Труды научно-технической конференции кластера пензенских предприятий, обеспечивающих
БЕЗОПАСНОСТЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ.
Том 10, Пенза-2016, с. 10-14 (http://пниэи.рф/activity/science/BIT/T10-p10.pdf)

Елфимов А. В. (г. Пенза), Безяев А.В. (г. Пенза)
УДК: 519.2; 519.6.

Сравнение гипотезы нормального распределения и гипотезы бета распределения расстояний Хэмминга для выходных кодов нейросетевых преобразователей

Общие положения тестирования средств биометрической аутентификации

В современном мире огромное значение имеют проблемы аутентификации в информационных системах. Из-за общей слабости и инерционности человеческой и машиной составляющей информационные системы аутентификации сталкиваются с существенными уязвимостями парольной защиты ресурсов. Пользователи не способны генерировать, запоминать и использовать сильные пароли, пароли могут быть перехвачены злоумышленником, владельцы информационных ресурсов вынуждены применять сложные схемы аутентификации, такие как двухфакторная аутентификация, с помощью системы мобильных приложений и т.п.

Для усиления защиты доступа к информационным ресурсам в настоящее время разрабатываются технологии создания и воспроизведения парольных фраз пользователя без непосредственного запрашивания его у пользователя. В США, Канаде, Евросоюзе разрабатываются так называемые «нечеткие экстракторы», которые используют рисунки отпечатка пальца [1], человеческий голос [2], изображения радужной оболочки глаза [3] для генерации кода криптографического ключа доступа [4]. Стойкость к атакам подбора этих систем неизвестна, в работах о «нечетких экстракторах» ошибки первого и второго даются на основе эмпирических данных. Отсутствуют так же документы, которые регламентируют тестирование на больших баз биометрических образов «Чужой».

В России используется стандартизованная технология нейросетевого преобразования биометрии в код пароля доступа. Нейросетевые преобразователи биометрических образов личности человека в код аутентификации должны выполняться в соответствии с требованиями пакета стандартов серии ГОСТ Р 52633.xx-20xx. В частности, национальный стандарт ГОСТ Р 52633.3 [5] предлагает осуществлять тестирование вероятностей ошибок второго рода с помощью предоставления обученному нейросетевому преобразователю до 32 примеров образов «Чужой», которые не участвовали ранее при обучении. Результат указан на рисунке 1.

Распределения расстояний Хэмминга между кодом «Свой» и кодами «Чужие» для нейросетевых преобразователей биометрия-код с 256 выходами хорошо описывается нормальным законом распределения значений. В связи с этим ГОСТ Р 52633.3 рекомендует вычислить математическое ожидание расстояний Хэмминга:

$$E(h) = \frac{1}{32}\sum_{i=1}^{32} h_i \qquad (1)$$

и их стандартное отклонение:

$$\sigma(h) = \sqrt{\frac{1}{31}\sum_{i=1}^{32} \left(E(h) - h_i\right)^2} \qquad (2).$$

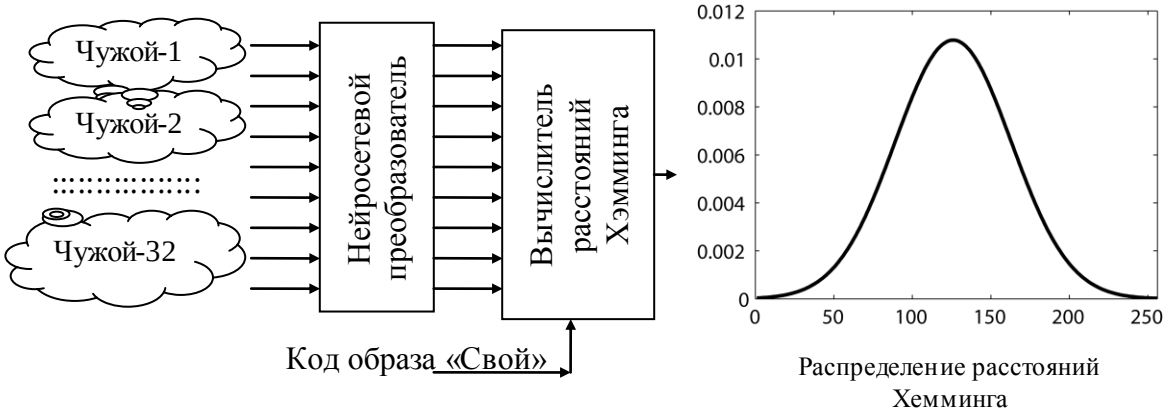


Рис. 1 Быстрое тестирование вероятности ошибок второго рода

Выборка из 32 примеров разных образов «Чужой» предоставляет достаточно данных для вполне точной оценки математического ожидания - E(h) и стандартного отклонения - σ(h).

В результате мы можем получить оценку вероятности ошибок второго рода для нейросетевых преобразователей биометрия-код, работающих совместно с криптографическими алгоритмами парольной аутентификации:

$$P_2 \approx \frac{1}{\sigma(h)\sqrt{2\pi}} \int_{-\infty}^{1} \exp\left\{\frac{-(E(h)-u)^2}{2\sigma(h)^2}\right\} \cdot du \qquad (3).$$

Если речь идет о применении «нечетких экстракторов», построенных на использовании избыточных кодов, исправляющих k ошибок, то интервал интегрирования следует увеличить:

$$P_2 \approx \frac{1}{\sigma(h)\sqrt{2\pi}} \int_{-\infty}^{k} \exp\left\{\frac{-(E(h)-u)^2}{2\sigma(h)^2}\right\} \cdot du \qquad (4).$$

Использование гипотезы бета распределения расстояний Хэмминга

Следует сказать, что оценки (3) и (4) всегда дают вероятность выше, чем она реально имеется. Это происходит из-за интегрирования в полубесконечных пределах, в то время как расстояние Хэмминга в случае нейросетевых преобразователей не может быть меньше нуля и больше 256. Если уменьшить интервал интегрирования:

$$P_2 \approx \frac{1}{\sigma(h)\sqrt{2\pi}} \int_{0}^{1} \exp\left\{\frac{-(E(h)-u)^2}{2\sigma(h)^2}\right\} \cdot du \qquad (5),$$

то мы получаем заниженную оценку вероятности ошибок второго рода.

Повысить точность оценок удается в случае, если отказаться от нормального распределения и применить гипотезу бета распределения [6, 7]:

$$p(x,\beta_1,\beta_2) = \frac{(\beta_1+\beta_2+1)!}{\beta_1!\cdot\beta_2!} \cdot x^{\beta_1} \cdot (1-x)^{\beta_2} \qquad (6),$$

где $x = \frac{h}{\max(h)}$ - нормированное расстояние Хэмминга; β_1, β_2 - первый и второй параметры настройки бета распределения.

Принципиальным преимуществом бета распределения является то, что интервал существования нормированных расстояний Хэмминга совпадает с интервалом этого распределения $0 < x < 1$.

Использование гипотезы бета распределения при оценке вероятности ошибок первого рода

Еще одним преимуществом использования гипотезы бета распределения является то, что она пригодна для оценки вероятности ошибок первого рода. Для этого необходимо изменить схему численного эксперимента. Вместо образов «Чужой» необходимо предъявлять примеры образа «Свой», ранее не использованные при обучении нейронной сети. Блок-схема численного эксперимента приведена на рисунке 2.

Рис. 2 Тестирование вероятности ошибок первого рода

На рисунке 2 отображена ситуация при которой 30 примеров тестовой выборки дали нулевое расстояние Хэмминга h=0. Два опыта дали расстояние Хэмминга h=1, один опыт дал расстояние h=3. Это приводит к значению математического ожидания E(h)=0.151 и дисперсии $\sigma^2(h) = 0.32$. Пользуясь тем, что бета распределение хорошо приближает асимметричные данные, мы можем перейти в нормированную систему расстояний Хэмминга и там найти параметры β_1, β_2. Зная их, не трудно найти вероятность ошибок первого рода для нейросетевого преобразователя с 256 выходами:

$$P_1 \approx \frac{(\beta_1 + \beta_2 + 1)!}{\beta_1! \cdot \beta_2!} \cdot \int_{1/256}^{1} x^{\beta_1} \cdot (1 - x)^{\beta_2} \cdot dx \qquad (7).$$

В случае тестирования «нечетких экстракторов» вероятность ошибок первого рода вычисляется по близкой формуле:

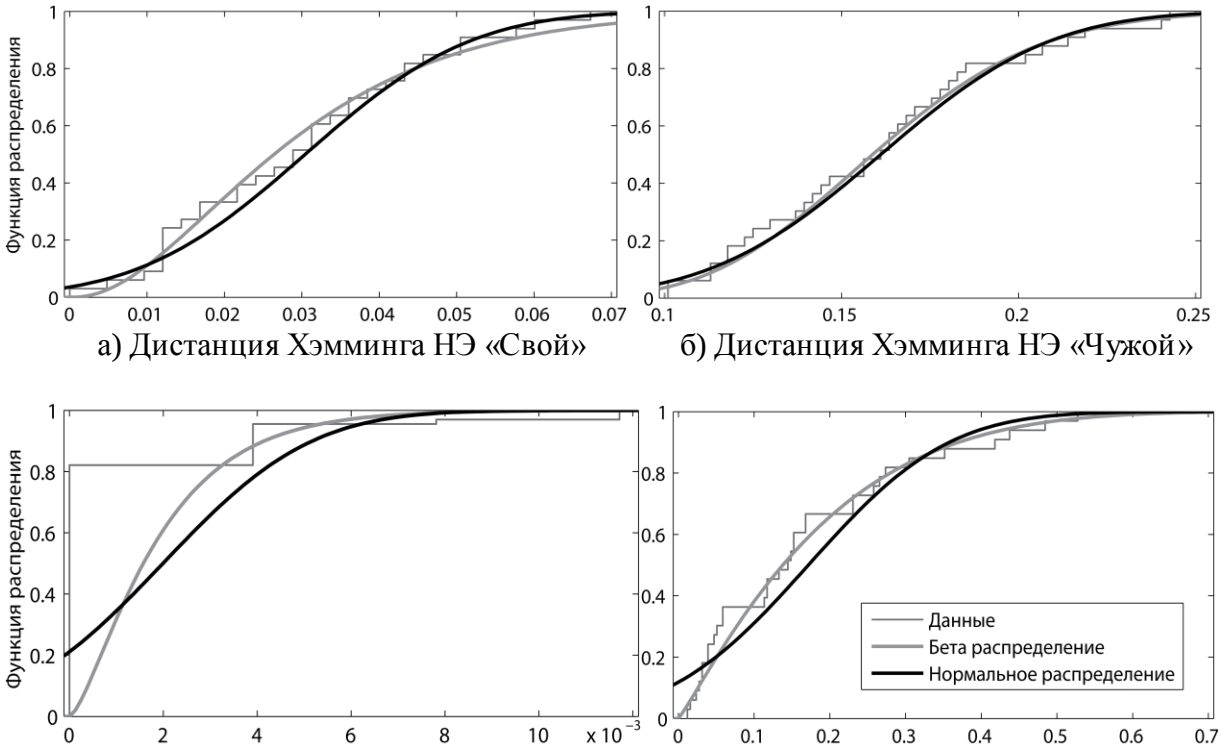
$$P_1 \approx \frac{(\beta_1 + \beta_2 + 1)!}{\beta_1! \cdot \beta_2!} \cdot \int_{k/N}^{1} x^{\beta_1} \cdot (1 - x)^{\beta_2} \cdot dx \qquad (8).$$

Получается, что гипотеза бета распределения является гораздо более универсальной по сравнению с гипотезой нормальности, она пригодна для описания выходных расстояний Хэмминга и искусственных нейронных сетей и «нечетких экстракторов». Гистограмма, изображенная на рисунке 2, существенно асимметрична, тем не менее, гипотеза бета распределения для нее работает. Результаты расчетов получаются точнее, чем оценка вероятности через вычисление отношения числа ошибок к общему числу проведения опытов.

Экспериментальное сравнение гипотез бета и нормального распределения

Рассмотрим данные расстояний Хемминга для нейросетевого преобразователя и «нечеткого экстрактора» для ввода рукописной парольной фразы в программе моделирования «БиоНейроАвтограф»[9]. Для режима «нечетких экстракторов» получаем на тестовой выборке расстояния Хемминга для суммы фактов вхождений анализируемых 416 параметров p рукописного воспроизведения в участок $E(p) \pm 3\sigma(p)$ распределения

каждого параметра *p* для тестовых образов «Свой». Для режима нейросетевого преобразователя - количество полученных на выходе нейросетевого преобразователя единиц для вводимых рукописных образов. Проведем по 32 опыта для получения расстояний Хэмминга «Свой» и «Чужой» для «нечеткого экстрактора» и «нейросетевого преобразователя» на примерах рукописного ввода тестового слова «Пенза», как эталона «Свой». Построим их функции распределения и наложим на них функции нормального распределения и бета распределения, описывающие распределения опытных данных. Для тестирования «Чужих» нейросетевого преобразователя использовались максимально похожие рукописные начертания «Чужих», отличающиеся от «Пенза» последней буквой, для как можно большего смещения расстояний Хэмминга «Чужих» к нулю координат. Функции распределения вероятностей (серая тонкая ступенчатая линия) с наложенным нормальным распределением (черная непрерывная линия) с наложенным бета-распределением (серая непрерывная линия) отображены на рисунке 4:



а) Дистанция Хэмминга НЭ «Свой» б) Дистанция Хэмминга НЭ «Чужой»

г) Дистанция Хэмминга НС «Свой» г) Дистанция Хэмминга НС «Чужой»

Рис. 4 Тестирование гипотез нормального и бета распределения.

Как видно из рисунка выше, гипотеза бета–распределения дает более близкие к экспериментальным данным, по сравнению с нормальной, функции распределения как в случае тестирования «нечетких экстракторов» (на рисунке 4 части а) и б)), так для нейросетевых преобразователей биометрия–код (на рисунке 4 части в) и г)).

Заключение

С использованием для расстояний Хэмминга гипотезы бета распределения, для получения оценок вероятностей ошибок первого и второго рода с приемлемой точностью становится достаточно небольших объемов тестовых выборок, из примерно 32 опытов. Это возможно из-за того, что бета распределение более точно описывает распределение расстояний Хэмминга по сравнению с гипотезой нормального распределения. Бета распределение является более универсальным и позволяет оценивать не только ошибку первого рода, но и ошибку второго рода, а также позволяет оценить вероятности распределения расстояний Хэмминга для подходов «нечетких экстракторов».

Критически важным элементом описанных выше процедур является их квантово-континуальный характер. Несмотря на то, что распределения расстояний Хэмминга дискретны, мы при вычислении интеграла вероятностей (3), (4), (5), (7), (8) рассматриваем их как непрерывные. Уменьшение тестовой выборки до 32 опытов как раз связано с континуально-квантовыми переходами. Для многомерных континуумов их энтропия много выше, чем энтропия дискретных выходных кодов «нечетких экстракторов» и нейросетевых преобразователей. Возможность снизить объемы тестовых выборок и, вследствие этого многократно ускорить вычисления как раз достигается за счет многомерных континуально-квантовых переходов при вычислениях. Чем выше размерность решаемой биометрической задачи, тем больше будет выигрыш от таких преобразований.

В связи с тем, что бета аппроксимация распределений расстояний Хэмминга работает лучше, чем аппроксимация нормальным законом распределения значений, необходимо доработать стандарт ГОСТ Р 52633.3 и ввести в действие его новую редакцию взамен действующей.

ЛИТЕРАТУРА:

1. Ramírez-Ruiz J. Keys Generation Using FingerCodes. / J. Ramírez-Ruiz, C. Pfeiffer, J. Nolazco-Flores //Advances in Artificial Intelligence - IBERAMIA-SBIA. – 2006 (LNCS 4140). - P. 178-187.
2. Monrose F. Cryptographic key generation from voice. / F. Monrose, M. Reiter, Q. Li, S. Wetzel // In Proc. IEEE Symp. on Security and Privacy. pp. 202-213,– 2001.
3. Hao F., Anderson R., Daugman J. Crypto with Biometrics Effectively //IEEE TRANSACTIONS ON COMPUTERS, VOL. 55, NO. 9, SEPTEMBER, Page(s):1073 – 1074, 2006.
4. Dodis Y., Reyzin L., Smith A. Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy // Proc. EUROCRYPT, April 13, pages 523-540, 2004.
5. ГОСТ Р 52633.3-2011 «Защита информации. Техника защиты информации. Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора».
6. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. М.: ФИЗМАТЛИТ, 2006 г., 816 с.
7. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. — М.: Наука, 1985. — 640 с.
8. Вентцель Е.С. Теория вероятностей. — М.: Наука, 1969. — 576 с.
9. 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.