Труды научно-технической конференции кластера пензенских предприятий,
обеспечивающих **БЕЗОПАСНОСТЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ.**
Том 8, Пенза-2012, с. 18-21 (http://пниэи.рф/activity/science/BIT/T8-p18.pdf)

ОЦЕНКА ЭНТРОПИИ БИОМЕТРИЧЕСКИХ ОБРАЗОВ ЧЕРЕЗ ПЕРЕХОД К ДИСКРЕТНОМУ ПРЕДСТАВЛЕНИЮ С АССИМЕТРИЧНЫМ РАСПРЕДЕЛЕНИЕМ МЕРЫ ХЕММИНГА

Куликов С.В. (г. Пенза)

Задача оценки энтропии для биометрических образов, представляемых вектором n (к примеру, 480 в технологии рукописного ввода ОАО ПНИЭИ) параметров не может быть решена с помощью классического подхода. Классическое определение информационной энтропии и формулы для ее расчета были введены Шенноном [1, 2, 3]. По Шеннону для некоторой конечной системы кодов (например, кодов букв языка) может быть вычислена энтропия этой системы через вероятности появления того или иного кода в анализируемой последовательности. Энтропию одиночных кодов оценивают через вероятности их появления:

$$H(x) = -\sum_{i=1}^{S} P(x_i) \cdot \log_2(P(x_i)), \quad (1),$$

Для вычисления многомерной энтропии группы из «*n*» символов в соответствии с классическим подходом необходимо использовать следующее выражение:

$$H(x_1,...,x_n) = -\sum_{1i=1}^{S}\sum_{2i=1}^{S}.....\sum_{ni=1}^{S} P(x_{1i}...x_{ni}) \cdot \log_2(P(x_{1i}, x_{2i}...x_{ni})) \quad (2)$$

Классический подход к расчету многомерной энтропии, к сожалению, технически не реализуем при исследовании биометрических образов. Основной недостаток классического многомерного вычисления энтропии состоит в том, что требуются огромные размеры исходных данных. Как правило, требуется массив исходных данных, размером превышающий число возможных состояний исследуемого кода. Кроме того, классический метод вычисления многомерной энтропии требует огромных вычислительных затрат.

Выходом является вычисление многомерной энтропии в пространстве мер Хемминга. Это позволяет перенести задачу оценки биометрических образов энтропии из поля 2^n в пространство мер Хемминга размера n+1. Очевидно, что переход от обычной классической энтропии к энтропии в пространстве мер Хэмминга является нелинейным преобразованием. Энтропия в пространстве мер Хэмминга является нелинейной функцией при малых размерностях *n* менее 100, при более высоких размерностях она практически линейно зависит от размерности, как и классическая энтропия [4]. Энтропия в пространстве мер Хэмминга вычисляется следующим образом:

$$\widetilde{H}(n) = -\sum_{i=0}^{n} P(h_i) \cdot \log_2(P(h_i)) \quad (3)$$

Для перехода к вычислению энтропии в пространстве мер Хемминга необходимо иметь возможность представлять биометрические образы неким кодом. Предлагается для каждого параметра вектора параметров биометрического

образа установить пороговую функцию, преобразующую параметр в бит кода (рисунок 1)

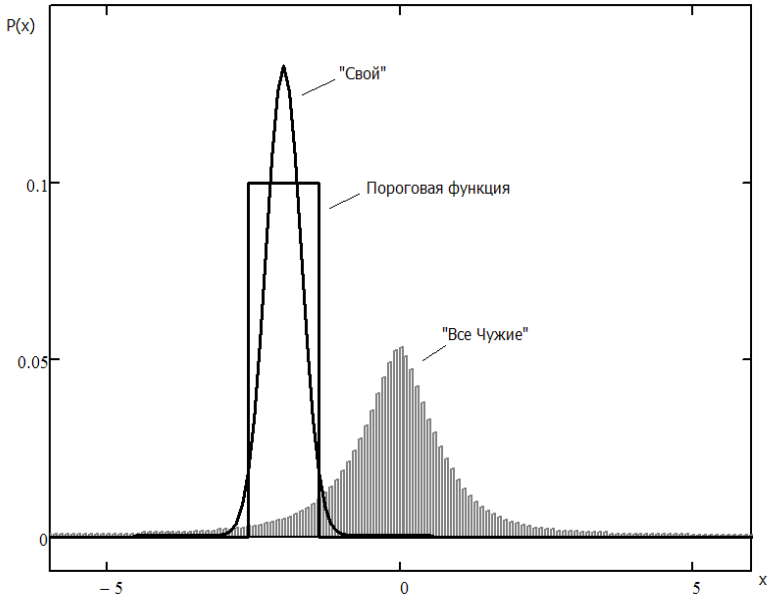

Рисунок 1 – Пороговая функция, используемая для разделения области значений параметра на «Свой» и «Чужой»

Тогда каждый образ «Чужой» относительно некоторого образа «Свой» представляется вектором бит $\mathbf{V}(n)$, где n – количество параметров образа. Вычисление меры Хемминга получившегося кода относительно кода «Свой» (все биты которого равны нулю) или, проще, подсчёт количества нулевых бит переводит биометрический образ в пространство кодов Хемминга. Такой переход позволяет проводить оценку энтропии биометрических образов в поле кодов Хемминга. При этом вычисляемая энтропия является энтропией биометрического образа «Свой» в пространстве биометрических образов, так как переход в поле кодов Хемминга осуществляется относительно образа «Свой».

Ошибка первого рода такого преобразования зависит от интервала пороговой функции, в котором она принимает значение «1». Рекомендуется выбирать интервал, равный

$$\left[E\left(p_i\right)-3D\left(p_i\right),E\left(p_i\right)+3D\left(p_i\right)\right] \quad (4)$$

где $E(p_i)$ – математическое ожидание параметра образа «Свой»; $D(p_i)$ – дисперсия параметра образа «Свой».

Распределение меры Хемминга для базы образов «Все Чужие» относительно некоторого образа «Свой» является ассиметричным (рисунок 2). На рисунке вертикальными линиями показана мера Хемминга для образов «Свой». Для выбранного интервала

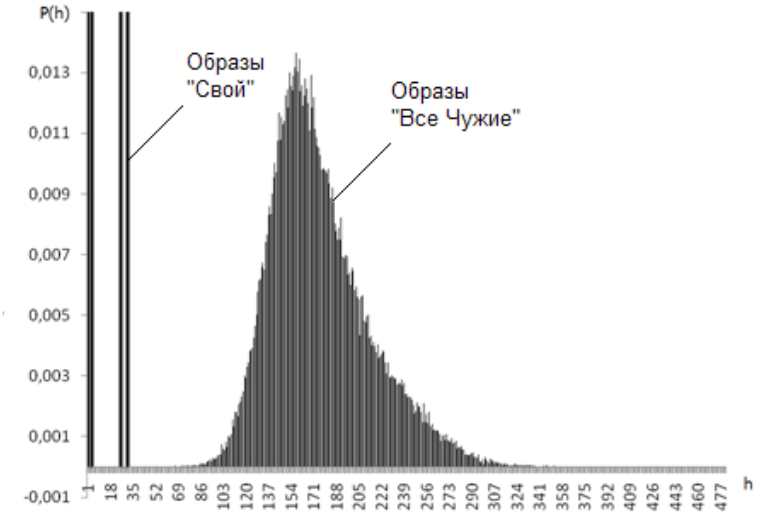

Рисунок 2 - Ассиметричное распределение меры Хемминга для базы образов «Все Чужие»

Подобное ассиметричное распределение может быть представлено смесью нормальных законов распределения. Существуют EM-алгоритмы, позволяющие итерационно вычислять параметры смеси нормальных законов при аппроксимации заданного распределения [4]. Смесь нормальных законов распределения является суммой взвешенных нормальных распределений.

В первом приближении энтропия биометрического образа «Свой» в пространстве образов «Все Чужие» может быть вычислена как информация события «образ является образом Свой». Для информации этого события необходимо оценить вероятность попадания образа «Чужой» в область значений «Свой». Областью значения «Свой» в пространстве меры Хемминга является интервал от 0 до максимальной среди образов «Свой» меры Хемминга *s*. Для каждого компонента смеси – нормального закона распределения вероятность попадания образа в область «Свой» определяется следующим выражением:

$$P_i = \frac{1}{\sigma_i \sqrt{2\pi}} \int_0^s \exp\left(- \frac{(x - \mu_i)^2}{2\sigma_i^2} \right) dx \qquad (5)$$

Положение и соотношение компонентов смеси нормальных законов распределения остаются постоянными для различных образов «Свой». Первый компонент смеси имеет наибольший вес и настолько большую вероятность попадания в область «Свой», что остальными компонентами смеси можно пренебречь (рисунок 3).

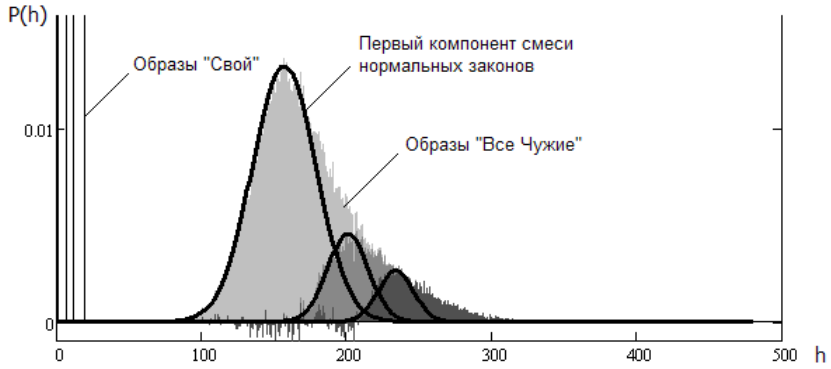

Рисунок 3 – Аппроксимация распределения мер Хемминга смесью нормальных законов

Таким образом, нет необходимости находить компоненты смеси с помощью EM-алгоритма, математическое ожидание первого компонента смеси m_1 может быть найдено по положению максимума плотности распределения $max(\mathbf{P}(h))$. Дисперсия первого компонента находится как:

$$d_1 = 2 \cdot \sum_{h=0}^{m} (m - P(h))^2 \qquad (6)$$

Энтропия биометрического образа может быть определена через вероятность угадывания его случайной попыткой «Чужого»:

$$H \approx -P_1 \cdot \log_2(P_1) \qquad (7),$$

где вероятность ошибок второго рода первого компонента нормальных смесей находится как:

$$P_1 = \frac{1}{\sqrt{2\pi \cdot d_1}} \int_0^h \exp\left(-\frac{(x - m_1)^2}{2d_1} \right) dx \qquad (8).$$

Получается, что приведенные выше вычисления примитивны, однако они позволяют достаточно точно осуществлять оценку энтропии непрерывных биометрических образов еще до их дискретизации преобразователем биометрия-код. Быстрое и корректное вычисление многомерной энтропии непрерывных биометрических образов удалось осуществить, только благодаря некоторой псевдодискретизации (рис. 1) биометрического образа по каждому из его непрерывных параметров. Прямое вычисление энтропии непрерывных 416-мерных биометрических образов через вычисление 416 вложенных интегралов с неизвестной 416-мерной подинтегральной плотностью распределения значений технически невыполнимо.

Литература

1. Стратанович Р.Л. Теория информации. //Москва: Советское радио, 1975, 424с.
2. Кульбак С. Теория информации и статистика. //Москва: Наука, 1967, 408с.
3. Яглом А.М., Яглом И.М. Вероятность и информация. //Москва: ДомКниги, 2007, 512с.
4. Королёв В.Ю. EM-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. Теоретический обзор. //Москва: Издательство ИПИ РАН, 2007, 102с.

Материалы поступили 08.03.2012, опубликовано в Интернет 20.05.2012 по положительной рецензии д.т.н., доц. Иванова А.И. (Пенза).