Труды научно-технической конференции кластера пензенских предприятий,
обеспечивающих **БЕЗОПАСНОСТЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ.**
Том 9, Пенза-2014, с. 32-35 (http://www.pniei.penza.ru/RV-conf/T9/C32)

ПОВЫШЕНИЕ УСТОЙЧИВОСТИ ОБУЧЕНИЯ БОЛЬШИХ НЕЙРОННЫХ СЕТЕЙ ДОПОЛНЕНИЕМ МАЛЫХ ОБУЧАЮЩИХ ВЫБОРОК ПРИМЕРОВ-РОДИТЕЛЕЙ, СИНТЕЗИРОВАННЫМИ БИОМЕТРИЧЕСКИМИ ПРИМЕРАМИ-ПОТОМКАМИ

Качалин С.В. (г. Пенза)

Россия является мировым лидером по технологии использования больших искусственных нейронных сетей для преобразования биометрических данных человека в код его личного криптографического ключа [1] или длинного пароля доступа, состоящего из случайных знаков. Главным в технологии является быстрое автоматическое обучение больших искусственных нейронных сетей по ГОСТ Р 52633.5-2011 [2]. Практика показывает, что маленькие нейронные сети хорошо учатся, но бесполезны, так как принимаемые ими решения оказываются много хуже, чем решения человека. Большие нейронные сети очень плохо учатся. Необходимы специальные меры [1, 2], делающие алгоритмы обучения быстрыми и устойчивыми.

Одной из причин неустойчивости множества существующих алгоритмов обучения искусственных нейронных сетей [1, 3] является ошибка дискретизации непрерывных (континуальных) биометрических данных, возникающая из-за их представления малым числом примеров в обучающей выборке. Практика показывает, что стандартный алгоритм обучения [2] начинает хорошо работать при наличии 20 -:- 30 примеров биометрического образа «Свой» в обучающей выборке. В этом отношении стандартный обучающий автомат [1, 2] работает много хуже в сравнении с обучением человека. Человеку для эффективного обучения достаточно предъявить 2-:-3 примера одного и того же биометрического образа «Свой». В этом отношении нейросетевой искусственный интеллект имеет существенные резервы по снижению размеров обучающей выборки.

В связи с вышеизложенным возникает задача создания специальных математических приемов, позволяющих корректно увеличивать размеры обучающей выборки, например, скрещиванием примеров-родителей и получения примеров-потоков по ГОСТ Р 52633.2-2010 [4]. Так же необходимо создавать автоматы, усиливающие интеллектуальную составляющую алгоритмов обучения больших искусственных нейронных сетей.

Будем исходить из того, что в режиме обучения пользователь предъявил 21 пример биометрического образа «Свой», из которых были извлечены 416 биометрических параметров[1]. Разобьем динамический диапазон первого биометрического параметра – v_1 на 10 интервалов, исходя из усредненного 2-кратного попадания контролируемых значений в каждый из интервалов. Пример одной из получившихся гистограмм распределения биометрических данных приведен на рисунке 1.

[1] Все изложенное в работе каждый может проверить самостоятельно, воспользовавшись бесплатной для университетов России средой моделирования «БиоНейроАвтограф», которую можно скачать по адресу http://пниэи.рф/activity/science/... Программа анализирует 416 биометрических параметра, хранящиеся в виде доступного для чтения файла.

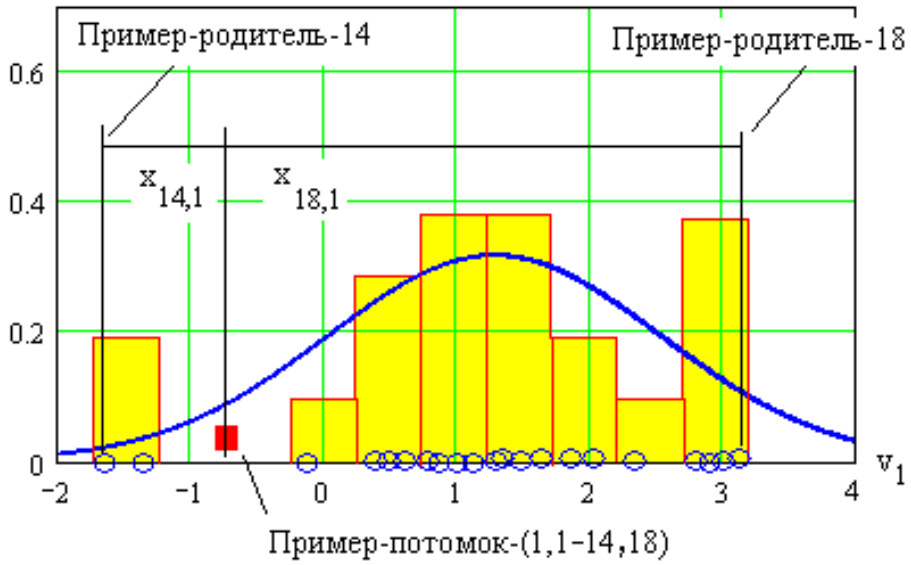

Рис. 1. Получение одного примера-потомка от наиболее далеких примера-родителя-14 и примера-родителя-18 биометрического образа «Свой»

Из рисунка 1 видно, что второй и третий интервалы гистограммы оказались пустыми (не содержат отсчетов). Необходимо получить новый пример-потомок-22 для которого параметр - v_1 попадет в центр пустующего интервала гистограммы. На рисунке 1 положение синтезированного примера-потомока-22 отображено квадратом с темной заливкой.

ГОСТ Р 52633.2-2010 рекомендует скрещивать наиболее разнесенные (не похожие) данные. Для первого биометрического параметра наибольшее значение дает 18-пример, наименьшее значение параметра v_1 дает пример-14. Мы легко можем вычислить расстояние между крайними примерами ($v_{1,18}$-$v_{1,14}$), а так же расстояния между примером-потомком-22 и примерами-родителями. То есть мы можем вычислить коэффициент похожести потомка-22 на первого и второго родителя:

$$\begin{cases} \beta_1 = \dfrac{\left|v_{1,14} - v_{1,22}\right|}{\left|v_{1,14} - v_{1,18}\right|}, \\ \beta_2 = (1-\beta_1) = \dfrac{\left|v_{1,18} - v_{1,22}\right|}{\left|v_{1,14} - v_{1,18}\right|} \end{cases} \qquad (1).$$

Одним из требований к синтезируемым примерам-потомкам является то, что добавление новых (синтетических) образов должно сохранять характерные для данных образа «Свой» корреляционные связи. Этого удается добиться, если синтезировать все данные одного примера с расстояниями пропорциональными расстояниям по первому параметру – v_1. Это удается сделать, если образ-потомок всегда располагать между данными примеров-родителей на расстояниях пропорциональных коэффициентам подобия (1). Пример работы алгоритма такого получения данных для второго параметра – v_2 приведен на рисунке 2.

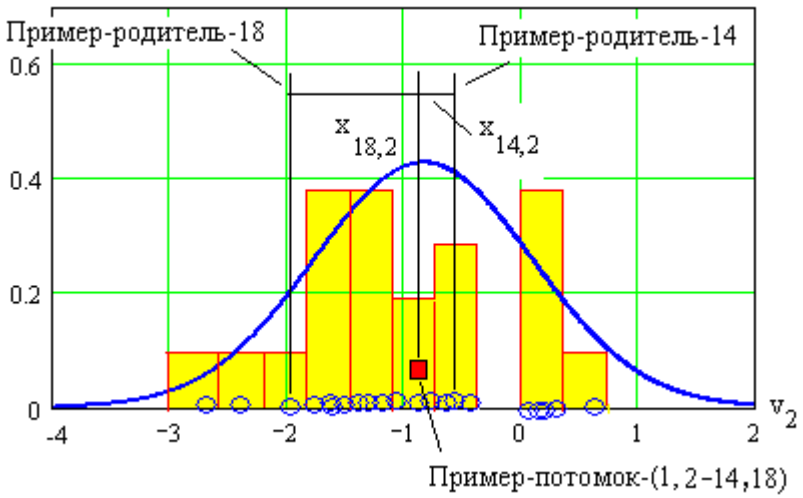

Рис. 2. Получение примера-потомка от примера-родителя-14 и примера-родителя-18 биометрического образа «Свой» при наследовании расстояний до родителей от биометрического параметра v_1

Расстояние по i-му параметру – v_i до образа потомка легко вычисляемо. Если минимальным оказывается значение i-го параметров первого родителя, то для примера-потомка-22 следует задать значение:

$$v_{i,22} = v_{i,14} + \left|v_{i,14} - v_{i,18}\right| \cdot \beta_1 \qquad (2).$$

Если минимальным оказывается значение i-го параметров второго родителя, то для примера-потомка-22 следует задать значение:

$$v_{i,22} = v_{i,18} + \left|v_{i,14} - v_{i,18}\right| \cdot (1 - \beta_1) \qquad (3).$$

Очевидно, что подобные вычисления можно проделать для каждого из оставшихся 415 параметров примера-потомка-22. Затем мы можем повторить процедуру, синтезировав морфингом 23-тий пример-потомок. При этом нужно построить новые гистограммы для уже имеющихся 22 примеров и ориентироваться на заполнение пустот в любой из гистограмм, исключая первую гистограмму. Каждый новый искусственный пример-потомок следует создавать, заполняя пробел в одной из 416 гистограмм. Пользуясь этой тактикой мы можем создать дополнительных 416 примеров-потомков, каждый раз заполняя пробел в какой-то из гистограмм распределения 416-ти параметров. Практика показала, что пробелы в гистограммах практически исчезают при 30-:-50 примерах образа «Свой».

Описанную выше процедуру размножения биометрических данных следует рассматривать как один из вариантов бутстрап предобработки [5] или многомерную морфинг интерполяцию. Именно интерполяцию, так как новые данные всегда размещаются между данными примеров-родителей.

В первом приближении разные алгоритмы обучения искусственных нейронных сетей можно сравнивать по тому сколько примеров образа «Свой» им нужно для качественного распознавания биометрии хозяина. Очевидно, что тот алгоритм, который способен обучаться на 20 примерах образа «Свой» устойчивее другого алгоритма, которому для обучения нужно 40 образов того же образа «Свой».

В этом контексте получается, что повышать устойчивость алгоритма обучения и, соответственно, повышать качество обучения можно путем введения в состав средства биометрической аутентификации блока размножения примеров как это показано на рисунке 3.

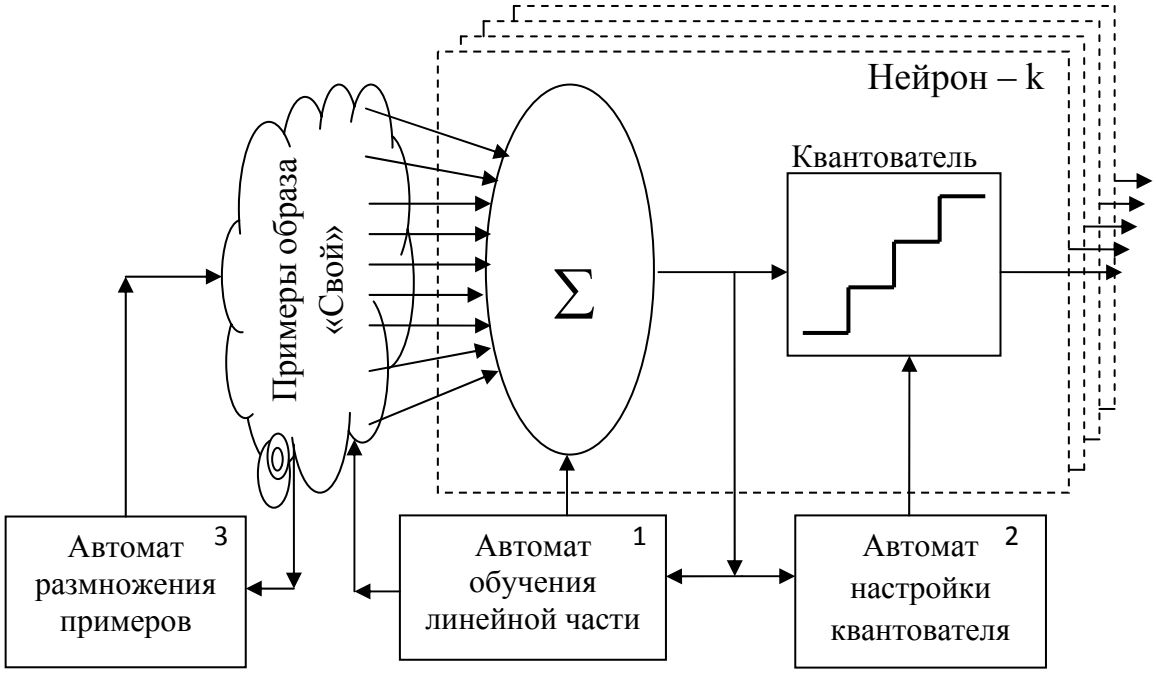

Рис 3. Блок-схема средства с повышенной устойчивостью обучения за счет дополнения примеров-родителей в обучающей выборке примерами-потомками

Из рисунка 3 видно, что в дополнение к двум автоматам обучения одиночных искусственных нейронов нейросети (к блоку-1 и блоку-2) необходимо добавить третий автомат размножения примеров образа «Свой».

Естественно, что увеличивать многократно обучающую выборку нельзя. То есть внутри автомата, синтезирующего примеры-потомки, должны стоять заранее установленные ограничители числа добавленных данных. Предположительно, что подобные ограничители будут учитывать несколько параметров биометрического образа «Свой» и позволять увеличивать обучающую выборку с 11 примеров до 31 примера, что эквивалентно почти трехкратному увеличению устойчивости обучения.

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.

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

3. Хайкин С. Нейронные сети. Полный курс. М.: Вильямс. 2006 г. 1042 с. ISBN 5-8459-0890-6.

4. ГОСТ Р 52633.2-2010 «Защита информации. Техника защиты информации. Требования к формированию синтетических биометрических образов, предназначенных для тестирования средств высоконадежной биометрической аутентификации»

5. Болл Р.М., Коннел Дж.Х., Панканти Ш., Ратха Н.К., Сеньор Э.У. Руководство по биометрии. Москва: Техносфера, 2007. -368 с., ISBN 978-594836-109-3

Материал поступил 29.04.2014, опубликовано по положительной рецензии доктора технических наук Малыгина А.Ю.