами защиты для биометрических механизмов защиты также можно указать размер эквивалентного ключа. Искусственная нейронная сеть может содержать 256 выходов и порождать ключи длиной 256 бит, однако эквивалент симметричного ключа для биометрического ключа в 256 бит будет в несколько раз короче. Для каждого биометрического образа необходимо оценивать длину эквивалентного ключа процедурами, описанными ранее в параграфах 3.7, …, 3.10. При этом каждый раз мы получаем некоторую оценку информативности конкретного био-

метрического образа, которая и является оценкой его собственной защитной информации.

Принципиально важным моментом является то, что длина эквивалентного симметричного ключа для биометрического образа или мера содержащейся в нем защитной информации является не абсолютной, а относительной величиной. Если мы будем использовать более эффективный преобразователь биометрия/код, то информативность биометрического образа вырастет, т. е. по мере совершенствования технологии нейросетевого извлечения защитной информации из биометрического образа его информативность будет расти.

В этом плане все оценки защитной информации, содержащейся в том или ином биометрическом образе, являются относительными и имеют смысл только в контексте конкретного продукта биометрической защиты информации. По этой причине в таблицах приложения «2А» и «3А» проекта стандарта [11], связывающих длину биометрических паролей с длиной эквивалентного ключа, даны ссылки на организацию, отвечающую за достоверность информации, а также указан период времени, в какой эта информация была получена.

Нейросетевые технологии извлечения защитной информации из биометрических образов – это не что иное, как технологии обогащения информации, технологии добычи знаний [47, 48]. Предположительно, эта ветвь информационных технологий в ближайшей перспективе будет активно развиваться и, соответственно, придется корректировать таблицы информативности биометрических образов под новые возможности более совершенных нейросетевых продуктов.

6.3. Информативность доступности распознаваемых биометрических образов

Одной из весьма интересных особенностей информационного подхода к сравнительной оценке разнородных механизмов защиты информации является то, что применительно к ним появляются новые понятия. В частности, «защитная информация» или «информация о защите» – J_2, определяемая через функционал (6.1), может быть дополнена «информацией о доступности»:

$$J_1 = -\log_2(P_1), \qquad (6.2)$$

где J_1 – уровень доступности, измеряемый в битах (длина битового дополнения эквивалентного симметричного двоичного ключа или логарифмическая мера числа возможных состояний эквивалентного ключевого поля, утраченных из-за необходимости увеличения доступности); P_1 – вероятность отказа в доступе «Своему» с первой попытки из-за помех в канале связи или естественной нестабильности биометрических образов.

В качестве примера рассмотрим биометрико-криптографические механизмы аутентификации. В них можно пожертвовать частью длины ключа, используя ее для обнаружения и исправления ошибок. Естественно, при этом мы снижаем уровень защищенности, но поднимаем уровень доступности. На данный момент типовой уровень доступности биометрических механизмов составляет значение $J_1 \leq 7$ бит, что соответствует вероятностям ошибок первого рода $P_1 \geq 0{,}01$. Если в будущем будет поставлена задача по существенному повышению уровня доступности, то для этого потребуется тем или иным способом увеличить расход дополнительных бит кода по обнаружению и исправлению ошибок. Естественно, что связь числа дополнительных бит, идущих на обнаружение и исправление ошибок, с реальным ростом уровня доступности существенно нелинейна. Для каждого из известных кодов с обнаружением и исправлением ошибок получается своя нелинейная зависимость. Уже сейчас ясно, что при увеличении уровня доступности на 1 бит до значения $J_1 \leq 8$ приходится жертвовать порядка 12 бит кода эквивалентного ключа на процедуры обнаружения и исправления ошибок. Тем не менее движение в сторону повышения уровня доступности вполне возможно и рано или поздно будет востребованным. Тогда функционалы вида (6.2) будут востребованы.

6.4. Балансировка информативности биометрических средств по доступности и защищенности

В рамках решения задач высоконадежной биометрической аутентификации считаются вполне допустимыми сравнительно высокие вероятности ошибок первого рода (отказа «Своему») и очень низкие вероятности ошибок второго рода (пропуска «Чужого»). Это связано

с тем, что пользователь может несколько раз предъявить свой тайный биометрический образ. Если оказывается, что вероятности ошибок первого и второго рода должны быть сопоставимы, мы приходим к очень важной характеристике

$$P_{EE} = P_1 = P_2 . \tag{6.3}$$

Системы защиты информации, обладающие свойством (6.3), одновременно являются и высоконадежными, и высокодоступными. Их параметры должны специальным образом балансироваться с тем, чтобы выполнялось (точно или приближенно) соотношение (6.3). По индукции для сравнений этого типа систем следует использовать соответствующий логарифмический показатель универсального качества

$$J_{EE} = -\log_2(P_{EE}) . \tag{6.4}$$

В случае, когда P_1 и P_2 достаточно близки, универсальный показатель качества может быть приближенно вычислен через их усреднение

$$J_{EE} \approx -\log_2\left(\frac{P_1 + P_2}{2}\right). \tag{6.5}$$

Показатели (6.4), (6.5) отражают потенциальную возможность механизма защиты информации оставаться дружественным к пользователю при усилении его защитных свойств. Если J_{EE} много меньше J_2, то усиление защитных свойств администратором безопасности будет приводить к отторжению пользователями биометрико-криптографического механизма защиты. Фактически это означает, что этот механизм не имеет широких возможностей по регулировкам уровня защищенности и уровня доступности. Для универсальных по качеству механизмов защиты регулировки одного из уровней не должны приводить к катастрофическому снижению другого показателя.

Очевидно, что информативность равновероятных систем оказывается всегда намного меньше, чем средняя информативность доступности и защищенности

$$J_{EE} \le (-\log_2(P_1) - \log_2(P_2) + 1) . \tag{6.6}$$

Попытки балансировки средств высоконадежной биометрической аутентификации обязательно приводят к ухудшению среднего значе-

ния информативности. По крайней мере, это наблюдается у современных биометрико-нейросетевых преобразователей. Если мы имеем для типичного биометрического преобразователя $J_1 = 7$, $J_2 = 60$, то при попытках его балансировки мы должны получить $J_{EE} \langle\langle 32$. Предварительные эксперименты показывают, что информативность падает в несколько раз $J_{EE} \approx 12$. Практически трехкратное снижение информативности свидетельствует о существенной нелинейности J_{EE} (J_1, J_2) от ее аргументов. Связано ли это с особенностями тестируемых нейросетевых механизмов или это общая закономерность, пока неизвестно.

6.5. Информационное описание высокоинтеллектуальных систем распознавания множества биометрических образов

Рассмотренные выше системы биометрической аутентификации достаточно просты в своем информационном описании. Они имеют только два состояния «Свой» и «Чужой», соответственно, информационное описание таких систем дает простейший граф, отображенный на рисунке 6.1.

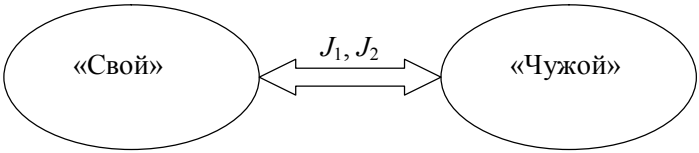


Рисунок 6.1 – Простейший граф информационных связей однопользовательской биометрической защиты

Состояния простейшего графа меняются с интенсивностью J_1 и J_2. Если мы делаем систему многопользовательской и обучаем ее распознаванию несколько пользователей, то размерность ее полного информационного описания резко увеличивается. Задача становится многомерной, система имеет гораздо более сложный граф информационного описания, приведенный на рисунке 6.2.

Как видно из рисунка 6.2, каждое из возможных состояний системы будет связано с любым другим состоянием системы своим информационным потоком J_1 и J_2. Для того, чтобы получить полное

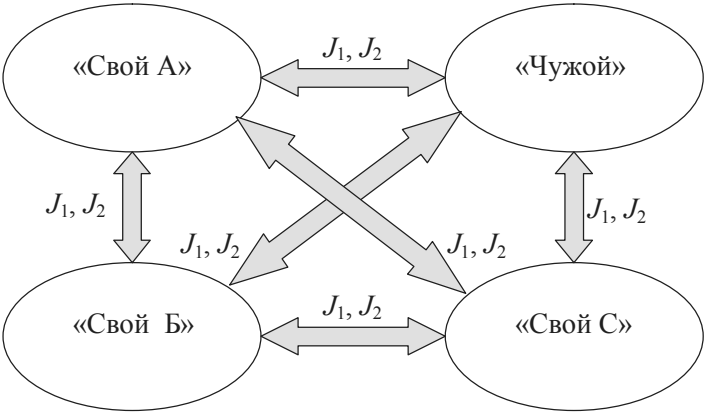


Рисунок 6.2 – Информационный граф
многопользовательской биометрической системы

описание сложной биометрической системы, только двух вероятностных характеристик P_1 и P_2 или их информационных аналогов уже недостаточно. Такие характеристики приходится вводить для каждого из возможных состояний биометрической защиты. Сложность полного тестирования многократно возрастает, приходится вводить упрощения и пренебрегать некоторыми связями полного графа.

Одним из наиболее эффективных упрощений является пренебрежение ошибочным перепутыванием «Своих». В этом случае мы должны контролировать каждого из «Своих» по отношению к «Чужим», но не можем гарантировать надежного различения «Своих». Отсутствие подобных проверок, тестов и гарантий может привести к коллизиям, следующего рода: когда один из группы «Свой» будет выдавать себя за другого из этой же группы. Насколько страшны подобные коллизии, могут решать только пользователи системы или ее владелец. На данный момент ясно одно: сложность тестирования высоконадежных многопользовательских систем многократно возрастает, соответственно, очень быстро должна расти их стоимость или стоимость их сертификации.

Подчеркнем, что привлечение теории информации к описанию искусственных нейронных сетей активно развивается [35], однако описание сложных нейросетевых систем через энтропию и взаимную

информацию для нейросетевых приложений информационной безопасности, видимо, нецелесообразно. Авторы придерживаются мнения, что информационные графы по своей сути являются тем же самым, что и энтропийные ряды и множество показателей взаимной информации, однако они гораздо более наглядны и удобны для практических выводов и интерпретаций. То, что защитную и доступную информацию можно рассматривать как некоторые эквивалентные ключи, позволяет переносить ряд полезных тестовых приемов криптографии. Это как раз тот случай, когда форма оказывается удобна для описания одновременно нейросетей и классической криптографии. Для высоконадежной биометрии именно организация безопасного стыка биометрии и криптографии является основным предметом исследований.

Глава 7

Гарантии безопасности использования нейросетевых технологий защиты информации

Сейчас уже нет сомнений, что биометрии принадлежит будущее. Вот почему правительства развитых западных держав и крупные корпорации инвестируют огромные суммы в развитие биометрических технологий и, что становится особенно важным в последнее время, в их тестирование и практическое внедрение.

На сегодняшний день признанным лидером разработки и внедрения биометрических технологий являются США [12]. Исследования в данной области возведены в США в ранг государственной задачи. Национальные институты стандартизации США (NIST и ANSI) за последние 7 лет разработали порядка 15 национальных биометрических стандартов, большинство из которых в данный момент используется как основа при разработке международных биометрических стандартов, специально созданным подкомитетом ISO/IEC JTC1 SC37. Правительством США в период с 1998 г. по настоящее время организована подготовка специалистов по биометрии в пяти различных университетах страны.

Все вышеперечисленное свидетельствует о значительном внимании к биометрии и развитию биометрических технологий. Столь значительное общественное внимание к биометрии привело к тому, что Международная организация по стандартизации (ISO) в лице ее 27-го подкомитета официально рассматривает усиление паролей и персональных кодов биометрией как одну из основных тенденций развития систем информационной безопасности [www.din.de/ni/sc27].

На международном уровне усилия по созданию биометрических стандартов координирует ISO (International Organization of Standartization), 13 декабря 2002 г. был создан специальный подкомитет SC37 по вопросам биометрии при первом объединенном комитете – JTC1 (joint technical committee 1), занимающемся информационными технологиями. Необходимость в создании нового специального подкомитета ISO/IEC JTC1 SC37 продиктована тем, что после решения

лидеров ГРУППЫ ВОСЬМИ ведущих стран в 2002 г. возникла необходимость унификации биометрических данных в национальных электронных паспортах разных стран. По сути дела именно это обстоятельство и форсировало работу недавно созданного биометрического подкомитета ISO/IEC JTC1 SC37.

Производители биометрических устройств и технологий объединены в рамках международной ассоциации IBIA (International Biometric Industry Association), которая активно влияет на процессы подготовки новых стандартов. Через IBIA производители регистрируют свои форматы представления биометрических данных, которые далее гармонизируются и обобщаются в виде международных стандартов и рекомендаций. На данный момент зарегистрировано 27 форматов данных, используемых в биометрических устройствах и технологиях различных компаний.

Естественно, что Россия не может оставаться в стороне от наметившихся процессов объединения международных усилий и стандартизации биометрических технологий. Россия в лице Госстандарта является полноправным членом ISO/IEC, в феврале 2003 г. при ГОСТ Р ТК355 (технический комитет «Автоматическая идентификация») был создан 7-й подкомитет, занимающийся только вопросами биометрической идентификации. На ГОСТ Р ТК355/ПК7 в настоящее время лежит вся тяжесть работы по гармонизации международных биометрических стандартов. В 2006 г. планируется завершить работу по гармонизации 3 проектов международных биометрических стандартов. В течение нескольких ближайших лет к русскому языку будут адаптированы и последующие международные биометрические стандарты.

Создавая средства высоконадежной биометрической защиты, Россия в лице Гостехкомиссии при президенте РФ первой из развитых стран столкнулась с проблемой сертификации такого типа новых информационных технологий [49].

В этой области развития информационных технологий в настоящее время Россия лидирует и, в принципе, не может воспользоваться существующим международным опытом (его просто нет). Видимо, России придется самостоятельно двигаться по достаточно дорогостоящему пути создания своей системы аттестации и сертификации высоконадежных биометрических технологий, по пути создания Российских методик и стандартов тестирования высоконадежных био-

метрических систем защиты, созданных с учетом высоконадежной поддержки отечественных средств криптографии. Только после того, как этот пласт работ будет выполнен, возможно будет обобщение накопленного Российского опыта в виде переноса содержательной части отечественных рекомендаций и документов во фрагменты международных стандартов по организации и тестированию систем высоконадежной биометрической защиты.

Следует подчеркнуть, что международные стандарты ISO/IEC JTC1 SC37 следует отнести к «слабой биометрии», имеющей высокие вероятности ошибок первого и второго рода от 0,01 до 0,0001. Если такого типа системы использовать в совокупности с криптографическими механизмами защиты, то они практически ничего не дают с позиций теории защиты информации. Для создания эффективных механизмов биометрико-криптографической защиты необходимо вероятности ошибок второго рода (ошибочного признания «Чужого» как «Своего») для биометрических механизмов снизить на 10, …, 20 порядков [50, 51].

С 2004 г. в межведомственной лаборатории тестирования биометрических устройств и технологий при факультете военного обучения Пензенского государственного университета проводятся исследования по созданию больших и сверхбольших баз биометрических образов написания слов-паролей и записи голосовых фраз-паролей, их систематизации и изучению. Тестирование относительно слабых биометрических устройств планируется осуществлять по международным биометрическим стандартам, которые уже приняты или проходят процесс рассмотрения и будут приняты в ближайшее время [52]. Официальные номера и переводы на русский язык названий этих стандартов даны в таблице 7.1.

Тестирование же высоконадежной российской биометрии будет осуществляться в соответствии с национальным стандартом ГОСТ Р [11]. Однако уже сейчас видно, что указанная выше окончательная редакция проекта ГОСТ Р (ТК 362) будет, по аналогии со стандартами таблицы 7.1, дополняться целым пакетом специальных стандартов.

Ознакомление с англоязычными версиями этих стандартов позволяет сделать однозначный вывод о том, что они не применимы к российским высоконадежным биометрическим технологиям со стойкостью к атакам подбора.

Таблица 7.1 – Стандарты по тестированию биометрии

Номер стандарта ISO/IEC	Название стандарта
1.37.19795 Ref. No.: NP 19795	Biometric Performance Testing and Reporting – процедуры выполнения тестирований и отчетов в биометрике
1.37.19795.1 Ref. No.: NP 19795	Biometric Performance Testing and Reporting – Part 1: Principles and Framework – процедуры выполнения тестирований и отчетов в биометрии. Часть 1: Принципы и структура
1.37.19795.2 Ref. No.: NP 19795	Biometric Performance Testing and Reporting – Part 2: Testing Methodologies – процедуры выполнения тестирований и отчетов в биометрии. Часть 2: Методики тестирования
1.37.19795.3 Ref. No.: NP 19795	Biometric Performance Testing and Reporting – Part 3: Specific Testing Methodologies – процедуры выполнения тестирований и отчетов в биометрии. Часть 3: Специальные методики тестирования
1.37.19795.4 Ref. No.: NP 19795	Biometric Performance Testing and Reporting – Part 4: Specific Test Programmes – процедуры выполнения тестирований и отчетов в биометрии. Часть 4: Специальные программы тестирования

Международный стандарт ISO/IEC 15408 адаптирован под условия России, пользуясь уже накопленным международным опытом и экономя при этом свои ресурсы. При разработке своих биометрических стандартов, своих профилей защиты высоконадежных биометрических систем нам придется самим тратить свои национальные ресурсы в завышенном объеме, проделывая эту работу не только для себя, но и для всего мирового сообщества.

Появляется реальная возможность возврата нашего морального долга мировому сообществу за предшествовавшее заимствование стандарта ISO/IEC 15408 и сопряженную с ним экономию национальных ресурсов в предыдущие несколько лет.

Отметим, что столь высокие требования (снижение вероятностей ошибок на десятки порядков) – это достаточно сложная техническая задача, однако ведущие страны все же ей активно занимаются. Наибольших успехов в этом направлении добились Россия и США [51, 53, 54]. Россия и США идут разными техническими путями: Россия идет по пути использования больших и сверхбольших искусственных нейронных сетей [55, 56]; исследователи США предлагают мировому сообществу использовать аппарат нечеткой (размытой)

логики для синтеза экстракторов (обогатителей) плохих, нечетких биометрических данных [7]. И в том, и другом случае речь идет о синтезе автоматического преобразователя биометрия/код, который вектор из нескольких сотен не однозначных, нестабильных биометрических параметров преобразует в однозначный криптографический ключ длинной в несколько сотен бит.

Применение новых нейросетевых технологий для защиты информации требует высокого доверия к программным продуктам, созданным с помощью этих технологий. Программные продукты должны соответствовать системе национальных и международных биометрических стандартов. Кроме того, они должны быть сертифицированы соответствующими органами международной или национальной сертификации. В настоящее время этой системы стандартов и сертификатов нет, она находится на стадии создания [56, 57].

Имеющиеся и разрабатываемые на сегодняшний день стандарты можно разделить на группы для того, чтобы оценить объем предстоящей работы. Вариант представлен на рисунке 7.1.

Рисунок 7.1 – Семь групп биометрических стандартов, обеспечивающих возможность сертификации биометрических продуктов

К первой группе стандартов можно отнести стандарты, описывающие операции предварительной обработки биометрических данных. В связи с тем, что реально для создания тайных биометрических образов могут быть использованы только три технологии, первая группа делится на три подгруппы:

1.1. Стандарты по предварительной обработке рукописного почерка.

1.2. Стандарты по предварительной обработке голосовых данных.

1.3. Стандарты по предварительной обработке динамики клавиатурного почерка.

В свою очередь, группа 1.1, видимо, должна объединять два стандарта:

1.1.1. Предварительная обработка данных динамики рукописной подписи.

1.1.2. Предварительная обработка данных статического следа рукописной подписи.

Во вторую подгруппу 1.2, видимо, должны входить три достаточно независимых стандарта:

1.2.1. Предварительная обработка голоса гребенкой узкополосных фильтров для последующей биометрической аутентификации.

1.2.2. Предварительная обработка речи линейными предсказателями для последующей биометрической аутентификации.

1.2.3. Нейросетевая предобработка речи перед биометрической аутентификацией.

Третья подгруппа 1.3 стандартов должна состоять из одного стандарта.

1.3.1. Предварительная обработка данных для аутентификации личности по динамике клавиатурного почерка.

Вторая, наиболее важная, группа стандартов должна состоять из двух стандартов:

2.1. Синтез и обучение нейронных сетей, предназначенных для безопасного хранения биометрических данных и информации о личном ключе пользователя.

2.2. Процедуры прогноза стойкости расширяющихся нейронных сетей для безопасного хранения биометрических данных и информа-

ции о личном ключе пользователя к атакам подбора, построенные на учете промежуточных данных обучения и на учете качества обучающей выборки.

Третья группа должна включать в себя руководящие документы или стандарты, регламентирующие основные требования по безопасному использованию биометрических паролей или ключей. Видимо, в эту группу могут входить три документа регламентирующих три технологии, позволяющие иметь тайные биометрические образы (рукописный почерк, голос, клавиатурный почерк).

Четвертая группа стандартов должна содержать требования по корректному объединению искусственных нейронных сетей с протоколами аутентификации, выполненными в соответствии с отечественными криптографическими стандартами (ГОСТ 28147–89, ГОСТ 34.10–95, ГОСТ 34.11–95) и международными криптографическими стандартами (DES, RSA, RC2, RC4, RC5, RC6, DSS, DSA, MD2, MD5). По нашему мнению, в данный момент разработка этой группы стандартов менее актуальна, чем разработка первой и второй групп стандартов.

Пятая группа стандартов должна содержать процедуры тестирования и сертификации систем биометрической аутентификации с расширяющимися ИНС к попыткам атак подбора со стороны входов или выходов, к попыткам извлечения информации из структуры и связей ИНС. Эта группа стандартов должна позволить производителям самостоятельно проверять заявляемую ими стойкость системы биометрической аутентификации. Стандарты должны содержать тексты программ автоматического тестирования и методики обработки реально полученных данных после многодневного тестирования заявленных производителями малых вероятностей ошибок второго рода. На данный момент трудно однозначно определить состав этой группы стандартов, их ориентировочные названия и смысловое содержание. Видимо, пятая группа стандартов должна включать, как минимум, четыре стандарта:

1.5.1. Имитаторы случайных правдоподобных атак на входы ИНС, хранящей в своих связях и параметрах конфиденциальную биометрическую и криптографическую информацию.

1.5.2. Процедуры ускоренного тестирования стойкости ИНС к атакам подбора по Хеммингу.

1.5.3. Процедуры ускоренной проверки степени неоднородности деления ИНС, ее выходного пространства (быстрые процедуры вычисления закона распределения мощностей множеств, порождающих истинный ключ и ложные ключи).

1.5.4. Процедуры ускоренной проверки степени хэширования (перемешивания) элементов пространства ИНС, хранящей в своих параметрах и связях конфиденциальную биометрическую и криптографическую информацию.

Шестая группа стандартов должна позволять измерять относительное и абсолютное качество (уникальность, стабильность) конкретных биометрических образов, воспроизведенных в рамках рукописной графики русского языка и кириллического рукописного алфавита, а также других языков и алфавитов. То же самое касается измерения абсолютного и относительного качества парольной фразы, воспроизведенной голосом на русском языке или других языках. Группа этих стандартов должна включать CD-ROM с образами рукописной графики и аналогичный CD-ROM с образами голосовых фраз, характерных для русскоязычной группы населения или других языковых групп населения России. Такие образцовые CD-ROM носители эталонных данных должны отражать среднестатистические характеристики конкретной группы населения и использоваться для тестирования качества (сертификации) систем предварительной биометрической обработки. В итоге, шестая группа стандартов должна включать, как минимум, три подгруппы документов с соответствующими CD-ROM приложениями:

6.1. Стандарты по оценке качества работы программ предварительной обработки рукописного почерка и по оценке качества биометрического образа конкретного пользователя. База данных (например, на CD-ROM носителе) с эталонными примерами рукописного почерка известного качества.

6.2. Стандарты по оценке качества работы программ предварительной обработки голоса и по оценке качества биометрического образа конкретного пользователя. База данных (например, на CD-ROM

носителе) с эталонными примерами произношения известного качества для разных языковых групп населения.

6.3. Стандарты по оценке качества работы программ предварительной обработки клавиатурного почерка и по оценке качества биометрического образа конкретного пользователя. База данных (например, на CD-ROM носителе) с эталонными примерами клавиатурных почерков известного качества.

К седьмой группе стандартов следует отнести документы, регламентирующие образовательный процесс по биометрии и применение ИНС в биометрии. Заметим, что формирование этой группы документов должно опираться на отечественный и зарубежный опыт (www.engr.sjsu.edu/biometuic) преподавания в высшей школе. За рубежом лидером по вопросам подготовки и переподготовки специалистов по биометрии является университет Сан-Хосе (США, Калифорния, Силиконовая долина, www.engr.sjsu.edu/biometuic). Государственный университет Сан-Хосе дает второе образование по биометрии, при нем организован институт «БИОМЕТРИЧЕСКОЙ АУТЕНТИФИКАЦИИ» и «НАЦИОНАЛЬНЫЙ ЦЕНТР ТЕСТИРОВАНИЯ И РАЗРАБОТКИ БИОМЕТРИЧЕСКИХ СИСТЕМ». И институт, и центр созданы на деньги Федерального правительства США, стремящегося сделать производителей биометрии США лидерами биометрических технологий.

В России лидирующее положение по научным исследованиям и подготовке кадров занимает Пензенский государственный университет, где поддерживаются четыре биометрические технологии аутентификации (по отпечаткам пальцев, голосу, клавиатурному и рукописному почерку). Силами университета (без поддержки федерального правительства) организован курс из 5 лабораторных работ по биометрической аутентификации, проводимый на кафедре «Безопасность информационных технологий» с конца прошлого века [21] для двух специальностей в рамках преподавания ряда профилирующих дисциплин.

При подготовке документов 7-й группы, видимо, имеет смысл взять за основу учебные программы университета Сан-Хосе по подготовке специалистов и дополнить их технологией удаленной биометрической аутентификации. На сегодня эта технология поддержи-

вается только в России Пензенским государственным университетом, о чем можно судить по содержанию международного стандарта BioAPI (www/bioapi.org; с. 19, раздел 1.8, где прямо говорится об отсутствии за рубежом нейросетевой технологии преобразования биометрических образов в криптографический ключ).

В целом получается достаточно большое поле увязанных между собой биометрических и нейросетевых стандартов. Разработать параллельно все эти стандарты, видимо, невозможно. Скорее всего, необходимо начать со стандартов второй группы и оформлять их как RFC стандарты разного уровня проработки или даже как Internet-Drafts документы. Целесообразно проводить эту работу на русском языке с последующим переводом документов на английский язык. В связи со значительным объемом работ перечисленные выше стандарты могут быть созданы только при целевой государственной поддержке и при поддержке работ по созданию стандартов фирмами-производителями биометрических устройств. Видимо, по аналогии с практикой США (www.engr.sjsu.edu/biometuic/…, www.biometrics.org) необходимо формировать специализированные университетские программы по развитию национальной биометрии и подготовке специалистов высокой квалификации, способных решать в будущем проблемы биометрии.

Предположительно, и это неполный список. В процессе исследований указанные выше документы могут корректироваться и дополняться.

Заключение

Разработка и применение средств, использующих биометрико-нейросетевые технологии, становятся реальностью сегодняшнего дня. При этом заявленная производителями стойкость этих технологий требует реального подтверждения путем тестирования.

В результате теоретических и практических исследований выясняется, что стоимость тестирования и подтверждения реальной стойкости средств, использующих высоконадежную биометрию, требует больше затрат таких, как временные, человеческие и финансовые, чем разработка и производство этих устройств. Поэтому для тестирования средств высоконадежной биометрии необходимо создание реальных сбалансированных баз биометрических образов достаточно большого размера, позволяющих подтвердить гипотезу о том или ином законе распределения значений с высокой надежностью. Знание же о законе распределения позволит проводить предварительное тестирование средств на небольших представительных выборках.

Знание о законе распределения позволит вместо термина «оценка» стойкости средств биометрической защиты использовать термин «измерение» биометрической стойкости средств защиты. Классический эталон и его погрешность в обычных измерениях замещаются на «знание» закона распределения значений контролируемой величины и доказанную (проверенную, аттестованную) погрешность этого «знания».

Таким образом, основной задачей аттестации (сертификации) средств биометрической защиты по их стойкости является получение знаний о законе распределения выходных значений преобразователей биометрия/код.

Данная монография является первой попыткой обобщения исследований, проводимых сотрудниками межведомственной лаборатории тестирования биометрических устройств и технологий при факультете военного обучения Пензенского государственного университета.

Список литературы

1. Волчихин В. И. Быстрые алгоритмы обучения нейросетевых механизмов биометрико-криптографической защиты информации / В. И. Волчихин, А. И. Иванов, В. А. Фунтиков. – Пенза: Изд-во Пенз. гос. ун-та, 2005. – 276 с.

2. ISO/IEC 1.37.19795 Процедуры выполнения тестирования и отчетов в биометрии.

3. ISO/IEC 1.37.19795.1. Процедуры выполнения тестирования и отчетов в биометрии. Часть 1: Принципы и структура.

4. ISO/IEC 1.37.19795.2. Процедуры выполнения тестирования и отчетов в биометрии. Часть 2: Методики тестирования.

5. ISO/IEC 1.37.19795.3. Процедуры выполнения тестирования и отчетов в биометрии. Часть 3: Специальные методики тестирования.

6. ISO/IEC 1.37.19795.4. Процедуры выполнения тестирования и отчетов в биометрии. Часть 4: Специальные программы тестирования.

7. Dodis Y. Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data / Yevgeni Dodis, Leonid Reyzin, Adam Smith // April 13, 2004. www.cs.bu.edu/~reyzin/fuzzy.html.

8. Малыгин А. Ю. Повышение качества решений через увеличение размерности выходного вектора нейросетевых преобразователей биометрия/код // А. Ю. Малыгин, А. И. Иванов, О. В. Ефимов // Надежность и качество – 2006: Сб. материалов Междунар. симп. – Пенза: Изд-во Пенз. гос. ун-та, 2006. – С. 15–17.

9. Волчихин В. И. Тестирование стойкости нейросетевых механизмов биометрико-криптографической защиты информации / В. И. Волчихин, А. Ю. Малыгин, Ю. И. Олейник // Современные технологии безопасности. – М.: ООО «Информ-Эстейт», 2005. – № 1 (12). – С. 36–38.

10. Иванов А. И. Противодействие “цифровому неравенству” граждан: тестирование стойкости программных средств безопасного хранения ключа ЭЦП / А. И. Иванов, А. Ю. Малыгин, О. С. Захаров // Современные технологии безопасности. – М.: ООО «Информ-Эстейт», 2006. – № 1 (16). – С. 33–35.

11. Окончательная редакция проекта ГОСТ Р ТК 362 «Защита информации. Техника защиты информации. Требования к высоконадежным средствам биометрической аутентификации».

12. Отличить «своего» от «чужого». Проблемы развития и тестирования высоконадежной биометрии / В. И. Волчихин, А. Ю. Малыгин, Ю. И. Олейник, В. Н. Щурков // Системы безопасности – межотраслевой каталог. – М.: Groteck, 2006. – № 4. – С. 164–168.

13. Малыгин А. Ю. Проблемы, возникающие при тестировании и сертификации высоконадежных биометрических средств / А. Ю. Малыгин, В. А. Фунтиков, Ю. И. Олейник // Надежность и качество – 2006: Сб. материалов Междунар. симп. – Пенза: Изд-во Пенз. гос. ун-та, 2006. – С. 15–17.

14. Галушкин А. И. Нейронные сети: история развития / А. И. Галушкин, Я. З. Ципкин. – М.: ИПРЖ, 2002. – Кн. 5-я сер. «Нейрокомпьютеры и их применение».

15. Галушкин А. И. Теория нейронных сетей. – М.: ИПРЖ, 2000. – Кн. 1-я сер. «Нейрокомпьютеры и их применение».

16. Горбань А. Н. Нейронные сети на персональном компьютере / А. Н. Горбань, Д. А. Россиев. – Новосибирск: Наука, 1996. – 276 с.

17. Оссовский С. Нейронные сети для обработки информации. – М.: Финансы и статистика, 2002.

18. Венцель Е. С. Теория вероятностей и ее инженерные приложения / Е. С. Венцель, Л. А. Овчаров. – М.: Наука, 1988. – 480 с.

19. Пугачев В. С. Теория вероятностей и математическая статистика. – М.: Наука, 1979. – 495 с.

20. Иванов А. И. Биометрическая идентификация личности по динамике подсознательных движений. – Пенза: Изд-во Пенз. гос. ун-та, 2000. – 188 с.

21. Волчихин В. И. Биометрия: быстрое обучение искусственных нейронных сетей / В. И. Волчихин, А. И. Иванов. – Пенза: Изд-во Пенз. гос. ун-та, 2000. – 40 с.

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

23. Иванов А. И. Нейросетевые алгоритмы биометрической идентификации личности. – М.: Радиотехника, 2004. – 144 с. – Кн. 15-я сер. «Нейрокомпьютеры и их применение».

24. Обзор методов измерения параметров динамических биометрических образов человека / А. Ю. Малыгин, Ю. И. Олейник, Е. А. Малыгина [и др.]. – М.: ЦВНИ МО РФ, 2006. – Деп. в центральном справочно-информационном фонде МО РФ, справка № 14503. Сер. Б. Вып. № 75, инв. № Б5837.

25. Обзор методов измерения параметров статических биометрических образов личности / А. Ю. Малыгин, Ю. И. Олейник, Е. А. Малыгина [и др.]. – М.: ЦВНИ МО РФ, 2006. – Деп. в центральном справочно-информационном фонде МО РФ, справка № 14504. Сер. Б. Вып. № 75, инв. № Б5838.

26. Малыгин А. Ю. Биометрия: проблемы тестирования / А. Ю. Малыгин, Ю. И. Олейник, Е. А. Малыгина // Надежность и качество – 2005: Сб. материалов Междунар. симп. – Пенза: Изд-во Пенз. гос. ун-та, 2005.

27. Малыгин А. Ю. Сокращение объемов тестовых выборок за счет знания параметров закона распределения выходных состояний биометрия/код / А. Ю. Малыгин, А. И. Иванов, Д. Н. Надеев // Безопасность информационных технологий: Сб. тр. науч.-техн. конф. – Пенза: ПНИЭИ, 2006. – Т. 6. – С. 10–12. – Секция-2.

28. Иванов А. И. Оценка границ корректности гипотезы нормального закона распределения значений выходных состояний преобразователя биометрия/код / А. И. Иванов, А. Ю. Малыгин, Д. Н. Надеев // Безопасность информационных технологий: Сб. тр. науч.-техн. конф. – Пенза: ПНИЭИ, 2005. – Т. 6. – С. 65–66. Секция-2. (http://beda.stup.ac.ru/RV-conf/v06/017).

29. Иванов А. И. Оценка погрешностей определения статистических моментов, обусловленных отсутствием представительной выборки / А. И. Иванов, Д. Н. Надеев // Безопасность информационных технологий: Сб. тр. науч.-техн. конф. – Пенза: ПНИЭИ, 2006. – Т. 6. – Секция-2.

30. Цветков Э. И. Основы теории статистических измерений. – Л.: Энергоатомиздат, 1986. – 256 с.

31. Цветков Э. И. Методические погрешности статистических измерений. – Л.: Энергоатомиздат, 1984. – 190 с.

32. ГОСТ Р 50779.10–2000 (ИСО 3534–1–93). Статистические методы. Вероятность и основы статистики. Термины и определения.

33. ГОСТ Р 50779.21–2004. Статистические методы. Правила определения и методы расчета статистических характеристик по выборочным данным. Часть 1. Нормальное распределение.

34. Волчихин В. И. Особые требования к обучению биометрико-нейросетевых преобразователей с большим числом выходов / В. И. Волчихин, А. И. Иванов, А. Ю. Малыгин // Надежность и качество – 2006: Сб. материалов Междунар. симп. – Пенза: Изд-во Пенз. гос. ун-та, 2006. – С. 17–18.

35. Иванов А. И. Оценка потенциальной информативности нейропреобразования биометрического образа в криптографический ключ доступа /А. И. Иванов, А. Ю. Малыгин, А. В. Семенов // Безопасность информационных технологий: Сб. тр. науч.-техн. конф. – Пенза: ПНИЭИ, 2006. – Т. 6. – С. 25–27. – Секция-2.

36. Надеев Д. Н. Симметрия биномиального зависимого закона распределения относительно среднего модуля коэффициентов корреляции входных данных преобразователя биометрия/код / Д. Н. Надеев, А. И. Иванов, А. Ю. Малыгин // Надежность и качество – 2006: Сб. материалов Междунар. симп. – Пенза: Изд-во Пенз. гос. ун-та, 2006. – С. 228–229.

37. О проблеме ресурсов при тестировании стойкости высоконадежных биометрических технологий / В. И. Волчихин, А. Ю. Малыгин, М. Ю. Лупанов, А. В. Семенов // Вопросы защиты информации. – М.: Изд-во ВИНИ, 2006. – № 4. – С. 15–16.

38. Иванов А. И. Оценка потенциальной информативности нейропреобразования биометрического образа в криптографический ключ доступа / А. И. Иванов, А. Ю. Малыгин, А. В. Семенов // Безопасность информационных технологий: Сб. тр. науч.-техн. конф. – Пенза: ПНИЭИ, 2005. – Т. 6. – С. 5–7. – Секция-2.

39. Кнут Д. Искусство программирования для ЭВМ. – Т. 2. Получисленные методы. – М.: Мир, 1977.

40. Бобнев М. П. Генерирование случайных сигналов. – М.: Энергия, 1971. – 240 с.

41. Шалыгин А. С. Прикладные методы статистического моделирования / А. С. Шалыгин, Ю. И. Палагин. – Л.: Машиностроение, 1986. – 320 с.

42. Пешаль М. Моделирование сигналов и систем. – М.: Мир, 1981. – 300 с.

43. Шеннон К. Работы по теории информации и кибернетике. – М.: Изд-во иностр. лит., 1963. – 829 с.

44. Колмогоров А. Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. – 1965. – Т. 1. – Вып. 1. – С. 3–11.

45. Кульбак С. Теория информации и статистика. – М.: Наука, 1967. – 408 с.

46. Смит Ричард Э. Аутентификация: от паролей до открытых ключей: Пер. с англ. – М.: Издат. дом «Вильямс», 2002. – 432 с.

47. Киселев М. Средства добычи знаний в бизнесе и финансах / М. Киселев, Е. Соломин // Открытые системы. – 1997. – № 4.

48. Степаненко В. В. Методы и модели анализа данных: OLAP и Data Mining: Хранилища данных; OLAP – оперативный анализ; Data Mining – интеллектуальный анализ; Методы решения задач классификации, кластеризации и поиска ассоциативных правил: Учеб. пособие для вузов / В. В. Степаненко, М. С. Куприянов, А. А. Барсегян. – СПб.: БХВ-Петербург, 2004. – 336 с.

49. О необходимости создания национальных стандартов по тестированию средств высоконадежной биометрии / А. Ю. Малыгин, Ю. И. Олейник, Е. А. Малыгина, А. В. Семенов. – М.: ЦВНИ МО РФ, 2006. – Деп. в центральном справочно-информационном фонде МО РФ, справка № 14505. Сер. Б. Вып. № 75, инв. № Б5839.

50. Малыгин А. Ю. Тестирование высоконадежной биометрии / А. Ю. Малыгин, Ю. И. Олейник, Е. А. Малыгина // Безопасность информационных технологий: Сб. тр. науч.-техн. конф. – Пенза: ПНИЭИ, 2005. – С. 94–98.

51. Саймон Хайкин Нейронные сети: полный курс: Пер. с англ. – 2-е изд. – М.: Издат. дом «Вильямс», 2006. – 1104 с.

52. Малыгин А. Ю. К вопросу тестирования высоконадежной биометрии / А. Ю. Малыгин, Ю. И. Олейник, Е. А. Малыгина // Сб.

XIII военно-научной конференции РАРАН, АВН, ВУ ПВО ВС РФ. – Смоленск: Изд-во ВУ ПВО, 2005. – С. 183–184.

53. Методика контроля механизма центрирования входных данных нейростетевого преобразователя биометрия/код ключа доступа / А. И. Иванов, А. Ю. Малыгин, Д. Н. Надеев [и др.] // Нейрокомпьютеры и их применение: XI Всерос. конф. – М.: ИПУ РАН, 2005.

54. Волчихин В. И. Использование понятий «теории информации» при сравнительной оценке эффективности разнородных биометрико-криптографических механизмов защиты информации / В. И. Волчихин, А. И. Иванов, А. Ю. Малыгин // Вопросы защиты информации. – М.: Изд-во ВИНИ, 2006. – № 4. – С. 11–15.

55. Малыгин А. Ю. Применение биометрических технологий для идентификации личности / А. Ю. Малыгин, Ю. И. Олейник, А. В. Семенов. – М.: ЦВНИ МО РФ, 2005. – Деп. в центральном справочно-информационном фонде МО РФ, справка № 14341. Сер. Б. Вып. № 73, инв. № Б5771.

56. Пакет дополнительных стандартов, регламентирующих применение нейросетевых биометрико-криптографических механизмов при дистанционной биометрической аутентификации / В. И. Волчихин, А. И. Иванов, А. Ю. Малыгин, О. В. Ефимов // Нейрокомпьютеры и их применение: XI Всерос. конф. – М.: ИПУ РАН, 2005.

57. Оценка потенциальной информационной эффективности нейросетевой машины обогащения данных преобразователя биометрии в криптографический ключ / А. И. Иванов, А. Ю. Малыгин, Н. В. Капитуров, А. В. Семенов // Нейрокомпьютеры и их применение: XI Всерос. конф. – М.: ИПУ РАН, 2005.

Термины и определения

В настоящей работе применены следующие термины с соответствующими определениями:

1. Автоматическое обучение – обучение, осуществляемое автоматически без вмешательства человека и осмысления им промежуточных результатов обучения.

2. Атака перехвата – атака, направленная на перехват конфиденциальной биометрической информации в виде физического биометрического образа, электронного биометрического образа, вектора биометрических параметров, получаемого из них пароля или криптографического ключа.

3. Атака случайного подбора – атака, состоящая в подстановке случайных биометрических образов на вход преобразователя биометрия/код, либо случайный подбор личного ключа (пароля), образующегося на выходах преобразователя.

4. База биометрических образов – база, в которой собраны биометрические образы, представленные их примерами.

5. База естественных биометрических образов – база биометрических образов, полученных естественным путем их снятия с людей.

6. База синтетических биометрических образов – база биометрических образов, полученных их синтезом из естественных образов, или случайных чисел, отражающих статистические распределения биометрических параметров, присутствующие в базах естественных биометрических образов.

7. Биометрическая аутентификация – аутентификация пользователя, осуществляемая путем предъявления им своего биометрического образа.

8. Биометрические данные – данные с выходов первичных измерительных преобразователей физических величин, совокупность которых образует биометрический образ конкретного человека.

9. Биометрическая идентификация – преобразование совокупности примеров биометрических образов человека, позволяющее описать их стационарную и случайную составляющие, например, в

виде математического ожидания и дисперсий контролируемых параметров или, например, в виде параметров обученной сети искусственных нейронов.

10. Биометрический образ – образ человека, полученный с выходов первичных измерительных преобразователей физических величин, подвергающийся далее масштабированию и иной первичной обработке с целью извлечения из него контролируемых биометрических параметров человека.

Примечание. Биометрический образ – это континуум множества биометрических примеров, однако с конечной погрешностью континуум примеров может быть представлен всего несколькими различающимися примерами.

11. Биометрический образ «Свой» – биометрический образ легального пользователя.

12. Биометрический образ «Чужой» – биометрический образ злоумышленника, пытающегося преодолеть биометрическую защиту.

13. Биометрические образы «Все чужие»: Совокупность множества биометрических образов «Чужой», верно отражающая статистику попыток подбора злоумышленниками образов «Свой».

14. Биометрический механизм – механизм преобразования физического биометрического образа в вектор биометрических параметров или код ключа (пароля).

Примечание. Биометрический механизм является функционально неполной частью системы защиты или средства защиты информации.

15. Биометрические параметры – параметры, полученные после предварительной обработки биометрических данных.

Примечание. Параметрами могут являться, например, коэффициенты Фурье кривых колебаний пера при воспроизведении человеком рукописного пароля.

16. Вероятность ошибки первого рода – вероятность ошибочного отказа «Своему» пользователю в биометрической аутентификации.

17. Вероятность ошибки второго рода – вероятность ошибочной аутентификации «Чужого» как «Своего» (ошибочная аутентификация).

18. Высоконадежная биометрическая аутентификация – биометрическая аутентификация с приемлемой вероятностью ошибок

первого рода и гарантированно малой вероятностью ошибок второго рода, сопоставимой по своему значению с вероятностью случайного подбора кода неизвестного криптографического ключа при малом числе попыток подбора.

19. Динамический биометрический образ – биометрический образ, изменяемый человеком по своему желанию, например, рукописный образ слова-пароля.

20. Естественный биометрический образ – биометрический образ, представленный рядом примеров, полученных от реального человека.

21. Механизм биометрической аутентификации – функционально неполный фрагмент средства биометрической аутентификации, преобразующий биометрические данные, но не способный принимать аутентификационные решения высокой надежности из-за низкой размерности анализируемых векторов или отсутствия механизма криптографической аутентификации.

22. Нейросетевой преобразователь биометрия/код – заранее обученная искусственная нейронная сеть с большим числом входов и выходов, преобразующая частично случайный вектор входных биометрических параметров «Свой» в однозначный код криптографического ключа (длинного пароля) и преобразующая любой иной случайный вектор входных данных в случайный выходной код.

23. Преобразователь биометрия/код – преобразователь, способный преобразовывать вектор нечетких, неоднозначных биометрических параметров «Свой» в четкий однозначный код ключа (пароля). Преобразователь, откликающийся случайным выходным кодом на воздействие случайного входного вектора, не принадлежащего множеству образов «Свой».

24. Синтетический биометрический образ – биометрический образ, представленный несколькими примерами, полученными из фрагментов естественных примеров, или сгенерированный с помощью генераторов случайных чисел так, чтобы верно отражать статистические параметрны множества реальных биометрических образов.

25. Средство биометрической аутентификации – средство биометрической аутентификации, способное принимать аутентификационное решение неопределенного уровня надежности.

26. Среднестатистический биометрический образ – биометрический образ, представленный достаточным числом примеров, соответствующий некоторому среднему статистическому параметру, например, средней стабильности воспроизведения биометрических параметров.

27. Средство высоконадежной биометрической аутентификации – средство биометрической аутентификации, способное принимать аутентификационное решение высокой надежности, имеющее в своем составе биометрические механизмы преобразования биометрических данных в векторы биометрических параметров большой размерности, преобразователь биометрия/код ключа (пароля), механизм криптографической аутентификации.

28. Статический биометрический образ – образ, данный человеку от рождения, не изменяемый по воле человека, например, рисунок отпечатка пальца.

29. Тайный биометрический образ – биометрический образ, сохраняемый пользователем в тайне.

П р и м е ч а н и е. Для сохранения в тайне динамического биометрического образа необходимо сохранить в тайне пароль, порождающий его; для сохранения в тайне статического биометрического образ необходимо обеспечить анонимность пользователя.

30. Открытый биометрический образ – биометрический образ человека, общедоступный для наблюдения.

П р и м е ч а н и е. Обычно открытые биометрические образы являются статическими, однако и динамические биометрические образы могут быть открытыми, например, рукописный автограф человека.

31. Обучение биометрического средства – обучение биометрического средства аутентифицировать человека с заданными вероятностями ошибок первого и второго рода на одном или нескольких примерах биометрических образов «Свой».

32. Физический муляж – муляж, выполненный на физическом уровне, исходя из знания физического эффекта, на котором работает датчик считывания биометрического средства защиты и знания ин-

дивидуальных особенностей, подделываемого на физическом уровне биометрического образа.

33. Электронный муляж – электронные данные, имитирующие биометрические данные пользователя при тестировании или попытках обхода системы защиты.

П р и м е ч а н и е. Различают неслучайный электронный муляж – электронные биометрические данные реального пользователя априорно известны, например, перехвачены. Случайный муляж – электронные данные генерируются случайно. Частично случайный муляж – частично или полностью известные электронные биометрические данные реального пользователя искусственно размываются случайным шумом.