Труды научно-технической конференции кластера пензенских предприятий,
обеспечивающих **БЕЗОПАСНОСТЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ.**
Том 10, Пенза-2016, с. 27-29 (http://пниэи.рф/activity/science/BIT/T10-p27.pdf)

УДК: 519.2, 612.087, 621.319.7

Серикова Ю.И., Банных А.Г.

**Бинормальная регуляризация оценки числа слабо коррелированных данных биометрического образа «Свой»**

Базовый национальный стандарт ГОСТ Р 52633.0-2006 [1] ориентирован на использование искусственных нейронных сетей с линейными функционалами обогащения данных из-за того, что известен устойчивый алгоритм обучения ГОСТ Р 52633.5-2011 [2], обладающий линейной вычислительной сложностью [3].

Предположительно эволюция алгоритмов обучений преобразователей биометрия-код будет идти по пути использования квадратичных форм высокой размерности и функционалов Байеса высокой размерности [4, 5]. И в том, и в другом случае, эволюция идет по пути учета коэффициентов корреляции при обучении преобразователей биометрия-код.

Следует отметить, что реальные коэффициенты парной корреляции биометрических данных для рукописных образов [6] имеют распределение значений с явно выраженной пологой вершиной, как это показано на рисунке 1.

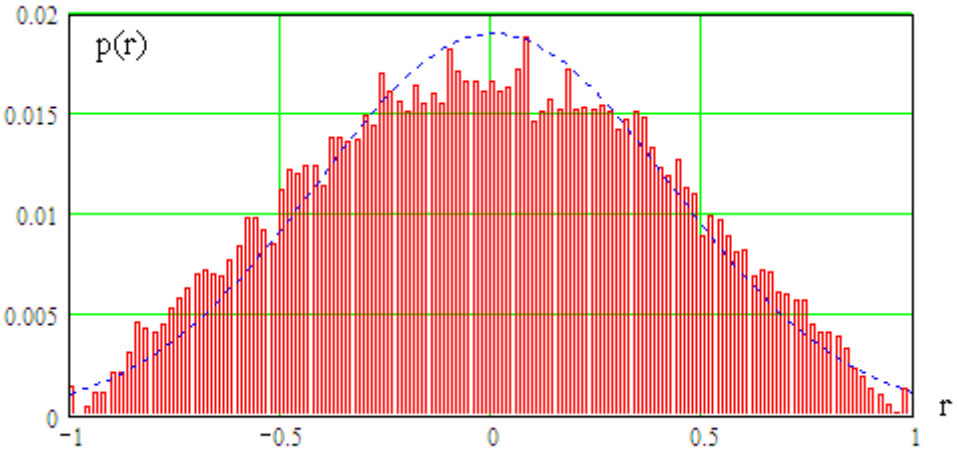


Рис. 1. Распределение значений коэффициентов корреляции 416 параметров рукописного образа «Пенза»

Практика показывает, что биометрические данные рукописных образов имеют симметричную плотность распределения значений. В итоге мы имеем ограниченное распределение значений коэффициентов корреляции в интервале от -1 до +1, обладающее тупой вершиной и симметрией относительно точки r =0.0.

Тупая вершина распределения свидетельствует о том, что достаточно много биометрических данных слабо коррелированны. Это открывает возможность применения сети квадратичных форм. Если пользоваться линейной алгеброй, то каждая из таких квадратичных форм будет описываться следующим соотношением:

$$y = (E(\overline{v}) - \overline{v})^T \cdot [R]^{-1} \cdot (E(\overline{v}) - \overline{v}) \qquad (1),$$

где $\overline{v}$ - вектор контролируемых нормированных биометрических параметров с единичным стандартным отклонением каждого биометрического параметра $\sigma(v_i) = 1$, $[R]^{-1}$ - обратная корреляционная матрица.

Очевидно, что проблема использования сетей квадратичных форм (1) технически ограничивается нашими возможностями по обращению корреляционных матриц. В общем случае эта задача является плохо обусловленной и нуждается в регуляризации [7, 8]. К сожалению, регуляризовать по Тихонову обращение корреляционных матриц 16 и более высоких порядков технически невозможно. Это связано с тем, что ошибка вычисления малых значений коэффициентов корреляции на малых выборках слишком высока. На рисунке 2 даны распределения коэффициентов корреляции для малых тестовых выборок.

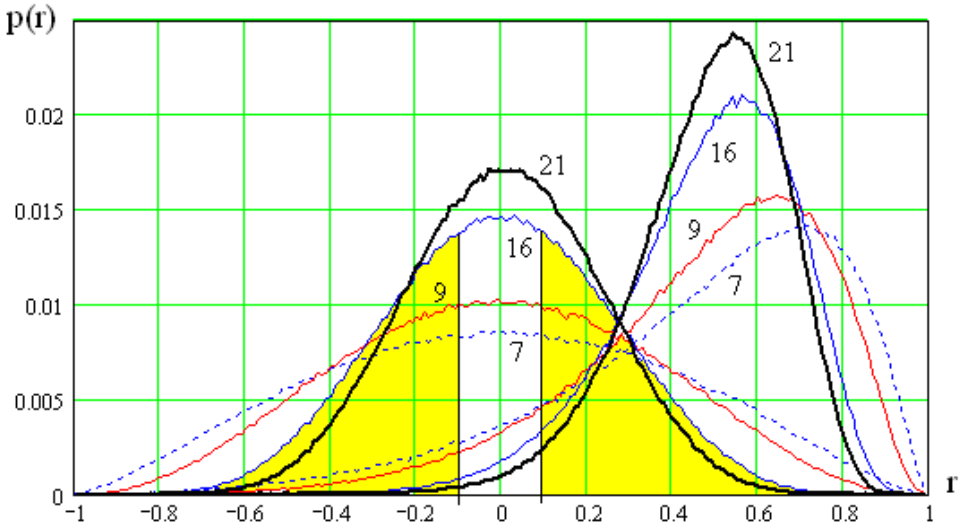


Рис. 2. Распределения значений коэффициентов корреляции, для выборок из 7, 9, 16, 21 примеров при двух заданных значениях коэффициентов корреляции r = 0 и r = 0.5

Из рисунка 2 видно, что самая большая погрешность возникает при определении коэффициентов корреляции независимых данных. Так для выборок из 16 опытов вычисленный нулевой коэффициент корреляции попадает в интервал от -0.1 до +0.1 с вероятностью только 0.2 (вероятность ошибок отмечена заливкой). Техническая трудность задачи регуляризации обусловлена тем, что ошибка вычисления коэффициентов корреляции в обращаемой матрице может составлять до 60%. При этом коэффициент обусловленности обращаемой матрицы может составлять несколько тысяч (коэффициент обусловленности $cond[R]$ в первом приближении можно рассматривать как коэффициент усиления шумов или ошибок входных данных).

Заметим, что проблемы с регуляризацией обращения корреляционной матрицы возникают только тогда, когда мы выбираем входные данные квадратичного функционала произвольно. Если же мы будем выбирать для обработки одним функционалом только слабо зависимые данные, то их коэффициентами корреляции можно пренебречь. То есть, выражение (1) вырождается до гораздо более простого соотношения:

$$y = (E(\overline{v}) - \overline{v})^T \cdot (E(\overline{v}) - \overline{v}) \qquad (2).$$

Если мы рассчитаем полную корреляционную матрицу 416x416, то некоторая часть коэффициентов парной корреляции попадет в интервал от -0.1 од +0.1. Определить число таких коэффициентов корреляции удается, если аппроксимировать распределение данных смесью двух нормальных законов распределения значений, сдвинутых относительно центра на некоторую величину - a со стандартным отклонением этих распределений уменьшенного не величину – a. На рисунке 3 пунктиром приведена плотность распределения бинормального распределения для параметра регуляризации a =0.3.

$$a := 0.3$$

$$y2_i := 0.5 \cdot dnorm\left[int_i, mean(r) - a, (1 - a) \cdot stdev(r)\right] + 0.5 \cdot dnorm\left[int_i, mean(r) + a, (1 - a) \cdot stdev(r)\right]$$

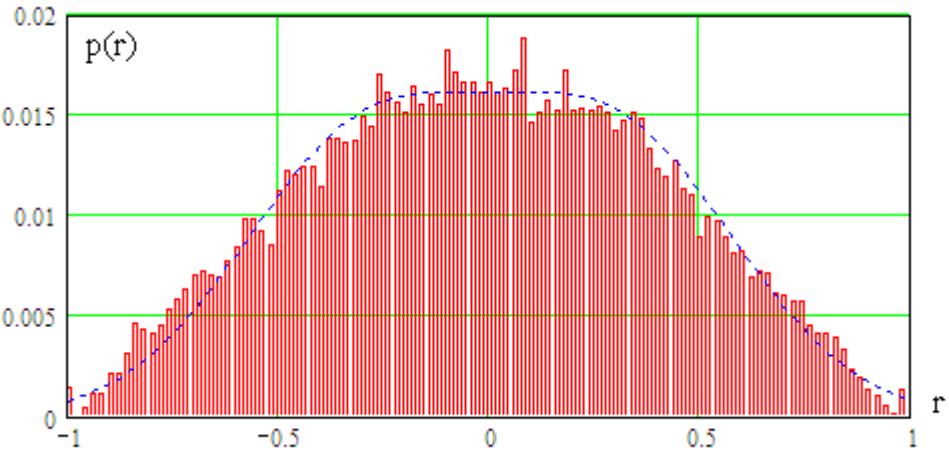


Рис. 3. Распределение значений коэффициентов корреляции 416 параметров рукописного образа «Пенза» при использовании бинормальной симметричной аппроксимации

Интегрирование плотности распределения значений рисунка 3 в интервале от - 0.1 од +0.1 дает вероятность 0.224. Так как матрица симметрична, общее число коэффициентов корреляции составит $(416^2\text{-}416)/2=86320$. Из них порядка 19336 пар параметров можно рассматривать как низкокоррелированные данные, пригодные для подстановки в квадратичные формы (2).

Каждая из 416 строк будет содержать как минимум 416x0.224=96 слабо коррелированных параметров. Для формирования первого 16-мерного квадратичного функционала достаточно выявить слабокоррелированные параметры по отношению к первому параметру (анализ первой строки корреляционной матрицы). Далее следует построить полную корреляционную матрицу слабокоррелированных данных размерностью 96x96 и осуществить удаление данных, имеющих самые большие по модулю коэффициенты корреляции.

Очевидно, что подобный сортировочный алгоритм имеет квадратичную вычислительную сложность настройки (обучения или формирования) функционалов вида (2). Предварительные расчеты показали, что обычная регуляризация [7, 8] позволяет обучать квадратичные функционалы до 8 порядка. Далее возникают технические проблемы плохой обусловленности обращения корреляционных матриц.

Описанный выше алгоритм регуляризации построен на выборе только слабокоррелированных данных и проверке минимальных значений взаимной коррелированности данных. Этот тип процедур регуляризации позволяет без технических трудностей настраивать квадратичные функционалы 16-го и более высоких порядков.

ЛИТЕРАТУРА:


1. ГОСТ Р 52633.0-2006 «Защита информации. Техника защиты информации. Требования к средствам высоконадежной биометрической аутентификации».
2. ГОСТ Р 52633.5-2011 «Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия-код доступа».
3. 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.
4. 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.
5. Качайкин Е.И., Иванов А.И. Идентификация авторства рукописных образов с использованием нейросетевого эмулятора квадратичных форм высокой размерности. «Вопросы кибербезопасности» № 4(12) 2015 с. 42-47.
6. Иванов А.И., Захаров О.С. Среда моделирования «БиоНейроАвтограф». Программный продукт создан лабораторией биометрических и нейросетевых технологий, размещен с 2009 г. на сайте АО «ПНИЭИ» http://пниэи.рф/activity/science/noc.htm для свободного использования университетами России, Белоруссии, Казахстана.
7. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. М.: Наука, 1979, 248 с.
8. Райс Дж. Матричные вычисления и математическое обеспечение. – М.: Мир, 1984 г. 412 с.



Статья поступила 10.09.2016, опубликована 23.09.2016 по положительной рецензии к.т.н. Безяева А.В.