Всероссийская научно-техническая конференция, посвященная 100-летию со дня рождения Б.И. Рамеева: «Информационно-управляющие и телекоммуникационные системы специального назначения». СЕКЦИЯ: «Биометрическая поддержка криптовалют и блокчейн реестров». Доклад состоится 16 мая, 2018 года, с 11^{45}-12^{00}, конференц-зал Технопарка «РАМЕЕВ», ул. Центральная, 1, г. Пенза.

Серикова Ю.И.

Двойная регуляризация процедур обучения нейронов Махаланобиса за счет симметризации корреляционных связей и компенсации ошибок вычисления коэффициентов парной корреляции биометрических данных

Аннотация.

Актуальность и цели. Целью работы является повышение качества решения задач биометрии за счет регуляризации обучения нейронов Махаланобиса.

Материалы и методы. Используется двухуровневая регуляризация вычислений. На первом уровне от произвольных корреляционных матриц переходят к симметричным матрицам одинаково коррелированных биометрических данных, на втором уровне проводят компенсацию ошибок вычисления коэффициентов корреляционных на малых обучающих выборках примеров образа «Свой».

Результаты. Регуляризация через симметризацию корреляционных связей биометрических данных позволяет исключить из вычислений плохо обусловленную процедуру обращения корреляционных связей. После симметризации задачи она легко табулируется при любом значении размерности. Компенсация ошибок вычисления коэффициентов корреляции между биометрическими данными позволяет многократно снизить требования к размерам обучающих выборок.

Выводы. Предложенные в работе процедуры регуляризации вычислений являются аналогами давно используемых процедур обучения нейронов. Регуляризация за счет симметризации корреляционных связей квадратичных форм является аналогом не итерационных алгоритмов автоматического обучения по ГОСТ Р 52633.5 нейронов с линейными свертками (персептронов). Итерационным алгоритмам обучения нейронов (персептронов) для квадратичных форм являются предложенные в статье итерационные алгоритмы компенсации ошибок вычисления коэффициентов корреляции.

Ключевые слова: коэффициент корреляции, нейроны квадратичных форм, метрика Махаланобиса, симметризация корреляционных связей.

Нейроны с обогащением входных данных их линейным свертыванием

В настоящее время активно идут процессы информатизации общества. Предполагается, что значительный объем информации будет храниться с привлечением облачных сервисов. При этом важнейшим требованием к хранению электронных документов в облаках является их надежная авторизация [1], а в некоторых случаях еще и обезличивание электронных документов [2]. Все эти требования удается выполнить, опираясь на использование нейросетевых преобразователей биометрических данных человека в код его личного криптографического ключа [3]. На данный момент нейросетевые преобразователи биометрия-код стандартизованы, для их автоматического обучения используется алгоритм, рекомендуемый ГОСТ Р 52633.5 [4]. Стандартизованный алгоритм обучения [4] абсолютно устойчив и имеет линейную вычислительную сложность, однако он ориентирован на использование нейронов с линейным обогащением входных биометрических данных. Формально все персептроны и обычные нейроны с иными гладкими функциями возбуждения следует рассматривать как линейную свертку пространства входных состояний с последующим нелинейным преобразованием уже свернутых (обогащенных) данных.

Нейроны с обогащением данных их свертыванием в квадратичном пространстве

Наряду с линейными свертками предварительного обогащения данных, в биометрии часто используются квадратичные свертки входных данных. За этими нейронами закрепилось название радиальных нейронов или радиально-базисных нейронов [5, 6]. Более корректно называть подобные конструкции нейронами квадратичных форм или нейронами Махалонобиса, эллиптическими нейронами. Описываются подобные конструкции следующей системой уравнений:

$$\begin{cases} e^2 = (\overline{v})^T \cdot [R]^{-1} \cdot \overline{v} \\ z(e^2) = \text{"0"} \ \ при \ \ e^2 \le k \\ z(e^2) = \text{"1"} \ \ при \ \ e^2 > k \end{cases} \qquad (1),$$

где k – порог срабатывания выходного квантов ателя нейрона, $\overline{v}$ - вектор нормированных и центрированных биометрических данных, $[R]$- матрица корреляционных связей, контролируемых биометрических данных.

Проблема обучения нейронов Махаланобиса (1) связана с тем, что приходится вычислять коэффициенты корреляции на малых обучающих выборках в 20 примеров образа «Свой». Подобная задача корректна для нейронов Махаланобиса с 2-4 входами. Попытки увеличения входной размерности нейрона приводят к утрате корректности вычислений, в этом случае ошибки вычисления элементов корреляционной матрицы $[\Delta R]$ оказывают большее влияние на результат, чем влияние реальных значений элементов корреляционной матрицы.

Задача становится плохо обусловленной, при ее решении необходимо пользоваться регуляризацией, например, по Тихонову [7].

Симметризация корреляционных связей квадратичных нейронов

Одним из эффективных путей регуляризации вычислений является симметризация корреляционных связей обращаемой матрицы [8, 9]. Общий подход к такой регуляризации построен на том, что для каждого из нейронов выбирают данные, имеющие одинаковые коэффициенты корреляции:

$$\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}, \begin{bmatrix} 1 & r & r \\ r & 1 & r \\ r & r & 1 \end{bmatrix}, \begin{bmatrix} 1 & r & r & r \\ r & 1 & r & r \\ r & r & 1 & r \\ r & r & r & 1 \end{bmatrix}, \begin{bmatrix} 1 & r & r & r & r \\ r & 1 & r & r & r \\ r & r & 1 & r & r \\ r & r & r & 1 & r \\ r & r & r & r & 1 \end{bmatrix}, \begin{bmatrix} 1 & r & r & r & r & r \\ r & 1 & r & r & r & r \\ r & r & 1 & r & r & r \\ r & r & r & 1 & r & r \\ r & r & r & r & 1 & r \\ r & r & r & r & r & 1 \end{bmatrix}, \ldots\ldots \quad (2).$$

Выбрать одинаково коррелированные биометрические параметры из полной не симметричной в выше и ниже диагонали корреляционной матрицы достаточно легко. Размерность полной корреляционной матрицы реальных биометрических данных высока. Так, если пользоваться средой моделирования «БиоНейроАвтограф» [10], то вне диагонали полной корреляционной матрицы размерности 416х416 будут находится 86 112 коэффициента корреляции. Пример распределения значений коэффициентов корреляции реальных биометрических данных приведен на рисунке 1.

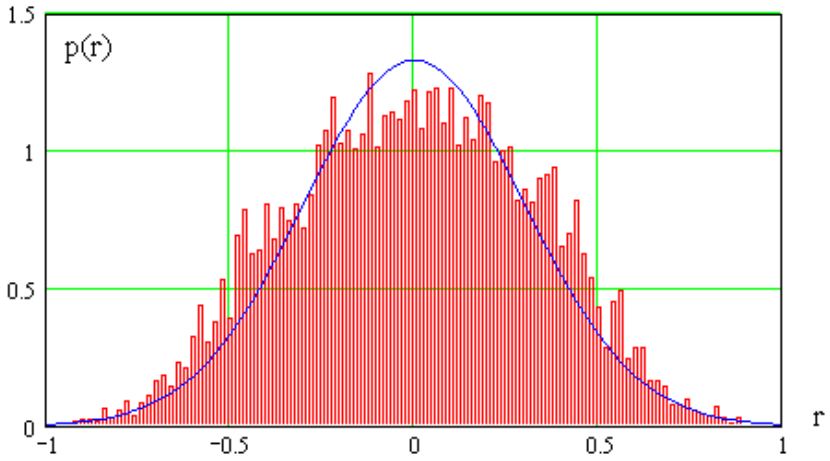

Рис. 1. Распределение значений коэффициентов корреляции биометрических данных, полученных в среде моделирования «БиоНейроАвтограф» [10] для рукописного слова *«Пенза»*

Идеальной является ситуация, когда выбираются данные, чьи коэффициенты корреляции близкие к нулю. В этом случае - мера Махалонобиса (1) будет иметь единичную корреляционную матрицу (проблема обращения корреляционной матрицы исчезает). Однако, по мере увеличения значений модулей коэффициентов равной корреляции вне диагонали матрицы ее коэффициент обусловленности увеличивается. Возникает иллюзия того, что обусловленность задачи обучения нейронов Махаланобиса быстро ухудшается (смотри рисунок 2), однако это не совсем так. Дело в том, что число обусловленности:

$$cond[R] = \frac{\max(\lambda_i)}{\min(\lambda_i)} \qquad (3),$$

является отношением собственных чисел матрицы и отражает проблему неустойчивости вычислений только в первом приближении. При инженерных вычислениях число обусловленности можно рассматривать как коэффициент усиления ошибок исходных данных:

$$\Delta(e^2) \approx cond[R] \cdot E(|\Delta r|) \qquad (4).$$

Если число обусловленности растет (рисунок 2), но одновременно снижается модуль ошибок вычисления исходных данных $|\Delta r| \to 0$, то устойчивость вычислений в целом сохраняется.

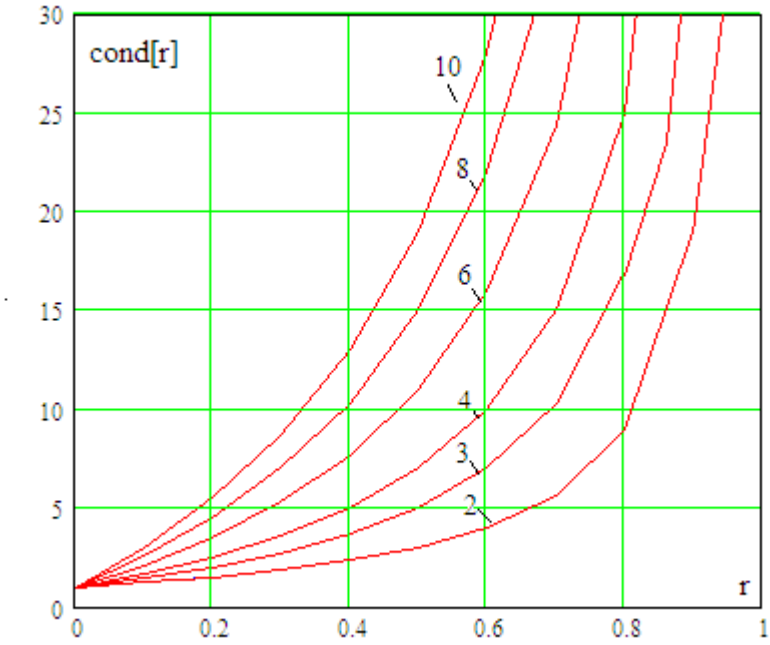

Рис. 2. Значения числа обусловленности симметричных матриц размерностей 2,3,.., 10 как функций равной коррелированности биометрических данных

Эта как раз та ситуация, которая наблюдается на практике. Ошибка вычислений коэффициентов корреляции из-за малого объема обучающей выборки оказывается наибольшей, когда данные слабо коррелированны. На рисунке 3 приведено распределения значений коэффициентов корреляции на малых выборках из 7, 9, …, 21 примеров.

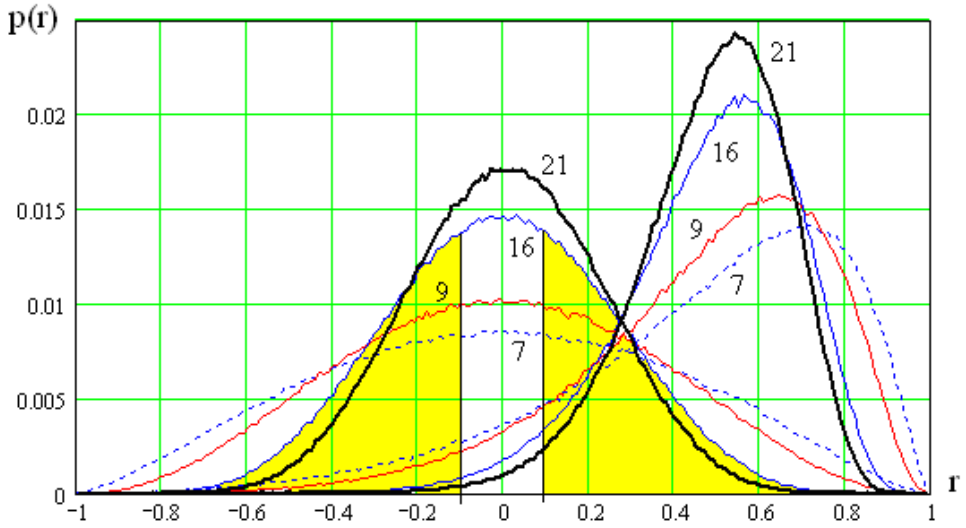

Рис. 3. Распределение значений коэффициентов корреляции на малых выборках для слабо коррелированных данных E(r) =0.0 и сильно коррелированных данных E(r) =0.5 при разных размерах обучающей выборки 7, 9,.., 21

Из рисунка 3 видно, что при обучающей выборке в 16 примеров коэффициент корреляции попадает в интервал от -0.1 до +0.1 с вероятностью только 0.2. Однако интервал неопределенности Δr снижается по мере увеличения модуля коэффициента корреляции. Более того, при росте $|r| \to 1.0$ происходит монотонное снижение

неопределенности ошибки вычисления $|\Delta r| \to 0.0$. Именно эти тенденции и являются тенденциями регуляризации вычислений в целом (4).

Регуляризация обучения за счет компенсации ошибок вычисления коэффициентов парной корреляции

Наиболее устойчивыми являются вычисления, когда из общей корреляционной матрицы выбираются значения $r_{1,i} = 0$. Например, эти значения могут выбираться из первой строки корреляционной матрицы, то есть мы выбираем параметры слабо коррелированные с первым биометрическим параметром. Так как коэффициенты корреляции вычислены с ошибкой, нет смысла находить их точно нулевое значение. Вполне достаточно группировать данные, для которых коэффициенты корреляции первой строки попадают в интервал от -0.05 до + 0.05. Если считать распределение коэффициентов корреляции биометрических данных нормальным (смотри рисунок 1), то в первой строке полной корреляционной матрицы будет обнаружено не менее

$$(\text{pnorm}(0.05,0,1)- \text{pnorm}(-0.05,0,1)\cdot 416=16.589 \qquad (5)$$

слабо коррелированных параметров. Вычисления по формуле (5) выполнено в среде моделирования MathCAD. Это означает, что мы можем использовать нейроны Махаланобиса 17-го порядка. Номера в группе, выделяемых биометрических параметров оказываются монотонно увеличивающимися со случайным интервалом между соседями, например, $\{\nu_1, \nu_{15}, \nu_{41}, ..., \nu_{387}\}$. Следует упростить задачу, осуществив смену нумерации параметров, добившись обычного приращения номера параметра на единицу $\{\nu_1, \nu_3, \nu_3, ..., \nu_{17}\}$. В этом случае формальная запись меры Махаланобиса (1) не меняется, и будет выглядеть следующим образом:

$$e^2 = [\nu_1\ \nu_2\ ....\nu_{17}] \times \begin{bmatrix} 1 & 0 & ::: & 0 \\ 0 & 1 & ::: & 0 \\ ::: & ::: & ::: & ::: \\ 0 & 0 & ::: & 1 \end{bmatrix} \times \begin{bmatrix} \nu_1 \\ \nu_2 \\ ::: \\ \nu_{17} \end{bmatrix} \qquad (6).$$

Однако мы знаем, что все коэффициенты корреляции вычислены с погрешностью, то есть единичная обратная корреляционная матрица в выражении (6) должна иметь некоторые ошибки в элементах, находящихся вне диагонали. Мы имеем право, стабилизировать вычисления, скомпенсировав ошибки вычисления коэффициентов корреляции:

$$e^2 = [\nu_1\ \nu_2\ ....\nu_{17}] \times \begin{bmatrix} 1 & \Delta_{1,2} & ::: & \Delta_{1,17} \\ \Delta_{1,2} & 1 & ::: & \Delta_{2,17} \\ ::: & ::: & ::: & ::: \\ \Delta_{1,17} & \Delta_{2,17} & ::: & 1 \end{bmatrix} \times \begin{bmatrix} \nu_1 \\ \nu_2 \\ ::: \\ \nu_{17} \end{bmatrix} \qquad (7).$$

Для того, что бы скомпенсировать ошибки вычислений коэффициентов корреляции необходимо подобрать значения 128-ти мерного компенсатора $\{\Delta_{1,2,}, \Delta_{1,3}, ..., \Delta_{1,17}, \Delta_{2,3}, \Delta_{2,4}, ..., \Delta_{2,17}, \Delta_{3,4}, ..., \ ,....., \Delta_{16,17}\}$. Эта операция выполняется одной из итерационных процедур направленного подбора, осуществляющей поиск максимума следующего показателя качества:

$$\max\left\{ q = \frac{\left| E(e^2(\overline{\nu})) - E(e^2(\overline{\xi})) \right|}{\sqrt{\sigma(e^2(\overline{\nu})) \cdot \sigma(e^2(\overline{\xi}))}} \right\} \qquad (8),$$

где $\overline{\nu}$ - вектора всех примеров обучающей выборки образа «Свой», $\overline{\xi}$ - вектора всех примеров обучающей выборки образов «Чужие».

В случае если мы выбираем биометрические данные с равной коррелированностью r = 0.05, то мы получим похожую метрику Махаланобиса:

$$e^2 = [\nu_1\ \nu_2\ ....\nu_{17}] \times \begin{bmatrix} 1 & 0.05 & ::: & 0.05 \\ 0.05 & 1 & ::: & 0.05 \\ ::: & ::: & ::: & ::: \\ 0.05 & 0.05 & ::: & 1 \end{bmatrix}^{-1} \times \begin{bmatrix} \nu_1 \\ \nu_2 \\ ::: \\ \nu_{17} \end{bmatrix} \qquad (9).$$

Обращение симметричной корреляционной матрицы в (9) приводит к получению матрицы с аналогичной симметрией:

$$\begin{bmatrix} 1 & r & ::: & r \\ r & 1 & ::: & r \\ ::: & ::: & ::: & ::: \\ r & r & ::: & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & \pi & ::: & \pi \\ \pi & 1 & ::: & \pi \\ ::: & ::: & ::: & ::: \\ \pi & \pi & ::: & 1 \end{bmatrix} \qquad (10).$$

Это означает, что компенсацию ошибок коэффициентов корреляции общей формы меры Махаланобиса, следует выполнять для следующей записи:

$$e^2 = [\nu_1\ \nu_2\ ....\nu_{17}] \times \begin{bmatrix} 1 & \pi+\Delta_{1,2} & ::: & \pi+\Delta_{1,17} \\ \pi+\Delta_{1,2} & 1 & ::: & \pi+\Delta_{2,17} \\ ::: & ::: & ::: & ::: \\ \pi+\Delta_{1,17} & \pi+\Delta_{2,17} & ::: & 1 \end{bmatrix} \times \begin{bmatrix} \nu_1 \\ \nu_2 \\ ::: \\ \nu_{17} \end{bmatrix} \qquad (11).$$

При использовании любых симметризованных корреляционных матриц суть регуляризации через подбор стабилизирующей вычисления матрицы не меняется. Для любой симметризованной матрицы одинаковых коэффициентов корреляции - r следует искать максимум функционала качества (8) вблизи значений элементов обратной корреляционной матрицы - π.

Заключение

Из приведенных выше аргументов следует, что регуляризация обучения нейронов Мехаланобиса должна быть двухступенчатой. Первая ступень регуляризации сводится к выбору одинаково коррелированных между собой биометрических параметров. Только эта процедура уже позволяет значительно снизить число обусловленности, решаемой задачи. Вторым этапом обучения является компенсация ошибок вычисления коэффициентов корреляции, обусловленная недостаточным объемом выборки примеров образа «Свой». Данную регулязарицию можно рассматривать как попытку скомпенсировать ошибки вычисления коэффициентов корреляции из-за малого объема примеров обучающей выборки образа «Свой».

Следует подчеркнуть, что первый этап регуляризации имеет квадратичную вычислительную сложность, так как построен на вычислении полной корреляционной матрицы. Оценка вычислительной сложности второго этапа регуляризации не определена, так как все итерационные процедуры подбора весовых коэффициентов нейронов имеют меняющуюся в процессе обучения вычислительную сложность. В начале обучения вычислительная сложность близка к линейной, однако ее показатель быстро увеличивается в плоть до экспоненциальной вычислительной сложности. Вторая часть регуляризации по своей сути является одним из вариантов «обычного» подбора весовых коэффициентов у нейрона.

ЛИТЕРАТУРА:

1. Ложников П.С. Биометрическая защита гибридного документооборота. /Новосибирск. Из-во СО РАН, 2017 г., 130 с.

2. Гулов В.П., Иванов А.И., Язов Ю.К., Корнеев О.В. Перспектива нейросетевой защиты облачных сервисов через биометрическое обезличивание персональной информации на примере медицинских электронных историй болезни. Вестник новых медицинских технологий (JORNAL OF NEW MEDICAL TECHNOLOGIEC) Том 24, №2 (июнь), 2017 г., с. 220-225.

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. ГОСТ Р 52633.5-2011 «Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия-код доступа».

5. Саймон Хайкин. Нейронные сети: полный курс. М.: «Вильямс», 2006. — С. 1104.

6. Болл Руд и др. Руководство по биометрии. / Болл Руд, Коннел Джонатан Х., Панканти Шарат, Ратха Налини К., Сеньор Эндрю У. // Москва: Техносфера, 2007. -368 с., (перевод с английского).

7. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. М.: Наука, 1979, 248 с.

8. 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.

9. A. I. Ivanov, P. S. Lozhnikov, Yu. I. Serikova Reducing the Size of a Sample Sufficient for Learning Due to the Symmetrization of Correlation Relationships Between Biometric Data // Cybernetics and Systems Analysis, *No. 3, May–June, 2016, pp. 49–56.* http://link.springer.com/article/10.1007/s10559-016-9838-x

10. Иванов А.И., Захаров О.С. Среда моделирования «БиоНейроАвтограф». Программный продукт размещен с 2009 года на сайте АО «ПНИЭИ» http://пниэи.рф/activity/science/noc./bioneuroautugraph.zi для свободного использования университетами России, Белоруссии, Казахстана.

Сведения об авторе:

Серикова Юлия Игоревна – инженер-программист АО НПП «Рубин» 44000, Пенза, ул. Байдукова, 2, E-mail: julia-ska@yandex.ru, ORCID: 0000-0002-4959-321X.