Труды научно-технической конференции кластера пензенских предприятий,
обеспечивающих **БЕЗОПАСНОСТЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ.**
Том 10, Пенза-2016, с. 5-9 (http://пниэи.рф/activity/science/BIT/T10-p5.pdf)

Банных Андрей Григорьевич (г. Пенза)
УДК: 519.2; 519.6.

Энтропийная оценка качества непрерывных биометрических данных малых обучающих выборок

В настоящее время активно идут процессы информатизации современного общества. Привычными стали услуги электронной почты, электронные услуги банков, электронные услуги государственных служб, осуществляемые через личные кабинеты граждан. Безопасность электронных услуг в основном строится на использовании коротких логинов и коротких, легко запоминаемых паролей доступа. К сожалению, пользователи электронных услуг не желают запоминать длинные, стойкие к атакам подбора пароли, состоящие из случайных символов.

Одним из путей усиления защиты персональных данных является использование биометрии личности человека при его аутентификации. В частности для этой цели могут быть использованы нейросетевые преобразователи рукописных образов в длинный случайно сгенерированный пароль доступа [1]. Запомнить рукописное слово, состоящее из нескольких букв на много проще, чем длинный пароль доступа, состоящий из 32 случайных символов, набираемых в двух регистрах клавиатуры.

Предположительно, что в ближайшем будущем появится достаточно большое число приложений, которые будут с использованием искусственных нейронных сетей преобразовывать различные биометрические образы человека (рукописные образы, голосовые образы, рисунки подкожных кровеносных сосудов, рисунки радужной оболочки глаза, …) в длинные коды доступа к личным электронным приложениям. В связи с этим возникает необходимость стандартизации преобразователей биометрия-код и тестирования качества их обучения [2].

В свою очередь качество обучения нейросетевого преобразователя биометрия-код сильно зависит от качества биометрического образа «Свой», использованного при обучении. Возникает задача предварительной оценки качества непрерывных биометрических данных. На данные момент стандарт [3] рекомендует оценивать такие параметры биометрических данных как их показатель стабильности, их показатель уникальности и их показатель качества. Очевидно, что трех перечисленных показателей одного биометрического параметра недостаточно. Как минимум, необходимо учитывать коэффициенты коррелированности биометрических параметров. Проблема их учета осложняется тем, что биометрических параметров много. Так приложение [1] учитывает 416 биометрических параметров и дает возможность наблюдать значения всех этих параметров для вводимых своею рукою рукописных знаков.

Второй проблемой является то, что оценку качества непрерывных биометрии параметров приходится выполнять на ограниченном числе примеров биометрического образа «Свой». То, что континуумы возможных состояний биометрических параметров представлены малым числом примеров приводит к эффекту квантования данных или к эффекту появления шума квантования. На рисунке 1 иллюстрируется влияние нежелательного эффект квантования, возникающего при использовании обучающей выборки, состоящей из 9 примеров.

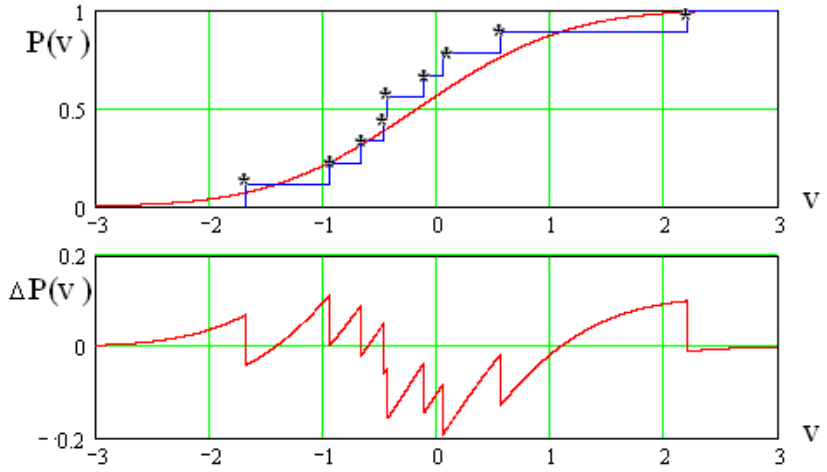

Рис. 1. Представление гладкой функции вероятности - P(v) ее ступенчатым приближением $\tilde{P}(v)$, приводящее к появлению случайного шума ошибок квантования – ΔP(v)

Очевидно, что шум квантования ΔP(v) непрерывных биометрических данных приводит к ощутимым ошибкам вычисления математического ожидания - $\Delta_{E(v)}$, ощутимым ошибкам вычисления стандартных отклонений - $\Delta\sigma_{(v)}$ и ощутимым ошибкам вычисления коэффициентов корреляции - Δ_r.

Одним из способов интегральной оценки качества биометрического образа «Свой» является вычисление условной энтропии вектора его непрерывных данных $\bar{v}$ относительно вектора биометрических данных образов «все Чужие» - $\bar{\xi}$. В простейшем одномерном случае мы можем воспользоваться гипотезой нормального закона распределения значений биометрических параметров данных образа «Свой» и множества образов «все Чужие». Иллюстрация процедуры вычисления одномерной энтропии дана на рисунке 2.

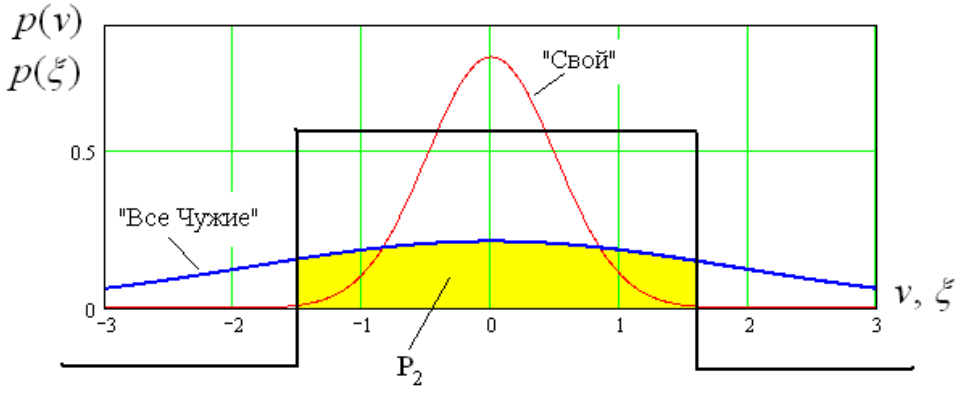

Рис. 2 Вычисление одномерной относительной энтропии

Из рисунка 2 видно, что стандартное отклонение данных «все Чужие» - $\sigma_{(\xi)}$ много больше, чем стандартное отклонение данных образа «Свой» - $\sigma_{(v)}$. Это позволяет по любому контролируемому биометрическому параметру осуществить квантование данных так, что бы все данные образа «Свой» давали состояние «1». При таком способе квантования данных вероятность ошибок первого рода отсутствует $P_1 = 0$. Присутствует только вероятность ошибок второго рода - $P_2 \neq 0$ (на рисунке 2 площадь, соответствующая вероятности

ошибок второго рода отмечена заливкой). Как следствие относительная энтропия биометрического параметра может быть оценена следующим образом:

$$H\left(\nu / \xi\right) \approx -\log_2(P_2) \qquad (1).$$

Знак приближения в выражении (1) появляется в силу того, что мы не можем точно указать пороги квантования из-за ошибок вычисления $\Delta_{E(\nu)}$ и $\Delta\sigma_{(\nu)}$.

Очевидно, что аналогичным образом мы можем вычислить двухмерную относительную энтропию для двух биометрических параметров:

$$H\left(\nu_1, \nu_2 / \xi_1, \xi_2\right) \approx -\log_2(P_2) = -\log_2(P("11")) \qquad (2).$$

Для оценки двухмерной энтропии (2) необходимо вычислить вероятность появления состояния пары разрядов «11» при воздействии на двухмерный квантователь примерами образов «все Чужие». Далее по индукции запишем выражение для m-мерных относительных совместных значений энтропии:

$$H\left(\nu_1, \nu_2, \ldots, \nu_m / \xi_1, \xi_2, \ldots, \xi_m\right) \approx -\log_2(P_2) = -\log_2(P("11\ldots1")) \qquad (3).$$

Оценки (3) технически осуществимы до m=16, далее начинаются технические трудности, обусловленные необходимостью иметь базы образов «все Чужие» очень больших размеров. Однако эта технологическая трудность вполне преодолима, если отказаться от ожидания редких событий и перейти к их предсказанию в пространстве расстояний Хэмминга [4, 5]. В итоге мы можем оценить условную энтропию (3) очень высокой размерности.

В частности применительно к данным среды моделирования «БиоНейроАвтограф» приходится иметь дело с размерностью m=416, соответственно, расстояния Хэмминга для кодов примеров образа «Свой» и «все Чужие» будут описываться следующими соотношениями:

$$\begin{cases} h(v) = \sum_{i=1}^{416} ("1") \oplus ("c_i") \\ h(\xi) = \sum_{i=1}^{416} ("1") \oplus ("x_i") \end{cases} \qquad (4),$$

где $"c_i"$ - i-тый разряд, кодов «Свой», $"x_i"$ -i-тый разряд выходных кодов образов «все Чужие».

Очевидно, что энтропия выходных кодов примеров образа «Свой» много меньше, чем энтропия выходных кодов, которые дают образы «все Чужие»:

$$H("\bar{c}") << H("\bar{x}") \qquad (5).$$

Соотношение энтропий (5) характерно для преобразователей биометрия-код любого типа. Если бы число примеров образа «Свой» было велико (порядка 200) и гипотеза нормальности законов распределения точно выполнялась, то энтропия кодов «Свой» должна быть практически нулевой. На самом деле данных в обучающей выборке мало и распределения данных описываются нормальным законом только в первом приближении [6]. Как итог, энтропия выходных кодов примеров образа «Свой» существенно отличается от нуля. Это приводит к тому, что расстояния Хэмминга (4) являются случайными величинами, каждая из которых имеет собственную плотность распределения значений. Характерные примеры таких плотностей распределения значений приведены на рисунке 3 (данные получены в среде моделирования «БиоНейроАвтограф» [1]).

Очевидно, что по нескольким примерам образа «Свой» мы можем получать для них последовательность расстояний Хэмминга. Далее для этой последовательности мы можем рассчитать математическое ожидание $E(h(v))$ и

стандартное отклонение $\sigma_{(h(v))}$. Пользуясь этими статистическими параметрами мы можем воспользоваться бета распределением и оценить максимальное значение расстояний Хэмминга возможное для его распределения - $p(h(v))$. В нашем случае оно составляет $\max(h(v)) = 62$ бита.

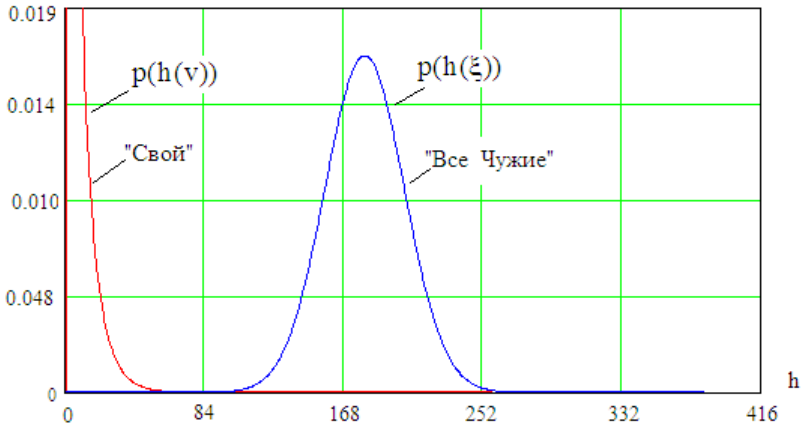

Рис. 3. Распределения расстояний Хэмминга для кодов «Свой» и кодов «все Чужие»

Кроме того, мы можем по данным о расстояниях Хэмминга найти математическое ожидание $E(h(\xi))$ и стандартное отклонение $\sigma_{(h(\xi))}$ для примеров образов «все Чужие». Из рисунка 3 видно, что плотность распределения значений $p(h(\xi))$ близка к нормальной. Это означает, что мы можем воспользоваться гипотезой о нормальном законе распределения значений расстояний Хэмминга для кодов «все Чужие» для предсказания вероятности ошибок второго рода:

$$P_2 \approx \frac{1}{\sigma_{(h(\xi))} \cdot \sqrt{2\pi}} \cdot \int_{-\infty}^{\max(h(v))} \exp\left\{ \frac{-\left(E(h(\xi)) - u\right)^2}{2\left(\sigma_{(h(\xi))}\right)^2} \right\} \cdot du \qquad (5).$$

Оценки (5) достаточно, что бы вычислить 416 мерную условную энтропию образа «Свой» по формуле (3). Получается, что переходя в пространство расстояний Хэмминга мы имеем возможность оценивать энтропию биометрических данных как угодно высокой размерности. Шеноновскими процедурами этого сделать нельзя, так как они построены на наблюдении редких событий. Описанные выше преобразования построены не по Шеннону. Процедуры вида (3), (4), (5) используют предсказания вероятности ошибок второго рода, их вычислительная сложность слабо зависит от размерности решаемой задачи.

Все это дает возможность сравнивать между собой биометрические образы по их информативности. Очевидно, что информативность биометрического образа тем выше, чем больше его энтропия. Как следствие качество и стабильность двух биометрических образов можно сравнить через отношение их энтропий при условии одинаковой их размерности. Тот образ будет информативнее, чья условная энтропия выше. Такой способ сравнения биометрических данных достаточно прост и дает устойчивые результаты. В частности применительно к данным рукописных образов среды моделирования «БиоНейроАвтограф» удается наблюдать значительные вариации их 416 мерной энтропии. Если самые плохие образы имеют энтропию порядка 10 бит, то самые хорошие дают энтропию от 70 бит до 90 бит.

Следует отметить, что предложенная схема энтропийной оценки качества биометрических данных во многом повторяет идеологию построения так

называемых «нечетких экстракторов» [7, 8, 9]. Фактически нечеткие экстракторы осуществляют квантование сырых биометрических данных и вынуждены корректировать большое число ошибок. В рассмотренной нами ситуации «нечеткий экстрактор» должен иметь избыточный код, который должен быть способен исправлять до 62 бит. В этом случае «нечеткий экстрактор» будет работоспособен. Получается, что «нечеткие экстракторы» выгодно применять для сравнения качества биометрических образов в силу простоты, осуществляемых ими преобразований.

Следует особо отметить устойчивость приведенных в данной статье вычислений по отношению к размерам обучающих выборок. Практика показала, что вычисления имеют высокий уровень повторяемости, если используется выборка из 20 примеров образа «Свой» и порядка 200 примеров образов «все Чужие».

Литература:

1. Иванов А.И., Захаров О.С. Среда моделирования «БиоНейроАвтограф» размещена на сайте АО «ПНИЭИ» http://пниэи.рф/activity/science/noc.htm. Продукт создан лабораторией биометрических и нейросетевых технологий ОАО «ПНИЭИ» в период 2009-2015 г.г. для свободного использования университетами России, Белоруссии, Казахстана.
2. ГОСТ Р 52633.3-2011 «Защита информации. Техника защиты информации. Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора».
3. ГОСТ Р 52633.1-2009 «Защита информации. Техника защиты информации. Требования к формированию баз естественных биометрических образов, предназначенных для тестирования средств высоконадежной биометрической аутентификации»
4. Язов Ю.К. и др. Нейросетевая защита персональных биометрических данных. //Ю.К.Язов (редактор и автор), соавторы В.И. Волчихин, А.И. Иванов, В.А. Фунтиков, И.Г. Назаров // М.: Радиотехника, 2012 г. 157 с. IBSN 978-5-88070-044-8.
5. Ахметов Б.С., Надеев Д.Н., Фунтиков В.А., Иванов А.И., Малыгин А.Ю. Оценка рисков высоконадежной биометрии. Монография. Алматы: Из-во КазНТУ им. К.И. Сатпаева, 2014 г.- 108 с.
6. Назаров И.Г., Иванов С.М., Ефимов О.В., Плотников В.Г. Согласование шага дискретизация входных биометрических данных и их динамического диапазона для универсального нейросетевого преобразователя биометрия-код «Нейрокомпьютеры: разработка, применение» №6, 2009 с. 28-30
7. Feng Hao, Ross Anderson, and John Daugman. Crypto with Biometrics Effectively, IEEE TRANSACTIONS ON COMPUTERS, VOL. 55, NO. 9, SEPTEMBER 2006.
8. Juels A., Wattenberg M. A Fuzzy Commitment Scheme // Proc. ACM Conf. Computer and Communications Security, 1999, p. 28–36
9. Y. Dodis, L. Reyzin, A. Smith Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy, Data April 13, In EUROCRYPT, pages 523-540, 2004.

Статья поступила 20.01.2016, опубликована 20.05.2016 по положительной рецензии д.т.н. Иванова А.И.