Труды научно-технической конференции кластера пензенских предприятий,
обеспечивающих **БЕЗОПАСНОСТЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ.**
Том 10, Пенза-2016, с. 78-83 (http://пниэи.рф/activity/science/BIT/T10-p78.pdf)

УДК: 519.2, 612.087, 621.319.7

Газин А.И. (г. Липецк), Ахметов Б.Б. (Казахстан, г. Туркестан),
Сериков А.В., Серикова Ю.И. (г. Пенза).

Оценка соотношения методической и случайной составляющих погрешности вычисления коэффициентов корреляции для малых выборок биометрических данных

В настоящее время в России [1] и в Казахстане [2], активно используются искусственные нейронные сети для преобразования биометрических параметров в код доступа. Обучение искусственных нейронных сетей выполняется по алгоритму ГОСТ Р 52633.5 [3], который имеет линейную вычислительную сложность. Естественно, что для ряда приложений качество принимаемых нейронными сетями решений недостаточно. В связи с этим, стандартизованные нейросетевые преобразователи биометрия-код [4] могут быть усилены сетями функционалов Байеса [5, 6], для обучения которых необходимо знать математические ожидания - $E(\nu)$ биометрических параметров, их стандартные отклонения - $\sigma(\nu)$ и значения парных коэффициентов корреляции – $r(\nu_i,\nu_j)$.

К сожалению, ошибка вычислений накапливается $|\Delta E(\nu)| < |\Delta\sigma(\nu)| < |\Delta r(\nu_i,\nu_j)|$. Необходимо выполнять некоторые процедуры регуляризации вычислений, которые уменьшают результирующую погрешность - $|\Delta r(\nu_i,\nu_j)|$ на малых выборках биометрических данных.

Очевидно, что регуляризация вычислений во многом должна напоминать известные методы борьбы с погрешностями. Так как методы компенсации различных составляющих погрешности разные, то необходимо уметь вычленять из результирующей погрешности вычислений ее случайную и методическую составляющие. Положение усугубляется тем, что для каждого размера выборки, погрешности оказываются разными. Примеры распределений значений вычисленных коэффициентов корреляции для выборок разного размера даных на рисунке 1 и рисунке 2.

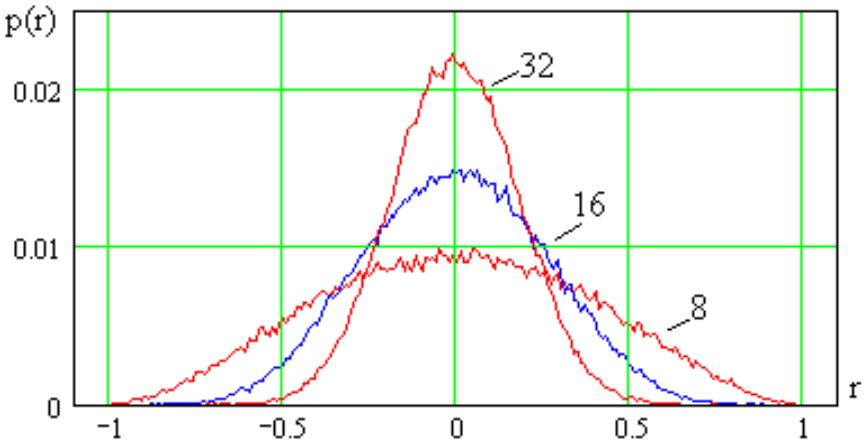

Рис. 1. Распределения значений вычисленных коэффициентов корреляции на выборках из 8, 16, 32 примеров для независимых данных

Из рисунка 1 видно, что для выборки из 8 примеров значения независимых данных попадают в интервал от -1 до +1. Уже при выборке из 16 примеров интервал вычисленных значений падает, попадая в интервал от -0.75 до +0.75. Следующее удвоение размеров выборки до 32 примеров сужает интервал от -0.6 до +0.6. Наблюдается замедление скорости сжатия интервала возможных состояний вычисленного корреляционного функционала, по мере увеличения размеров тестовой выборки. Для малых выборок от 2 до 32 примеров, каждое из распределений имеет свой закон распределения значений. Происходит монотонная нормализация этих законов. Для выборок, состоящих из 32 и более примеров, закон распределения значений можно считать нормальным. Исходя из этого, для больших выборок независимых данных интервал значений будет сжиматься пропорционально $\sqrt{n}$, где n – число примеров в выборке.

Для независимых данных математическое ожидание распределений рисунка 1 нулевое, следовательно полностью отсутствует методическая погрешность вычисления классических коэффициентов корреляции.

К сожалению, эта ситуация меняется как только мы будем вычислять коэффициент корреляции зависимых данных. На рисунке 2 приведены распределения значений для разных выборок сильно зависимых данных с коэффициентом корреляции r = 0.99.

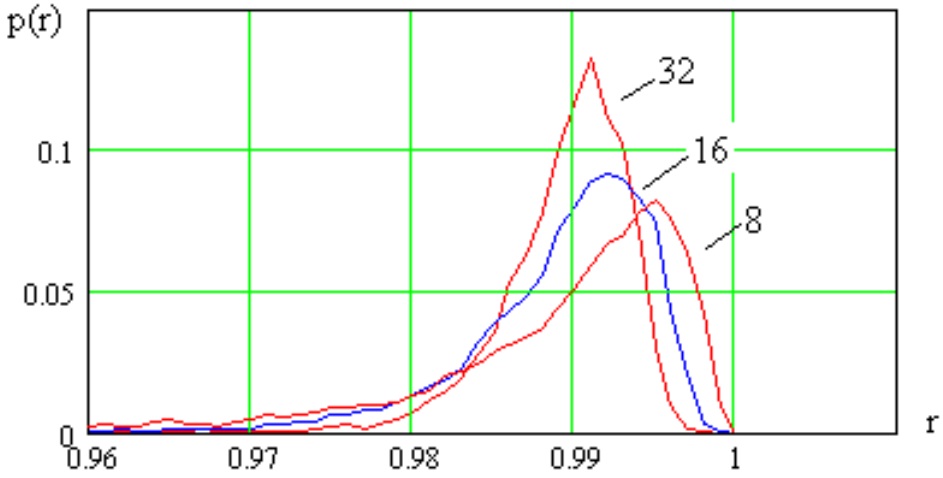

Рис. 2. . Распределения значений вычисленных коэффициентов корреляции на выборках из 8, 16, 32 примеров для сильно зависимых данных

Из рисунка 2 видно, что эти распределения асимметричны, их математическое ожидание не совпадает со значением коэффициента корреляции, что приводит к методической погрешности вычислений. Оказывается, что только в трех точках при вычислениях коэффициентов корреляции, полностью отсутствует методическая погрешность:

$$\begin{cases} E(r) = r \quad если \quad r = -1, \\ E(r) = r \quad если \quad r = \pm 0, \\ E(r) = r \quad если \quad r = +1 \end{cases} \qquad (1).$$

Во всех иных случаях совпадение отсутствует:

$$E(r) \neq r \qquad (2),$$

что эквивалентно наличию методической погрешности вычислений.

Естественно, что небаланс (2) для разных выборок оказывается разным. Функции изменения значений методической погрешности вычисления коэффициента корреляции, полученные численным моделированием, даны на рисунке 2.

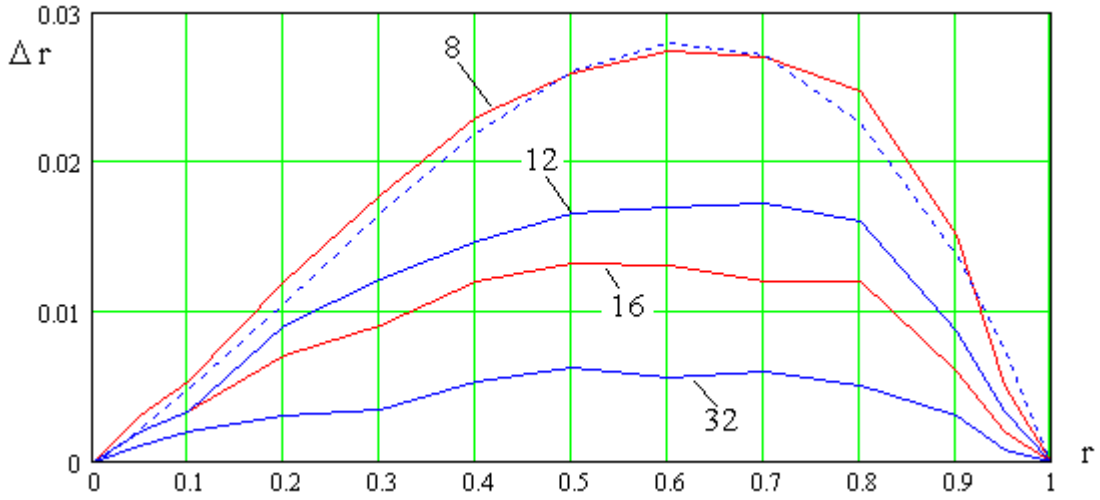

Рис. 2. Значения методической ошибки - Δr, полученные численным моделированием (сплошные линии) для выборок из 8, 12, 16, 32 примеров.

Результаты численного моделирования не являются гладкими функциями только из-за ограниченного числа использованных состояний программного генератора псевдослучайных чисел (10 тысяч выборок) в каждой из 13 точек моделирования коэффициентов корреляции. Увеличение числа точек моделирования значений коэффициентов корреляции и увеличение числа учитываемых состояний программного генератора псевдослучайных чисел приводит к сглаживанию наблюдаемых кривых.

При численном моделировании коэффициенты корреляции воспроизводились путем умножения вектора псевдослучайных чисел - $\bar{x}$ на симметричную связывающую симметричную матрицу [7, 8]:

$$\begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix} \cdot \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix} \qquad (2).$$

Данные, получаемые от псевдослучайных программных генераторов, всегда независимы, их коэффициенты корреляции имеют нулевое математическое ожидание $E(r(\bar{x}_1, \bar{x}_2)) = 0$. После умножения данных на связывающую матрицу с не нулевыми коэффициентами вне диагонали данные оказываются зависимыми $E(r(\bar{y}_1, \bar{y}_2)) \neq 0$. Размер выборки $\bar{x}_1 = \{x_{1,1}, x_{1,2}, ..., x_{1,n}\}$ может быть любым.

Так как мы можем заранее вычислить значения методической погрешности (рис. 2), она является устранимой. Могут быть использованы соответствующие поправочные таблицы или их аналитическое приближение. В частности, может использоваться аппроксимация данных полиномом третьего порядка (на рисунке 2 след приближающего полинома дан пунктиром):

$$\Delta r(r) = 0.052 \cdot \left[0.5 - \left[1.42\left(0.5 - r\right)\right]^2\right] - 0.125r \cdot \left(0.5 - r\right) \cdot \left(1 - r\right) \qquad (3).$$

Приближение (3) построено для выборки из 8 примеров, если выборка оказывается больше, то необходимо уменьшить масштаб соотношения (3):

$$\Delta r(r,n) = \frac{M(n)}{0.038} \cdot \Delta r(r) \qquad (4).$$

Масштабирующая приближающий полином функция M(n), полученная численным экспериментом, дана сплошной линией на рисунке 3.

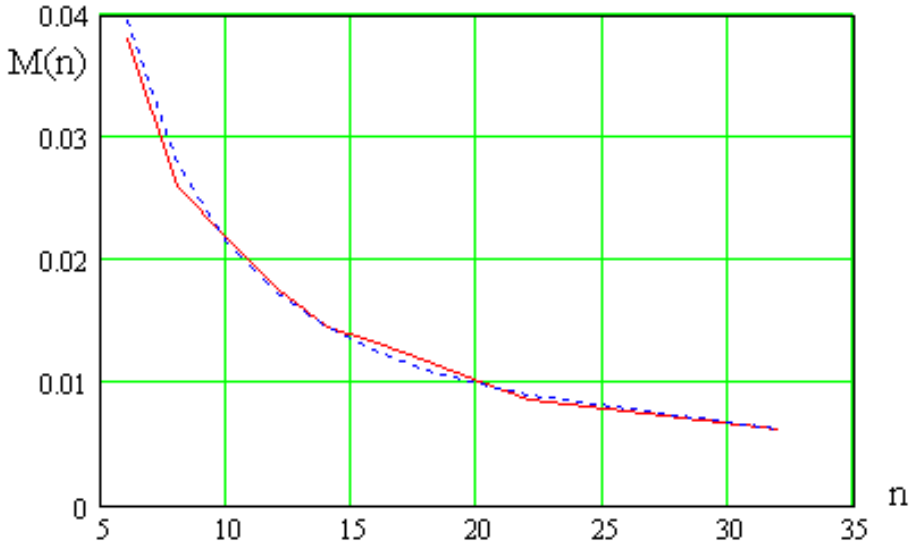

Рис. 3. Изменение масштаба приближающего полинома с ростом объема выборки

Достаточно хорошее приближение масштабирующей функции дает гипербола степени 1.29, след которой на рисунке 3 дан пунктиром:

$$M(n) := \frac{1 \cdot 0.38}{n^{1.29}} + 0.0019 \qquad (5).$$

В связи с тем, что методическая погрешность вычисления коэффициентов корреляции может быть скомпенсирована, возникает вопрос о том, когда такая компенсация целесообразна. Очевидно, что компенсировать методическую погрешность вычислений имеет смысл только, если она больше или сопоставима со случайной погрешностью. В свою очередь, интервал случайной погрешности можно оценивать по инженерному правилу, как три стандартных отклонения в правую и левую сторону от математического ожидания.

Используя средства имитационного моделирования можно убедиться в том, что для каждого размера выборки получается своя зависимость $\sigma(r,n)$. Примеры таких функций даны на рисунке 4.

Из рисунка 4 видно, что случайная ошибка максимальная для независимых данных. Повышение уровня корреляционной связанности данных приводит к монотонному снижению значения стандартного отклонения. В точках $r = \pm 1$ стандартное отклонение становится нулевым. Рост числа примеров в выборке приводит к снижению случайной составляющей ошибки вычисления корреляционных функционалов.

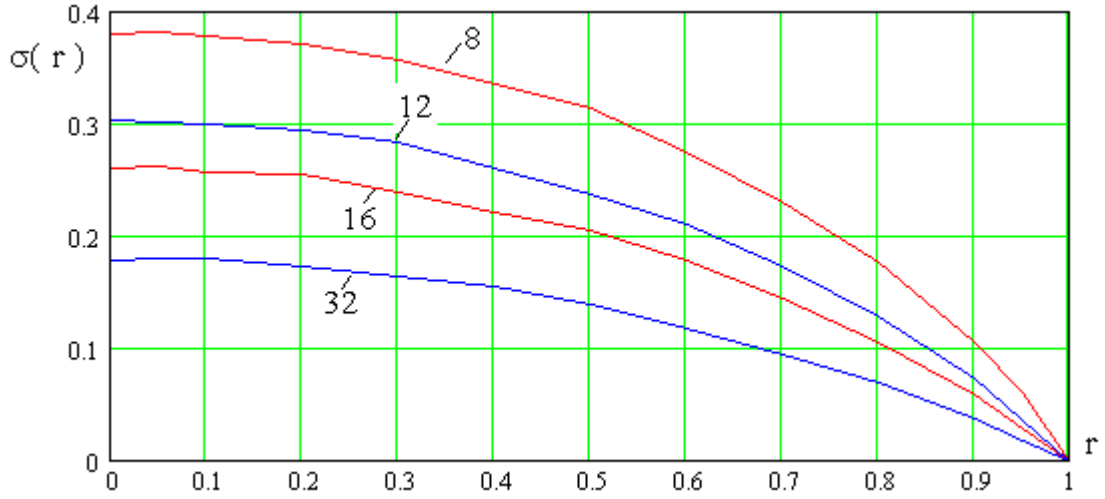

Рис. 4. Зависимости стандартных отклонений для выборок разного объема от коррелированности данных в выборках

Если сравнивать данные рисунка 2 и рисунка 4, то оказывается, что методическая ошибка от 3 до 25 раз меньше случайной составляющей погрешности. То есть корректировка методической составляющей погрешность должна снизить общую погрешность от 3% до 33%. Эффективность корректировки методической погрешности различна при разных выборках и разном уровне корреляции, однако, эффект понижения результирующей ошибки всегда присутствует. Для реализации этого эффекта требует только внести изменения в программную реализацию вычислений.

В свою очередь, отсутствие (компенсация) методической погрешности упрощает процедуры борьбы со случайной составляющей погрешности. Для этой цели необходимо использовать несколько разных линейно независимых способов вычисления коэффициентов корреляции [9]. Если при этом данные будут не полностью коррелированны, то всегда есть возможность создать линейное обобщение, усредняющее (подавляющее) случайные составляющие погрешности.

ЛИТЕРАТУРА:

1. 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.
2. Ахметов Б.С., Иванов А.И., Фунтиков В.А., Безяев А.В., Малыгина Е.А. Технология использования больших нейронных сетей для преобразования нечетких биометрических данных в код ключа доступа. Монография, Казахстан, г. Алматы, ТОО «Издательство LEM», 2014 г. -144 с., находится в открытом доступе (http://portal.kaznu.kz/files/publicate/2014-06-27-11940.pdf)
3. ГОСТ Р 52633.5-2011 «Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия-код доступа».
4. ГОСТ Р 52633.0-2006 «Защита информации. Техника защиты информации. Требования к средствам высоконадежной биометрической аутентификации».

5. Иванов А.И., Ложников П.С., Качайкин Е.И. Идентификация подлинности рукописных автографов сетями Байеса-Хэмминга и сетями квадратичных форм. «Вопросы защиты информации» №2 2015 г., с. 28-34.
6. Иванов А.И., Ложников П.С., Качайкин Е.И., Сулавко А.Е. Биометрическая идентификация рукописных образов с использованием корреляционного аналога правила Байеса. «Вопросы защиты информации» №3 2015 г., с. 48-54.
7. Ахметов Б.Б., Иванов А.И. Многомерные статистики существенно зависимых биометрических данных, порождаемые нейросетевыми эмуляторами квадратичных форм: Монография. Казахстан – Алматы. Из-во LEM, 2016. 86 с.
8. 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.
9. Волчихин В.И., Иванов А.И., Ахметов Б.Б., Серикова Ю.И. "Фрактально-корреляционный функционал, используемый при поиске пар слабо зависимых биометрических данных в малых выборках" //«Вестник высших учебных заведений. Поволжский регион. Технические науки» №4, 2016 г., с. 25 – 31.

Статья поступила 11.12.2016, опубликована 12.12.2016
по положительной рецензии д.т.н. Малыгиным А.Ю.