Всероссийская научно-техническая конференция, посвященная 100-летию со дня рождения одного из основоположников советской вычислительной техники Б.И. Рамеева, на тему: «Информационно-управляющие и телекоммуникационные системы специального назначения». Секция «Биометрическая поддержка криптовалют, блокчейн реестров, облачных сервисов», проводится в конференц-зале Технопарка высоких технологий «Рамеев», доклад 16 мая 2018 года с 13^{30} до 13^{45} (ул. Центральная, д. 1), г. Пенза.

Сериков А.В., Качалин С.В.

Корреляционная молекула с эллиптическими квантователями для вычислений на малых обучающих выборках

Аннотация.

Актуальность и цели. Целью работы является повышение корректности вычисления коэффициентов корреляции.

Материалы и методы. Классические процедуры вычисления коэффициентов корреляции на малых выборках дают большую ошибку. Для повышения точности вычислений необходимо иметь возможность определять коэффициенты корреляции разными способами. Предложено для вычислений использовать корреляционный нейрон, имеющий два эллиптических квантователя.

Результаты. Показано, что новый метод вычисления корреляции дает погрешность слабо коррелированную с погрешностью вычисления корреляции классическим методом. То есть новые данные могут быть использованы для корректировки погрешности классических вычислений.

Выводы. Синтезированный корреляционный нейрон следует рассматривать как корреляционную молекулу, имеющую свой уникальный выходной дискретный спектр состояний. Таким образом, на текущей момент созданы два варианта корреляционных молекул, каждая из которых имеет свой уникальный спектр выходных состояний. Вычисления, имеющие одинаковый математический смысл могут соответствовать нескольким реализациям математических молекул.

Ключевые слова: вычисление коэффициентов корреляции на малых выборках, регуляризация вычислений, многообразие корреляционных молекул.

Проблема учета коэффициентов корреляции при обучении нейросетевых преобразователей биометрия-код

В настоящее время Россия создает собственную цифровую экономику. Одной из проблем цифровой экономики является создание эффективных механизмов учета цифровых затрат и цифровых ресурсов. Подобные механизмы учета могут быть реализованы, опираясь на возможность массового применения криптографии. При этом возникает серьезное технологическое ограничение, обусловленное тем, что обычный человек не может запомнить множество своих личных криптографических ключей.

Для того, что бы снять это ограничение в США и странах НАТО создаются так называемые «нечеткие экстракторы» [1, 2, 3], преобразующие биометрию человека в код пароля доступа. Проблемой нечетких экстракторов является то, что они применяют классические коды с высокой избыточностью для обнаружения и исправления ошибок [4]. При 30-ти кратной избыточность длина выходного кода «нечеткого экстрактора» становится в 30 раз меньше, чем число контролируемых биометрических параметров человека. По этой причине для рукописных образов, голосовых образов, рисунков отпечатка пальцев, образов лица человека «нечеткие экстракторы» дают коды длинной до 20 бит. Это не позволяет объединять «нечеткие экстракторы» с сильной криптографией длинных ключей.

Россия идет иным путем, создавая и стандартизируя нейросетевые преобразователи биометрия-код [5, 6, 7]. Один нейрон в таких преобразователях создает один бит ключа, то

есть их выходной код может иметь любую длину [8, 9] из-за того, что в преобразователях можно использовать много искусственных нейронов. Обучение нейросетевых преобразователей биометрия-код должно быть полностью автоматически и иметь низкую вычислительную сложность. В частности стандартизованный в России алгоритм обучения [7] имеет линейную вычислительную сложность за счет того, что он не является итерационным и вычисляет весовые коэффициенты нейронов как простую функцию младших статистических моментов двух переменных:

$$|\mu_i| = f(E(\nu_i), E(\xi_i), \sigma(\nu_i), \sigma(\xi_i)) \qquad (1),$$

где ν_i - биометрический параметр образа «Свой», ξ_i - биометрический параметр образа «Чужой», $E(.)$ - оператор вычисления математического ожидания, $\sigma(.)$ - оператор вычисления стандартного отклонения.

Ожидается, что следующее поколение стандартизованных нейросетевых преобразователей биометрия-код будет автоматически обучаться алгоритмом, имеющим квадратичную вычислительную сложность за счет дополнительного учета одинаковых корреляционных связей [10, 11, 12] биометрических данных одного нейрона:

$$|\mu_i| = f(E(\nu_i), E(\xi_i), \sigma(\nu_i), \sigma(\xi_i), r_i(\nu_i, \nu_j)) \qquad (2),$$

где $r_i(\nu_i, \nu_j) = r_i(\nu_i, \nu_{j+1}) = r_i(\nu_i, \nu_{j+2}) = \ldots$ одинаково коррелированные биометрические данные одного нейрона, полученные специальной процедурой симметризации корреляционных связей.

Проблема определения корреляционных связей при использовании малых тестовых выборок

К сожалению, классическая процедура вычисления коэффициентов корреляции по формуле Пирсона неустойчива:

$$r(\nu_i, \nu_j) = \frac{1}{n}\sum_{k=1}^{n} \frac{(E(\nu_i) - \nu_{i,k}) \cdot (E(\nu_j) - \nu_{j,k})}{\sigma(\nu_i) \cdot \sigma(\nu_j)} \qquad (3).$$

На малых выборках ошибки вычисления коэффициентов корреляции оказываются велики. Эта ситуация отображена на рисунке 1.

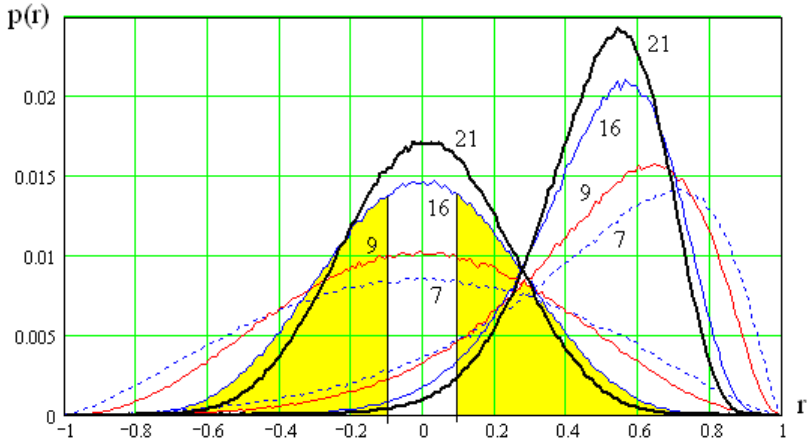

Рис. 1. Распределения значений коэффициентов корреляции при разных тестовых выборках в 7, 9, 16, 21 примеров

Из рисунка 1 видно, что самая большая погрешность возникает при определении коэффициентов корреляции независимых данных. Так для выборок из 16 независимых опытов вычисленный коэффициент корреляции попадает в интервал от -0.1 до +0.1 с вероятностью только 0.2. Становится актуальной задача повышения точности вычисления

коэффициентов корреляции на малых выборках. Причина плохой обусловленности вычисления коэффициентов корреляции состоит в том, что накапливаются четыре ошибки вычисления статистических моментов, входящих в формулу Пирсона (3):

$$\Delta r(\nu_i, \nu_j) = f(\Delta E(\nu_i), \Delta E(\nu_j), \Delta\sigma(\nu_i), \Delta\sigma(\nu_j)) \qquad (4).$$

Функция (4) обычно монотонна. В связи с этим ошибку вычисления коэффициентов корреляции удается снизить, уменьшая значения ошибок статистических моментов двух контролируемых биометрических параметров [13, 14]. Еще одним способом повышения точности вычислений является синтез новых процедур оценки [15, 16] коэффициентов корреляции и усреднения их данных с данными классической формулы Пирсона (3).

Синтез новой процедуры оценки коэффициента корреляционной связи

При обработке данных реальных стрельб военными [17] практиковалось описание нормальных распределений эллипсами, как это показано на рисунке 2.

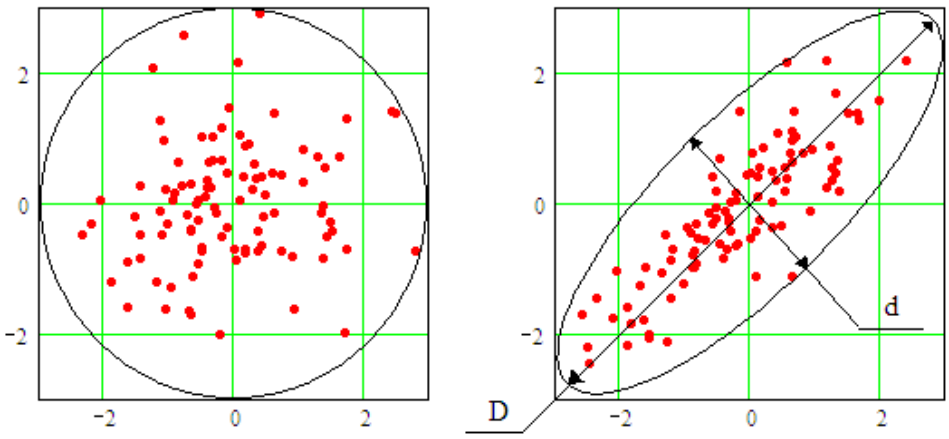

Рис. 2 Приближение двухмерного нормального распределения данных эллипсами

На рисунке 2 приведены 2 приближения нормальных распределений, полученные ручным способом. После подбора параметров эллипсов в ручном режиме оценивают соотношение их большого и малого диаметров:

$$r \approx 1 - \frac{d}{D} \qquad (5).$$

Расчет коэффициентов корреляции по формуле (5) дает значение корреляции близкое к нулю для распределения в левой части рисунка 2. Для распределения данных в правой части рисунка корреляция составит r=1/3.

Следует отметить, что применение эллипсов для описания распределений эквивалентно использованию симметричных эллиптических квантователей:

$$\begin{cases} y_i^2 = \begin{bmatrix} E(\nu_1) - \nu_{1,i} \\ E(\nu_2) - \nu_{2,i} \end{bmatrix}^T \cdot \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} E(\nu_1) - \nu_{1,i} \\ E(\nu_2) - \nu_{2,i} \end{bmatrix} \\ z(y_i^2) = \text{"0"} \quad \text{если} \quad y_i^2 \le k \\ z(y_i^2) = \text{"1"} \quad \text{если} \quad y_i^2 > k \end{cases} \qquad (6),$$

где k –порог квантователя.

Если использовать два квантователя с параметрами r=1/3 и r=-1/3, мы получим корреляционную молекулу, работа которой иллюстрируется рисунком 3.

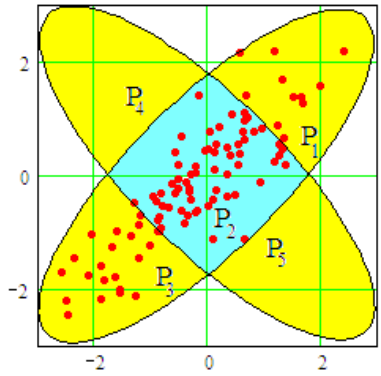

Рис. 3. Работа двух эллиптических квантователей корреляционной молекулы

Работа молекулы построена на том, что число опытов попавших внутрь эллипсов первого и второго квантователя разное:

$$P_1 + P_2 + P_3 > P_2 + P_4 + P_5 \qquad (7)$$

Если исследуемое распределение положительно коррелированно, то число опытов обнаруженное внутри первого квантователя должно быть больше чем число опытов обнаруженное внутри второго квантователя:

$$N_1 > N_2 \qquad (8).$$

Если корреляция отрицательна, то соотношение (8) меняется на противоположное.

Учитывая это, мы получим следующие соотношения для вычисления коэффициентов корреляции:

$$\begin{cases} \tilde{r} \approx 1 - \dfrac{N_1 - N_2}{N_1} & \text{при } N_1 > N_2 \\ \tilde{r} \approx -\left(1 - \dfrac{N_2 - N_1}{N_2}\right) & \text{при } N_2 > N_1 \end{cases} \qquad (9).$$

Очевидно, что вычисления коэффициента корреляции по формуле (3) и по формулам (9) являются разными вычислительными процедурами. Как следствие мы получаем разные значения ошибок Δr и $\Delta\tilde{r}$. Значения этих ошибок положительно коррелированны $r(\Delta r, \Delta\tilde{r}) = 0.519$. То, что их корреляция не единична является предпосылкой для использования вычислений (9) совместно с результатами, полученными по классической формуле Пирсона (3).

Еще одним важным результатом является то, что спектры состояний корреляционной молекулы с линейными квантователями [14] и корреляционной молекулы эллиптическими квантователями (6) имеют разную структуру. Так для ситуации отображенной на рисунке 3 квантовые состояния для двух молекул различаются даже по длине векторов дискретных состояний. Для молекулы с двумя линейными квантователями спектр описывается вектором из 4-х параметров {P_1=31, P_2=9, P_3=52, P_4=8}. Для молекулы с двумя эллиптическими квантователями получается вектор из 5 параметров {P_1=11, P_2=67, P_3=21, P_4=0, P_5=0}. Если корреляция менее 0.333 два последних компонента вектора становятся не нулевыми.

Литература:

1. Monrose F., Reiter M., Li Q., Wetzel S. Cryptographic key generation from voice. //Proc. IEEE Symp. on Security and Privacy, pp. 202-213, 2001. URL: https://www.cs.unc.edu/~reiter/papers/2001/SP2.pdf
2. Ramírez-Ruiz J., Pfeiffer C., Nolazco-Flores J. Cryptographic Keys Generation Using FingerCodes. //Advances in Artificial Intelligence - IBERAMIA-SBIA 2006 (LNCS 4140), p. 178-187, 2006 URL: http://dl.acm.org/citation.cfm?id=2110882

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. Иванов А.И. Нечеткие экстракторы: проблема использования в биометрии и криптографии // Первая миля. № 1, 2015 г. с. 40-47. URL: http://www.lastmile.su/journal/article/4489.
5. ГОСТ Р 52633.0-2006 «Защита информации. Техника защиты информации. Требования к средствам высоконадежной биометрической аутентификации»
6. ГОСТ Р 52633.3-2011 «Защита информации. Техника защиты информации. Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора»
7. ГОСТ Р 52633.5-2011 «Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия-код доступа».
8. 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.kazntu.kz/files/publicate/2014-06-27-11940.pdf)
9. 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.
10. Ложников П.С. Биометрическая защита гибридного документооборота. /Новосибирск. Из-во СО РАН, 2017 г., 130 с.
11. 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.
12. 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.
13. Кулагин В.П., Иванов А.И., Серикова Ю.И. Корректировка методических и случайных составляющих погрешностей вычисления коэффициентов корреляции, возникающих на малых выборках биометрических данных // Информационные технологии. №9 Том. 22 -2016 г. с.705-710. URL: http://novtex.ru/IT/it2016/number09.html
14. Иванов А.И., Серикова Ю.И. Номограммы оценки погрешности, коэффициентов корреляции, вычисленных на малых выборках биометрических данных. //Вопросы радиоэлектроники, № 2, 2015, с 123-130.
15. 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. URL: http://izvuz_tn.pnzgu.ru/tn3416
16. Волчихин В.И., Иванов А.И., Сериков А.В., Серикова Ю.И. Квантовая суперпозиция дискретного спектра состояний математической молекулы корреляции для малых выборок биометрических данных // Вестник Мордовского университета. Т27. №2, 2017, с 230-243.
17. Абезгауз Г.Г., Тронь А.П., Копенкин Ю.Н., Коровина И.А. Справочник по вероятностным расчетами. М.: Воениздат, 1970 г., 536 с.