

ОЦЕНКА УСТОЙЧИВОСТИ АЛГОРИТМОВ ОБУЧЕНИЯ БОЛЬШИХ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ БИОМЕТРИЧЕСКИХ ПРИЛОЖЕНИЙ

С. В. Качалин

Пензенский государственный университет
Российская Федерация, 440026, г. Пенза, ул. Красная, 40
E-mail: cnit@pnzgu.ru

Рассматривается проблема повышения скорости, устойчивости и экономичности известных и перспективных алгоритмов обучения больших искусственных нейронных сетей. Для решения этой задачи предлагается создать максимально сбалансированные многокритериальные показатели сравнения разных алгоритмов обучения, ввести численную оценку параметров обучения и планомерно осуществлять работы по их повышению. Утверждается, что устойчивость обучения нейронных сетей является аналогом процедуры оценки числа обусловленности матриц линейной алгебры. Предложен показатель устойчивости обучения, дана связь этого показателя с вычислительной сложностью алгоритмов обучения, а также с числом примеров, на которых алгоритм работоспособен. Обсуждается эффект повышения устойчивости за счет дополнения обучающей выборки синтетическими примерами образа «Свой». Приводится оптимальное количество дополнительных синтетических примеров, добавление которых позволяет уменьшить количество примеров в обучающей выборке. Представлена блок-схема перспективного средства обучения нейросетевых преобразователей, в которой вводится третий автомат размножения примеров образа «Свой». Чтобы контролировать количество синтетических примеров, предлагается создать таблицу ограничений для первых автоматов с учетом мнения экспертов.

Ключевые слова: биометрия, нейронная сеть, обучение, распознавание образов, статистика, вероятность, гистограмма.

Vestnik SibGAU
2014, No. 3(55), P. 68–72

ASSESSMENT OF STABILITY LEARNING ALGORITHMS LARGE ARTIFICIAL NEURAL NETWORKS OF BIOMETRIC APPLICATION

S. V. Kachalin

Penza State University
40, Krasnaya str., Penza, 440026, Russian Federation
E-mail: cnit@pnzgu.ru

The problem of increasing the speed, stability and efficiency of known and emerging algorithms for training large artificial neural networks is considered. To solve this problem it is proposed to create the most balanced multi-criteria performance comparison of different learning algorithms, enter a numerical estimate of the learning parameters and systematically carry out work to improve them. It is argued that resistance training of neural networks is an analogue of the assessment procedure of the condition number of matrices of linear algebra. We propose a measure of stability training, given the connection between this indicator with the computational complexity of learning algorithms, as well as with a number of examples on which the algorithm is efficient. We discuss the effect of increasing the stability by supplementing the training sample synthetic image examples of "your". The optimum amount of additional synthetic examples, the addition of which can reduce the number of examples in the training set is provided. A block diagram of a promising means of training neural network converters, which introduced the third machine reproduction examples image of "Its" is given. To control the number of synthetic examples, it is proposed to create a table of restrictions for the first machines based on expert opinion.

Keywords: biometrics, neural network, learning, pattern recognition, statistics, probability, histogram.

Введение. На сегодняшний день известно порядка 100 алгоритмов обучения искусственных нейронных сетей [1; 2]. В конце XX в. было хорошим тоном предлагать эвристические алгоритмы обучения и их комбинации. При этом сравнить алгоритмы обучения искусственных нейронных сетей оказывается достаточно сложно. У тех, кто занимается обработкой изображений, существуют стандартные изображения, косящие примерно 20 лет из одной работы в другую. Это вполне понятный органолептический подход к сравнению алгоритмов обработки изображений. Читатель сам решает, какая обработка лучше, если на одной странице приведено стандартное изображение после двух разных процедур обработки.

К сожалению, в литературе по обучению искусственных нейронных сетей нет даже такого подхода. Отсутствуют стандартные (типовые) примеры распознаваемых образов, при обучении на которых можно было бы сравнить конкурирующие между собой алгоритмы обучения. Сложность проблемы сравнения алгоритмов обучения искусственных нейронных сетей обусловлена тем, что на них влияют как минимум пять составляющих:

- информативность распознаваемого образа;
- число учитываемых при распознавании параметров;
- число примеров образа, используемых при обучении;
- вычислительная сложность процедур обучения;
- устойчивость процедуры обучения.

Рассматривать влияние сразу всех факторов крайне трудно. Необходимо попытаться снизить число факторов, переопределив их один через другой. Желательно создать максимально сбалансированные многокритериальные показатели сравнения разных алгоритмов обучения.

Актуальность этой задачи высока уже сегодня, в будущем актуальность будет увеличиваться. Каждому придется лично столкнуться с программными роботами, задачей которых будет надежное распознавание биометрического образа «Свой» и еще более надежное распознавание образов «Чужой» при охране персональной информации гражданина, находящейся в личных кабинетах и облачных хранилищах. Быстрое, устойчивое, экономичное обучение возможно, если иметь возможность численной оценки этих параметров обучения и планомерно осуществлять работы по их повышению. Сегодня очевидно, что предстоит значительная работа по совершенствованию алгоритма автоматического обучения по ГОСТ Р 52633.5–2011 [3]. Вводить какие-либо модификации в стандартный алгоритм обучения можно только в том случае, если они улучшат часть показателей, сохранив при этом высокий уровень других показателей.

Обратимся к хорошо изученным преобразованиям линейной алгебры и математической статистики. Известно, что применительно к задаче биометрии можно использовать квадратичные формы [4]:

$$e^2 = (E(\bar{x}) - \bar{x})^T \cdot [\rho]^{-1} \cdot (E(\bar{x}) - \bar{x}), \quad (1)$$

где $\bar{x}$ – вектор из n биометрических параметров образа «Свой»; $E(\bar{x})$ – вектор математических ожиданий биометрических параметров; $[\rho]$ – квадратная матрица ковариационных функций контролируемых биометрических параметров.

Решающее правило строится путем сравнения квадрата эллиптического расстояния (1) с некоторым порогом Θ :

$$\begin{cases} z \leftarrow \langle 1 \rangle, & \text{если } e^2 \leq \Theta, \\ z \leftarrow \langle 0 \rangle, & \text{если } e^2 > \Theta. \end{cases} \quad (2)$$

Квадратичная форма (1) при использовании решающего правила (2) создает некоторую поверхность в форме гиперэллипса с центром в точке математических ожиданий $E(\bar{x})$ контролируемых биометрических параметров. Внутри объема гиперэллипса решающее правило дает состояние «1», соответствующее распределению данных образа «Свой». Вне поверхности решающее правило дает состояние «0», соответствующее множеству данных образов «Чужой».

В терминах биометрии обучению соответствуют процедуры вычисления элементов матрицы коэффициентов ковариации и процедура ее обращения. Очевидно, что для вычисления коэффициентов ковариации потребуется использовать k примеров образа «Свой». Чем больше требуется примеров в обучающей выборке, тем менее экономичной будет процедура обучения.

С другой стороны, число примеров в обучающей выборке можно оценить, зная число обусловленности матрицы коэффициентов ковариации – $\text{cond}[\rho]$. Чем выше число обусловленности, тем больше примеров нужно использовать для вычисления коэффициентов ковариации. В первом приближении будем считать, что связь линейна, число примеров и число обусловленности – это показатель неустойчивости:

$$k \approx \text{cond}[\rho]. \quad (3)$$

Тогда обратная величина будет иметь смысл показателя устойчивости (экономичности) процедур обучения:

$$D_{\text{уст}} \approx \frac{1}{k} \approx \frac{1}{\text{cond}[\rho]}. \quad (4)$$

Минимальное значение $\text{cond}[\rho]$ может быть единственным, т. е. предельно устойчивый алгоритм обучения должен обучаться на одном примере образа «Свой». Если для обучения необходимо 100 примеров, то показатель устойчивости алгоритма обучения будет 0,01. Если алгоритм обучения не работает, его устойчивость нулевая. Это эквивалентно необходимости использовать при обучении бесконечное число примеров $k = \infty$.

При реализации любого решающего правила возникают ошибки первого рода (отказ в доступе «Своему») и ошибки второго рода (ошибочный пропуск

«Чужого»). Биометрические системы легко сравниваются, если вероятности ошибок первого и второго рода одинаковы [5; 6]:

$$P_1 = P_2 = P_{EE}. \quad (5)$$

Однако выполнить условие равной вероятности ошибок (5) для ряда биометрических систем технически невозможно. В связи с этим возможно оценивать равновероятную ошибку как среднее геометрическое двух ошибок [4]:

$$P_{EE} \approx \sqrt{P_1 \cdot P_2}. \quad (6)$$

Чем меньше равновероятная ошибка, тем лучше биометрическая система и, соответственно, алгоритм ее обучения. С другой стороны, система и алгоритм тем хуже, чем больше он требует примеров для своего обучения. Учесть все это можно через использование следующего показателя качества:

$$Q = \frac{-\log_2(P_1) - \log_2(P_2)}{2 \cdot k}. \quad (7)$$

Очевидно, что комплексный показатель качества (7) оказывается способен учитывать множество параметров алгоритма обучения. Пользуясь им, можно сравнивать между собой любые алгоритмы обучения.

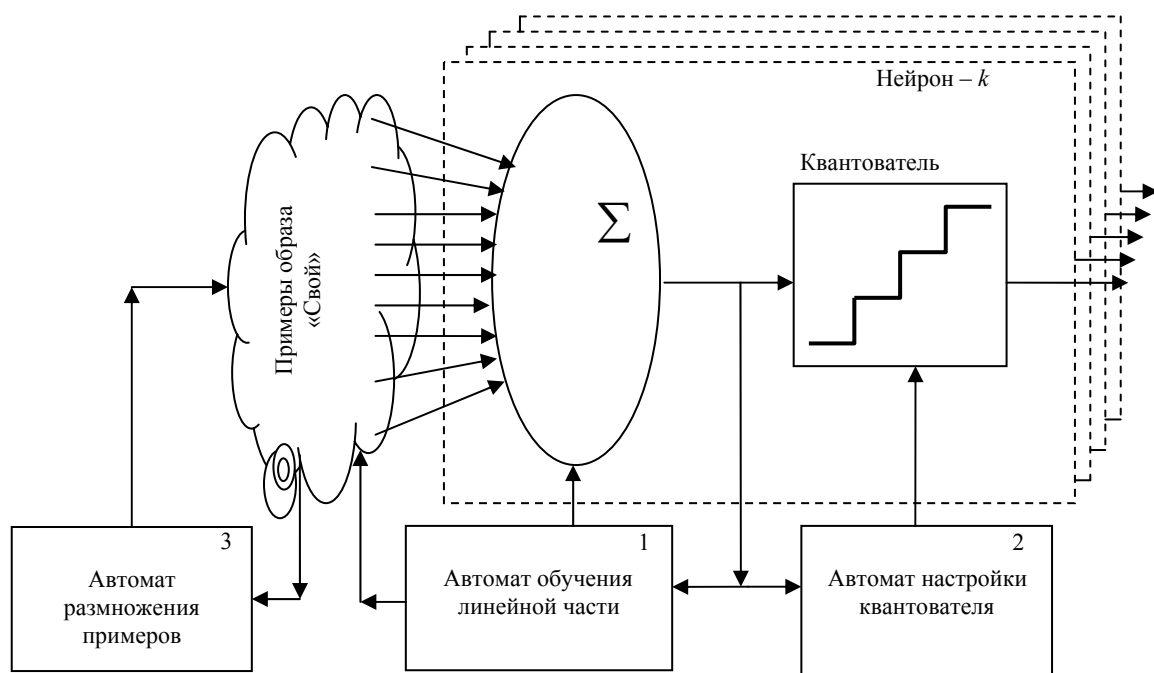
Одним из технических приемов, позволяющих улучшить показатель качества алгоритма, является искусственное увеличение данных в обучающей выборке. Синтезированные дополнительные данные за рубежом [7] имеют устоявшееся название бутстрап-подмножеств. В России ГОСТ Р 52633.2–2010 [8] описывает процедуру синтеза данных примеров – потомков образа «Свой» из двух примеров-родителей.

То есть сформировались два направления, использующие искусственное размножение данных в обучающей выборке.

Практика показывает, что незначительное увеличение данных в обучающей выборке практически всегда приводит к положительному результату [9–11]. Однако существует некоторый оптимум числа дополнительно синтезированных данных. Избыточное раздувание размеров обучающей выборки синтезированными примерами образа «Свой» вредно. В связи с этим возникает вопрос, на какое число синтетических примеров может быть увеличена обучающая выборка.

Предположим, что статистически доказано, что при 21 примере в обучающей выборке можно безопасно добавлять 4 дополнительных синтетических примера. То есть вместо 25 примеров мы можем достичь тех же самых показателей ошибок первого и второго рода на 21 примере. Последнее эквивалентно повышению показателя устойчивости алгоритма обучения (4) на 19 %.

Получается, что в будущем возможно появление дополнения для уже созданных автоматов обучения еще одним дополнительным автоматом, который будет увеличивать обучающие выборки примеров «Свой» некоторым числом дополнительных синтетических примеров. Намечается тенденция явного усложнения (усиления интеллектуальности) средств обучения нейросетевых преобразователей биометрия–код. Блок-схема перспективного средства обучения нейросетевых преобразователей представлена на рисунке.



Блок-схема средства с повышенной устойчивостью обучения за счет дополнения примеров в обучающей выборке

Из рисунка видно, что в дополнение к двум автоматам обучения одиночных искусственных нейронов нейросети необходимо иметь третий автомат размножения примеров образа «Свой».

На сегодняшний день не существует средств обучения искусственной нейронной сети, способных оптимально увеличивать размеры обучающей выборки. Предположительно, что первые автоматы дополнения обучающих выборок будут строиться путем заполнения пустот в гистограммах исходных биометрических данных и иметь жесткие ограничения на число дополнительно созданных примеров. Таблицу ограничений для первых автоматов предполагается создавать с учетом мнения экспертов. Фактически предполагается свернуть некоторую экспертную систему принятия решений до двухмерной таблицы ограничений, учитывающей число примеров-родителей и показатели качества всего образа «Свой», вычисленные по ГОСТ Р 52633.1–2009 [12].

Следует подчеркнуть, что все вышесказанное относится к обычным бинарным искусственным нейронным сетям. Эти сети состоят из бинарных нейронов (сумматора и бинарного квантователя на его выходе). Однако в работах [13–15] показано, что перспективой является переход к нейронам с более сложными квантователями. В частности, на рисунке показан нейрон с трехпороговым выходным квантователем, имеющим 4 выходных состояния «00», «01», «10», «11».

Переход к использованию квантователей с более чем двумя устойчивыми выходными состояниями приводит к значительному росту хеширующих свойств нейронов для образов «Чужой», однако за это приходится платить повышением требований к размерам обучающей выборки биометрических образов «Свой». Затронутая в данной статье проблема устойчивости алгоритмов обучения в ближайшем будущем будет обостряться. Таким образом, введение в биометрические приложения автомата размножения примеров образа «Свой» является одним из возможных решений задачи.

Библиографические ссылки

1. Галушкин А. И. Нейронные сети. Основы теории. М. : Горячая линия-Телеком, 2010. 496 с.
2. Саймон Хайкин. Нейронные сети. Полный курс. Вильямс, 2006. 1104 с.
3. ГОСТ Р 52633.5–2011. Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия – код доступа. М. : Стандартинформ, 2012. 20 с.
4. Дуда Р., Харт П. Распознавание образов и анализ сцен. М. : Мир. 1976. 507 с.
5. Руководство по биометрии / Болл Руд [и др.]. М. : Техносфера, 2007. 368 с.
6. Иванов А. И. Сопоставительный анализ конкурирующих технологий биометрико-криптографической аутентификации личности // Защита информации. INSAID. 2014. № 3. С. 3–9.

7. Efron B. Bootstrap methods: Another look at the Jackknife. Ann. Statistics, 7:1-26, 1979.

8. ГОСТ Р 52633.2–2010. Защита информации. Техника защиты информации. Требования к формированию синтетических биометрических образов, предназначенных для тестирования средств высоконадежной биометрической аутентификации. М. : Стандартинформ, 2011. 24 с.

9. Нейросетевая защита персональных биометрических данных / Ю. К. Язов [и др.]. М. : Радиотехника, 2012. 160 с.

10. Ахметов Б. С., Волчихин В. И., Иванов А. И., Малыгин А. Ю. Алгоритмы тестирования биометрико-нейросетевых механизмов защиты информации. [Электронный ресурс]. Алматы : КазНТУ им. Сатпаева, 2013. 152 с. ISBN 978-101-228-586-4, URL: <http://portal.kazntu.kz/files/publicate/2014-01-04-11940.pdf>.

11. Ахметов Б. С. [и др.]. Технология использования больших нейронных сетей для преобразования нечетких биометрических данных в код ключа доступа : монография. [Электронный ресурс]. Алматы : ТОО «Издательство LEM», 2014. 144 с. URL: <http://portal.kazntu.kz/files/publicate/2014-06-27-11940.pdf>.

12. ГОСТ Р 52633.1–2009. Защита информации. Техника защиты информации. Требования к формированию баз естественных биометрических образов, предназначенных для тестирования средств высоконадежной биометрической аутентификации. М. : Стандартинформ, 2010. 24 с.

13. Иванов А. И. Биометрическая идентификация личности по динамике подсознательных движений : монография. Пенза : ПГУ, 2000. 178 с.

14. Куликов С. В. [и др.]. Учет «тяжелых» хвостов ненормального закона распределения биометрических параметров все «Чужие» при настройке нелинейного элемента нейрона с несколькими дискретными состояниями // Нейрокомпьютеры: разработка, применение. 2012. № 3. С. 56–59.

15. Волчихин В. И. Перспективы использования искусственных нейронных сетей с многоуровневыми квантователями в технологии биометрико-нейросетевой аутентификации / В. И. Волчихин [и др.] // Известия высших учебных заведений. Поволжский регион. Технические науки. 2013. № 4(28). С. 88–99.

References

1. Galushkin A. I. *Neyronnye seti. Osnovy teorii*. [Neural networks. Fundamentals of the theory]. Moscow, Goryachaya liniya-Telekom Publ., 2010, 496 p.
2. Saymon Khaykin. *Neyronnye seti. Polnyy kurs*. [Neural networks. Full course]. Moscow, Vil'yams Publ., 2006, 1104 p.
3. *GOST R 52633.5–2011. Zashchita informatsii. Tekhnika zashchity informatsii. Avtomaticheskoe obuchenie neyrosetevykh preobrazovateley biometriya-kod dostupa* [Information protection. Information protection technology. The neural net biometry-code convertor

automatic training]. Moscow, Standartinform Publ., 2012, 20 p.

4. Richard O. Duda, Peter E. Hart. *Raspoznavanie obrazov i analiz stszen.* [Pattern classification and scene analysis]. Moscow, Mir Publ., 1976, 507 p.

5. Ruud Bolle, Jonathan H. Connell, Sharanth Pankanti, Nalini K. Ratha, Andrew W. Senior. *Rukovodstvo po biometrii* [Guide to Biometrics]. Moscow, Technosfera Publ., 2007, 368 p.

6. Ivanov A. I. [Comparative analysis of competing technologies, biometrics cryptographic authentication identity]. *Zashchita informatsii. Insaid.* 2014, no. 3, p. 3–9. (In Russ.)

7. Efron B. Bootstrap methods: Another look at the Jackknife. *Ann. Statistics*, 7:1-26, 1979.

8. GOST R 52633.2–2010. *Zashchita informatsii. Tehnika zashchity informatsii. Trebovaniya k formirovaniyu sinteticheskikh biometricheskikh obrazov, prednaznachennykh dlya testirovaniya sredstv vysokonadezhnoy biometricheskoy autentifikatsii.* [Information protection. Information protection technology. Requirements for creation procedures for bases of synthetic biometric images, intended for high-reliability biometric authentication means testing]. Moscow, Standartinform Publ., 2011, 24 p.

9. Volchikhin V. I., Ivanov A. I., Funtikov V. A., Nazarov I. G., Yazov Y. K. *Nejrosetevaya zashchita personalnykh biometricheskikh dannykh.* [Neural network protection of personal biometric data]. Moscow, Radio-tekhnika Publ., 2012., 160 p.

10. Akhmetov B. S., Volchikhin V. I., Ivanov A. I., Malygin A. Yu. *Algoritmy testirovaniya biometricheskikh mekhanizmov zashchity informatsii.* [Biometrics-testing algorithms neural network information protection mechanisms]. Kazakhstan, Almaty, KazNTU im. Satpaeva Publ., 2013, 152 p.

11. Akhmetov B. S., Ivanov A. I., Funtikov V. A., Bezyaev A. V., Malygina E. A. *Tekhnologiya is-*

pol'zovaniya bol'shikh neyronnykh setey dlya preobrazovaniya nechetkikh biometricheskikh dannykh v kod klyucha dostupa. [The technology of using large neural networks to convert the fuzzy biometric data in the key code access]. Kazakhstan, Almaty, TOO "Izdatel'stvo LEM" Publ., 2014, 144 p.

12. GOST R 52633.1–2009. *Zashchita informatsii. Tehnika zashchity informatsii. Trebovaniya k formirovaniyu baz estestvennykh biometricheskikh obrazov, prednaznachennykh dlya testirovaniya sredstv vysokonadezhnoy biometricheskoy autentifikatsii.* [Information protection. Information protection technology. Requirements for creation procedures for bases of natural biometric images, intended for high-reliability biometric authentication means testing]. Moscow, Standartinform Publ., 2010, 24 p.

13. Ivanov A. I. *Biometricheskaya identifikatsiya lichnosti po dinamike podsozhatel'nykh dvizheniy.* [Biometric identification on the dynamics of the unconscious movements]. Penza, Penzenskiy gosudarstvennyy universitet Publ., 2000, 178 p.

14. Kulikov S. V., Sekretov M. V., Zakharov O. S., Ivanov A. I., Mayorov A. V. [Accounting for the "heavy" tails of the distribution of abnormal biometrics all "Aliens" when setting up the nonlinear element of a neuron with multiple discrete states]. *Neyrokompyutery: razrabotka, primeneniye*, 2012, no. 3, p. 56–59 (In Russ.)

15. Volchikhin V. I., Ivanov A. I., Funtikov V. A., Malygina E. A. [Prospects for the use of artificial neural networks with multi-level quantizers in biometrics technology, neural network authentication]. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki.* 2013, vol. 28, no. 4, p. 88–89. (In Russ.)

© Качалин С. В., 2014