БИОМЕТРИЧЕСКАЯ ХЭШ-ФУНКЦИЯ

УДК: 004.032.26, 004.424.47

Майоров А.В

Дано определение нечеткой хэш-функции, указаны ее свойства и отличия от классической хэш-функции. Рассмотрена простая реализация нечеткой хэш-функции с помощью классических криптографических функций. Показаны недостатки нечеткой хэш-функции, построенной на основе нечеткой логики. Для хэширования биометрических данных предложена реализации нечеткой хэш-функции на основе искусственных нейронных сетей и выполнено доказательство ее свойств.

Введение

При работе с нечеткими данными, например, биометрическими интересной и практически значимой является задача преобразования их в стабильный код. Такая процедура называется нечетким хешированием и описывается как $H(x) = h$, где x – входное значение, h – выходной код или дайджест. Особенностью нечеткого хеширования является возможность получения на выходе некоторого подмножества $H_{Свой}$ только при подаче на вход вектора значений доверенного множества $X_{Свой}$ (множества «Своих» образов). Значения, не принадлежащие доверенному множеству (множество образов «Чужие»), дают выходное значение, отличное от $H_{Свой}$ с большой вероятностью.

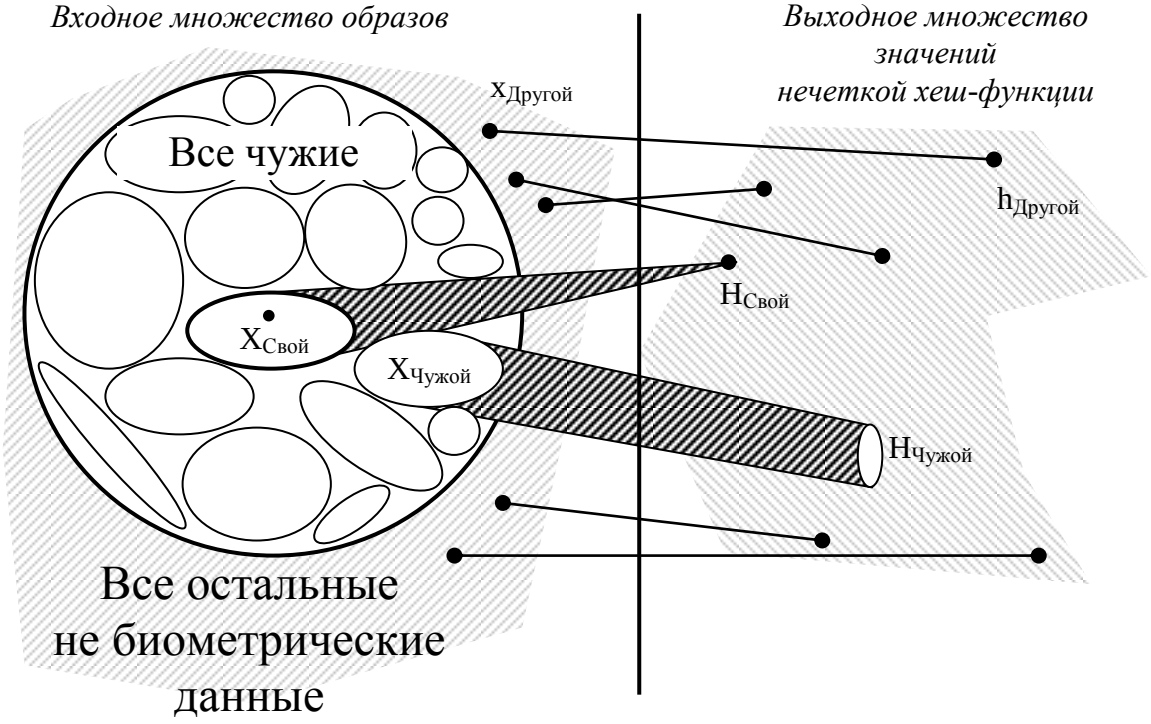


Рис 1. Проекция нечетких данных на выходное множество

Свойства нечеткой хеш-функции как классической хеш-функции:

1.1 Значительно уменьшает пространство поиска. $D(H) \geq E(H)$, где $D(H)$ - область определения хеш-функции, $E(H)$ – область значений хеш-функции;

1.2 Не изменяется во времени и пространстве. $\forall x_1 = x_2, \exists H(x_1) = H(x_2)$;

1.3 Имеет коллизии. $\exists H(x_1) = H(x_2)$, для $x_1 \neq x_2$ и $D(H)>E(H)$.

При решении задач по защите информации интересна реализация нечетких хеш-функций, обладающих свойствами криптографических хеш-функций [1]:

2.1 Является однонаправленной, то есть вычисление x по H(x) гарантированно сложно;

2.2 Обладает стойкостью к отысканию прообраза. По выходному коду h должно быть сложным отыскание такого значения x, для которого h = H(x);

2.3 Обладает стойкостью к отысканию второго прообраза. По входному значению x_1 должно быть сложным отыскание второго входа x_2, $x_1 \neq x_2$, такого, что $H(x_1) = H(x_2)$;

2.4 Обладает стойкостью к коллизиям. Должно быть сложным отыскание двух разных значений x_1 и x_2 таких, что $H(x_1) = H(x_2)$.

Проекция входных данных на выходное пространство значений нечеткой хэш-функции показано на Рис.1.

Простейшая реализация нечеткой хэш-функции

Простейший вариант нечеткой хеш-функции, можно реализовать с помощью классической хеш-функции по формуле:

$$H(X_{Свой}) + C = H_{Свой} \qquad (1)$$

где C – множество *значений-дополнений*, соответствующих каждому $x_i \in X_{Свой}$.

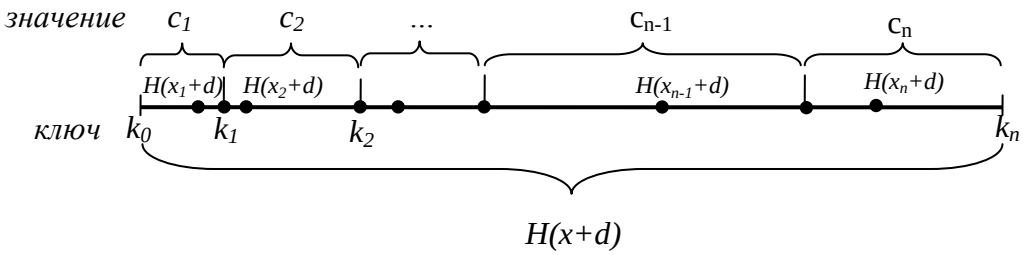


Рис 2. Бескомпроматное связывание $X_{Свой}$ и C с помощью ассоциативной таблицы

Для отображения $X_{Свой}$ в выходное множество значений $H_{Свой}$ без компрометации факта принадлежности входного значения множеству $X_{Свой}$ предлагается использовать дополнительное преобразование: сопоставить каждое значение $x_{Свой}$ с соответствующим $c_{Свой}$ при помощи ассоциативной таблицы *T*. *Ключом* таблицы являются значение $H(x+d)$, где d – фиксированное смещение; *значением* - число-дополнение c. Чтобы маскировать факт соответствия c_i конкретному значению x_i, выходное множество разбивается на n интервалов (по числу значений $X_{Свой}$) таким образом, чтобы в каждый из них проецировалось в случайную позицию только 1 $x_i \in X_{Свой}$, как это показано на Рис. 2. Каждому интервалу ставится в соответствие одно c_i.

Таким образом, формула для нечеткой хеш-функции запишется как:

$$T = \{([k_0,k_1],c_1),([k_1,k_2],c_2) \ldots , ([k_{n-1},k_n],c_n)\},$$
$$k=H(x + d),\ d = const \qquad (2)$$
$$H(x) = H(x) + T_k$$

Сложность реализации нечеткой хэш-функции при помощи множества классических хэш-функций заключается в необходимости предварительного формирования ассоциативной таблицы и поддержки функции поиска элементов в ней. Ограничение состоит в возможности работы только с дискретными множествами входных значений.

Нечеткая хэш-функция на основе нечеткой логики

Развитие предложенного способа построения нечеткой функции представляет схема получения стабильного выходного кода из зашумленных и биометрических данных [2]. В основе схемы лежит процедура выделения ключевых признаков во входном образе (*x* представляет собой вектор значений), реализованная с помощью нечеткой логики. Ожидая сильные искажения и зашумленность биометрических образов, в своих работах авторы [2, 3] предлагают получать достаточное для продукции стабильного кода число ключевых признаков, восстанавливая сильные признаки по слабым признакам при помощи кодов, исправляющих ошибки.

Однако, очевидно, такой подход не лишен недостатков. Во-первых, процедура отбора сильных признаков по входным образам требует формализации и является трудоемкой. Для выделения сильных признаков необходимы экспертные исследования и большая база входных образов. Во-вторых, процедура обучения нечеткого преобразователя не гарантирует успешный результат во всех случаях и, как следствие, не проводится в автоматическом режиме. В-третьих, авторы не учитывают то обстоятельство, что входные данные могут быть не только искажены и зашумлены, но и сильно коррелированны между собой, а значит, только малое число признаков входного образа линейно разделимо и может быть в явном виде использоваться для классификации. На примере динамического рукописного образа экспериментально подтверждено, что только около 5% данных биометрического образа обладает высоким показателем качества. В-четвертых, коды, исправляющие ошибки требуют значительной избыточности входного образа для контроля явных признаков [4].

Нечеткая хеш-функция на основе ИНС и последующим хешированием

Разработанные алгоритмы быстрого обучения искусственных нейронных сетей (ИНС), позволяют использовать для обучения и тестирования до 95% информации, размытой в биометрических образах. Поэтому заведомо «плохие» (сильно коррелированные) биометрические образы могут быть использованы для обучения ИНС. Процесс обучения полностью автоматизирован и позволяет проводить обучение даже на небольшой выборке образов доверенных множества $X_{Свой}$ [5].

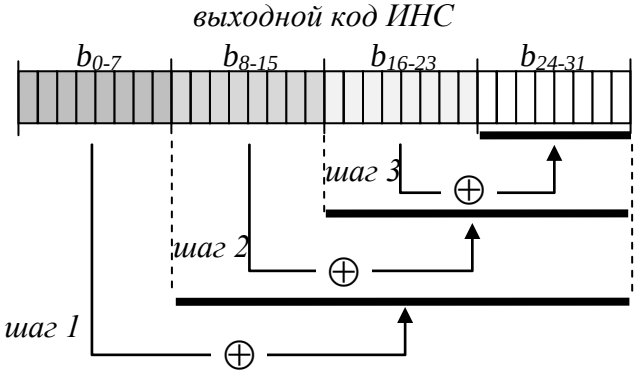


Рис 3. Схема смешивания стабильного кода для получения дайджеста хеш-функции

Предлагается реализовать нечеткую хеш-функцию с помощью однослойной ИНС. Биометрический образ подается на вход обученной ИНС. Получаемый на его выходе стабильный код подается на вход криптографической хеш-функции, осуществляющей смешивания его бит. В предложенной схеме нечеткий преобразователь служит для перехода от нелинейно разделимых множеств входных данных к линейно разделимым, а хеш-функция – для усложнения схемы восстановления исходных данных по выходному коду. В качестве

простой хеш-функции допустимо использовать блочную операцию сложения по модулю 2, применяемую со схемой на Рис. 3. Каждый следующий блок по ней используется для сокрытия оставшейся части выходов искусственной нейронной сети.

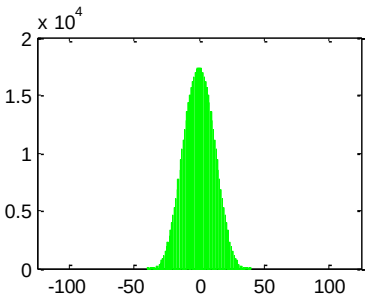
Достоинством ИНС является корректная работа в условиях частичного разрушения структуры сети. Теоретически возможно получение правильного кода на выходе при повреждении сети до 50%. Кроме того, ИНС обладают способностью к восстановлению информации, что позволяет получать корректный выходной код даже при преднамеренном искажении сигнала на входе. Преимущества ИНС поставили вопрос о возможности реализации нечетких хеш-функции, в том числе и обладающих свойствами криптографических, с использованием только математического аппарата ИНС.

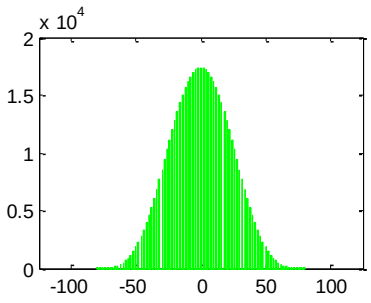
Нечеткая хэш-функция на основе ИНС

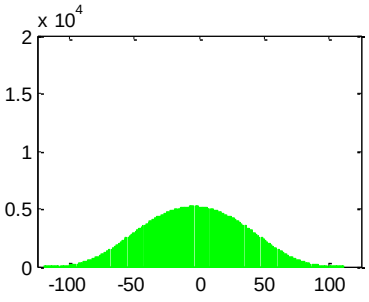
Важным вопросом при создании ИНС является вопрос о распределении коллизий в выходном пространстве значений. В идеальной хеш-функции, коллизии распределяются равномерно. В ИНС это не так. Приняв формулу для расчета выходного значения одного слоя ИНС за $\boldsymbol{a} = \boldsymbol{f}(\boldsymbol{Wp}+\boldsymbol{b})$, где $\boldsymbol{a}$ – выходной вектор значений ИНС; $\boldsymbol{f}$ – вектор нелинейных функций; $\boldsymbol{p}$ – входной вектор значений; $\boldsymbol{W}$ - матрица весов слоя; $\boldsymbol{b}$ – вектор значений-подставок, отметим ее свойства.

Комбинации всех возможных входных значений линейная часть формулы для каждого нейрона в случае сбалансированных весов (сумма весов равна 0) дает некоторое нормальное распределение, что показано на Рис. 4а-б. Плотность распределения коллизий выходных значений зависит от выбора весов: выбор значений весов, имеющих меньшее количество сомножителей приводит к уменьшению числа коллизий, что показано на Рис. 4в. Неравномерное изменение весовых коэффициентов приводит к смещению центра распределения и появлению волн коллизий. Изменение значения подставки b приводит к сдвигу распределения относительно начала координат и, соответственно, изменяет распределение коллизий на выходе дискретной нелинейной функции. Очевидно, что в распределении существуют *слабые* позиции, в которых число коллизий меньше, чем в соседних. Поэтому суперпозиция полей может ослаблять сложность расчета выходного решения.

Использование нелинейных дискретных (функция Хэвисайда, знака, остатка от деления, функций с насыщением), а также периодических функции позволяет сбалансировать распределение коллизий на выходом пространстве значений. Поскольку все нейронных дают коллизии, т.е. сужают пространство поиска, то свойство 1.1 для хеш-функций в нейронных сетях выполняется. Свойство 1.2 выполняется тогда, когда используются нейронные сети «без памяти», т.е. не динамические нейронные сети, обучаемые однократно на обучающей выборке. Свойство 1.3 выполняется по определению, что следует из Рис 4.

а) 

б) 



c)

Рис 4. Линейная часть нейрона: а) со сбалансированными весами; б) с увеличенным диапазоном весов; в) со сбалансированными весами из простых чисел;

Линейная часть ИНС представляет собой множество одиночных сумматоров, характеризуемых *нормальным* распределением коллизий. Рост числа коллизий зависит от диапазонов входных и выходных значений и является полиномиальным. Поэтому для ИНС, имеющего число входов большее числа выходов, получение единственного решения является маловероятным, следовательно, свойство в общем случае 2.1 выполняется. А поскольку дискретные и циклические нелинейных функций активации балансирует распределение коллизий в выходном поле значений, то свойство 2.1 выполняется и для ИНС.

Теперь покажем, что для больших ИНС с 2-мя и более слоями выполняется свойство 3.2. Для этого будем ужесточать требования к ИНС и исследовать ее на предмет выполнения свойства 3.2.

Рассмотрим однослойную ИНС с линейной функцией активации, содержащей m входов и n выходов (нейронов). Уравнение этой сети запишется в виде:

$$\begin{cases} w_{11}p_1 + \ldots + w_{1m}p_m + b_1 = a_1, \\ \ldots, \\ w_{n1}p_1 + \ldots + w_{nm}p_m + b_n = a_n, \end{cases} \qquad (3.1)$$

Очевидно, что нахождения корней $p_1..p_n$ СЛАУ просто: сложность равна $O(m^3)$. Если $n > m$, лишние уравнения можно исключить. Если m > n, то m-n уравнений отбрасываются, а $p_n..p_m$ заменяются нулями. Значит, свойство 3.2 для однослойной ИНС с линейной функцией активации не выполняется.

Возьмем ту же однослойную ИНС усложненную нелинейной функцией активации и ограничением выходных значений (счетным множеством или диапазоном), описываемую уравнением:

$$\begin{cases} f(w_{11}p_1 + \ldots + w_{1m}p_m + b_1) = a_1, \\ \ldots, \\ f(w_{n1}p_1 + \ldots + w_{nm}p_m + b_n) = a_n, \end{cases} \qquad (3.2)$$

Если обратная функция f^T для f вычислима, то уравнение *3.2* сведется к *3.1*. Но, если в качестве нелинейной функции используется дискретная или периодическая функция, не имеющая обратной, то f^T для нее запишется как множество частных решений и уравнение *3.2* примет новую форму:

$$\begin{cases} w_{11}p_1 + \ldots + w_{1m}p_m + b_1 = \boldsymbol{o}_1 = \{o_1,\ldots,o_k\}_1, \\ \ldots, \\ w_{n1}p_1 + \ldots + w_{nm}p_m + b_n = \boldsymbol{o}_n = \{o_1,\ldots,o_k\}_n, \end{cases} \qquad (3.3)$$

Поскольку все решения о удовлетворяют исходному уравнению, то для нахождения корней уравнения достаточно выбрать любой набор коэффициентов $\boldsymbol{o}$ и решить СЛАУ.

Ужесточим требования к ИНС и потребуем, чтобы имело место ограничение диапазона входных значений $p_1 \ldots p_m$. В этом случае не все комбинации решений f^T будут

удовлетворять допустимым входным значениям. Следовательно, выбор первого или случайного набора не будет являться строгим решением. Для нахождения решения в общем случае придется находить все допустимые решения уравнения. Перепишем СЛАУ 3.3, представив $\boldsymbol{o}_1 \ldots \boldsymbol{o}_n$ в качестве неизвестных:

$$\begin{cases} v_{11}\boldsymbol{o}_1 + \ldots + v_{1n}\boldsymbol{o}_n = p_1, \\ \ldots, \\ v_{m1}\boldsymbol{o}_1 + \ldots + v_{mn}\boldsymbol{o}_n = p_m, \end{cases} \qquad (3.4)$$

Поскольку по условию входные значения являются дискретными значения или имеют ограничение на область значений, то уравнение 3.4 преобразуется в систему неравенств. Ранее в литературе [6] Черниковой Н.Е. было показано, что сложность решения системы линейных неравенств можно оценить как:

$$O(n^{m/2+1}) \qquad (3.5)$$

где n – число неизвестных, m – число неравенств, в соответствии с 3.4. Следовательно, для ИНС сети, с дискретной или периодической нелинейной функцией, сложность гарантированного нахождения корней уравнения является экспоненциальной. Поскольку, чтобы реализовать фильтрацию входных значений по диапазону или дискретному множеству нужен второй слой ИНС, свойство 3.2 строго выполняется для двухслойных сетей ИНС с числом выходов, меньшим числа входов. Свойство 3.3 выполняется, поскольку сложность отыскания второго прообраза также зависит от сложности решения уравнения *3.4.* Свойство 3.4 выполняется, поскольку для нахождения коллизий придется разрешить уравнение *3.4.*

Мера близости хеш-функций, использующих коррелированные данные

Правильный выбор весовых коэффициентов и устранение частных случаев позволяется избежать ослабления стойкости систем к атакам подбора. Однако, на практике, из-за высокой коррелированности входных данных, стойкость нечеткой хеш-функции на базе ИНС может значительно снижаться. Для оценки близости нечеткой хеш-функции предлагается ввести меру близости к идеальной хеш-функции через вычисление средней стабильности выходных бит кода нечеткой хеш-функции.

Для идеальной криптографической хеш-функции незначительное изменение входного значения должно приводить к сильному изменению выходного значения. На Рис 5.а показан результат тестирования криптографической хеш-функции MD5 на большой базе образов. Каждый бит выходного значения принимает 1 и 0 практически равновероятно, что подтверждает высокое качество криптографической хеш-функции.

Для некачественной хеш-функции, в которой выходные биты слабо зависят от входных значений, будет иметь место преобладание отдельных значений битов. В этом случае распределение примет вид как на Рис 5.б. На рисунке показаны результаты тестирования, при котором 78 бит всегда принимали значения «0», а оставшиеся 256-78 бит 0 только «1».

Нечеткие ИНС хеш-функции по результатам тестирования будут лежать между крайними состояниями. Сеть, показанная на Рис 5.г, обладает худшими свойствами по сравнению с сетью Рис 5.д, поскольку вынос отдельных бит гораздо ближе к некачественной сети, чем у другой сети. Отметим таким образом, что дисперсия распределения вероятностей появления отдельных значений в выходном коде относительно равновероятного состояния может использоваться для сравнения хеш-функций, использующих коррелированные данные на входе.

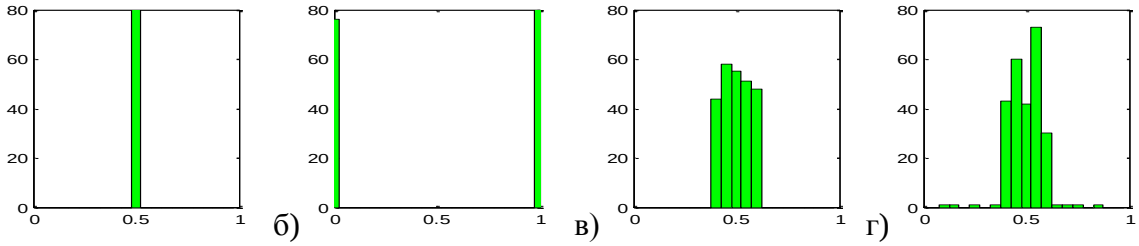


Рис 5. Распределение стабильности бит выходного кода для: а) MD5 б) некачественной хеш-функций в) нечеткой ИНС функции №1 г) нечеткой ИНС функции №2

Качество нечеткой ИНС хеш-функции сильно зависит не только от выбора правильной настройки баланса весов сети на «Своих» образах, но и от базы образов «Чужой», в частности от их корреляции. На Рис. 6.а-д показано как изменяется распределение стабильности нечеткой хеш-функции, при сдвиге входных параметров на некоторое значение. Чем больше одновременный сдвиг, тем больше корреляция образа. Из рисунка следует, что сильная корреляция может приводить к значительному ослаблению нечеткой хеш-функции.

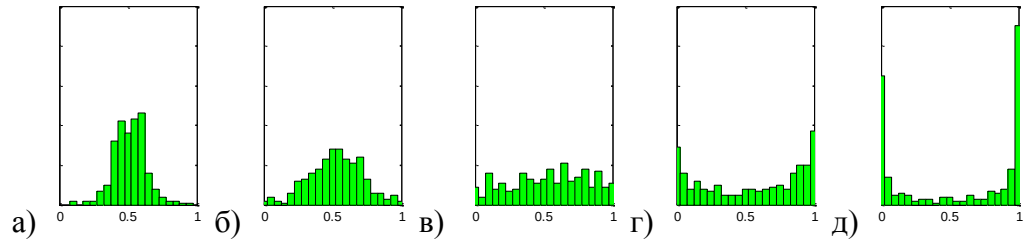


Рис 6. Усиление корреляции через смещение входов на: а) 0.05 б) 0.1 в) 0.2 г) 0.4 д) 1.0

Заключение

Таким образом рассмотрены нечеткие хеш-функции, дано их определение и основные свойства. Описаны различные способы их реализации с использованием классических хеш-функций, нечеткой логики, искусственных нейронных сетей. Доказано, что ИНС могут использоваться в качестве нечетких хеш-функции, обладающих свойствами как классических, так и криптографических хеш-функций. Для измерения качества нечеткой хеш-функции на базе ИНС предложена мера близости к идеальным хеш-функциям.

Предполагается развить работу в направлении углубленного исследования свойств нечетких хеш-функций на базе ИНС и практического их использования при решении задач защиты информации в динамических системах.

СПИСОК ЛИТЕРАТУРЫ

1. *Киви Бердж.* Хеш-пятилетка: Объявлен конкурс на новый стандарт хеш-функции //Компьютерра, №9 от 09.03.2007;
2. *Cavoukian Ann, Stoianov Alex.* Biometric Ecryption: A Positive-Sum Technology that Archieves Strong Authentication, Security and Privace // Toronto, Ontario, March 2007;
3. *Dodis Y., Reyzin L., Smith A.* Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data// April 2004;
4. *Питерсон У., Уэлдон Э.* Коды, исправлющие ошибки /монография, под ред. Добрушиной Р.Л., Самойленко С.И. //-М.:МИР, 1976;
5. *Иванов А.И.* Быстрые алгоритмы обучения нейросетевых механизмов биометрико-криптографической защиты информации;
6. *Черников С.Н.* Линейные неравенства //монография. М.: Наука, 1968;