Труды научно-технической конференции кластера пензенских предприятий,
обеспечивающих **БЕЗОПАСНОСТЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ.**
Том 9, Пенза-2014, с. 14-17 (http://www.pniei.penza.ru/RV-conf/T9/C14)

ЭКСПЕРТНАЯ СИСТЕМА ДЛЯ АВТОМАТИЧЕСКОГО ОБНАРУЖЕНИЯ И УДАЛЕНИЯ ГРУБЫХ ОШИБОК В МАЛОЙ ОБУЧАЮЩЕЙ ВЫБОРКЕ БИОМЕТРИЧЕСКИХ ДАННЫХ

Гришин В.М. (Пенза)

Автоматическое обучение больших искусственных нейронных сетей [1] предполагает вычисление математического ожидания и среднеквадратического отклонения по малым обучающим выборкам. Чем меньше обучающая выборка, тем результат обучения на ней больше зависит от так называемых грубых ошибок [2]. Возникает необходимость создания экспертной системы, способной автоматически (без участия человека-эксперта) обнаруживать грубые ошибки в обучающих выборках биометрических данных и удалять их. Полностью автоматическое обнаружение и удаление грубых ошибок требуется из-за того, что стороннего человека-эксперта по статистической обработке данных привлекать нельзя [3] по условиям обеспечения информационной безопасности.

Экспертные системы предназначены для решения качественных задач с помощью накапливаемых знаний и получения логических выводов. Последние могут вырабатываться как с помощью формализации собранной от экспертов-специалистов в данной предметной области информации, так и с помощью извлечения знаний из других информационных источников. Кроме того, экспертные системы способны фиксировать неудачные решения и учитывать их в дальнейшем.

Современные математические алгоритмы не могут быть применены в описываемой экспертной системе, т.к они не учитывают априорную информацию, которая включает в себя сведения об используемом законе распределения данных и количестве примеров для работы. Данная экспертная система, благодаря учету всей априорной информации, будет функционировать намного эффективнее, следовательно сделает более точным расчет математического ожидания и среднеквадратического отклонения по малым обучающим выборкам.

Будем считать, что закон распределения данных нормальный, а количество примеров равно 21, математическое ожидание нулевое, среднеквадратическое отклонение единичное.

Проводя оценку математического ожидания и дисперсии для такой случайной выборки из 21 примера, получим оценку интервала значения ошибок из-за малого числа исходных данных:

$$\Delta E(X) \approx \pm 3 \cdot 0{,}22$$

$$\Delta \sigma(X) \approx \pm 3 \cdot 0{,}14$$

Следует подчеркнуть, что столь значительные ошибки вычислений могут быть уменьшены, если заранее точно знать закон распределения значений и выбросить так называемые «грубые ошибки». Обычно человек-эксперт легко обнаруживает и выбрасывает грубые ошибки из обрабатываемой выборки. Однако этот технический прием зачастую вызывает отторжение из-за своего субъективизма.

В связи с этим, возникают следующие вопросы:

1. Можно ли создать автомат, выбрасывающий один самый плохой пример из выборки. Уменьшится ли при этом ошибка вычисления

математического ожидания и среднеквадратического отклонения (т.е., точнее ли будет результат)?

2. Как принять решение о необходимости выбросить «грубую ошибку»? Как отличить хорошие данные от плохих, нуждающихся в цензурированиии?

3. Как отыскать один наихудший пример в обучающей выборке?

На все эти вопросы цензор-человек находит ответ интуитивно, однако привлечение людей-цензоров при обучении средств биометрической аутентификации недопустимо из-за угрозы компрометации конфиденциальных данных. Необходимо создать экспертную систему, которая будет способна выполнять роль цензора-человека и удалять из плохих данных «грубые ошибки».

Именно экспертная система должна давать ответы на поставленные выше вопросы путем решения определенного набора правил, воспроизводящих поведение эксперта при решении аналогичных задач.

Проще всего создать автомат дающий ответ на второй вопрос. Известно, что гистограмма нормального закона распределения входной последовательности симметрична относительно центра.

Асимметричность нормального закона распределения проще всего заметить и распознать на краях последовательности. Следовательно, анализируя гистограммы распределения последовательности, поступающей на вход системы, можно сделать вывод о том, откуда отбрасывать плохой пример – из начала или из конца.

В приведенном ниже случае, дающем ответ на первый вопрос, все расчеты проводились для последовательности из 21 примера, приведенного на рисунке 1.

0	1	2	3	4	5	6	7	8
-1.883	-1.653	-1.249	-0.797	-0.545	-0.525	-0.353	-0.209	-0.012

12	13	14	15	16	17	18	19	20
0.607	0.709	0.831	1.024	1.114	1.161	1.21	1.323	1.433

Рисунок 1 – Входная последовательность из 21 примера

Вычисленные для данной последовательности математическое ожидание и дисперсия равны соответственно:

$$\Delta E(X) \approx +0{,}13$$

$$\Delta\sigma(X) \approx -0{,}04$$

Кроме того, общая ошибка в данном случае будет равна:

$$\sqrt{\Delta E^2(X) + \Delta\sigma^2(X)} \approx 0{,}13$$

Далее, построив гистограмму распределения данной последовательности, изображенную на рисунке 2, можно заметить, что асимметрию можно уменьшить, снизив значение крайней правой величины, другими словами, отбросив последний пример из последовательности, как показано на рисунке 3.

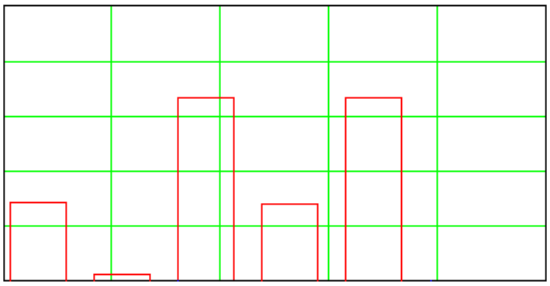

Рис. 2 – Гистограмма распределения входной последовательности из 21 примера, полученная делением динамического диапазона данных на 5 равномерных интервалов

Из рисунка 2 видно, что правый столбик гистограммы много меньше левого, что не соответствует гипотезе о симметричности распределения. Как следствие повысить симметричность распределения можно удалив из выборки наибольшее значение биометрического параметра (рис. 3).

0	1	2	3	4	5	6	7	8
-1.883	-1.653	-1.249	-0.797	-0.545	-0.525	-0.353	-0.209	-0.012

12	13	14	15	16	17	18	19	
0.607	0.709	0.831	1.024	1.114	1.161	1.21	1.323	1.433

Рис. 3 – Процедура отбрасывания последнего примера цензором из плохой последовательности

После этой процедуры, получив последовательность из 20 примеров, был заново осуществлен пересчет основных показателей.

$$\Delta E(X) \approx -0{,}06$$

$$\Delta\sigma(X) \approx -0{,}06$$

$$\sqrt{\Delta E^2(X) + \Delta\sigma^2(X)} \approx 0{,}09$$

Видно, что после процедуры отбрасывания одного плохого примера из плохой последовательности, математическое ожидание уменьшилось практически в два раза, дисперсия увеличилась несущественно, в то время, как величина общей ошибки снизилась. И, что немаловажно, полученные результаты намного ниже, чем для последовательности, подчиняющейся нормальному закону распределения.

Кроме того, экспертная система должна решать задачи по классификации входных данных (хорошие или плохие), а также принимать решения, стоит ли производить отбрасывание одного примера из последовательности (будет ли данная операция эффективна в данном конкретном случае). Для этого в экспертной системе должен быть нейрон, который решает задачу классификации входных последовательностей на «хорошие» и «плохие». И только во втором случае будет проводиться процедура отбрасывания специальным автоматом-цензором. Также необходима база знаний, содержащая алфавит «хороших» данных, выделенных ранее при оптимизации человеком-цензором. Концептуальная модель экспертной системы представлена на рисунке 4.

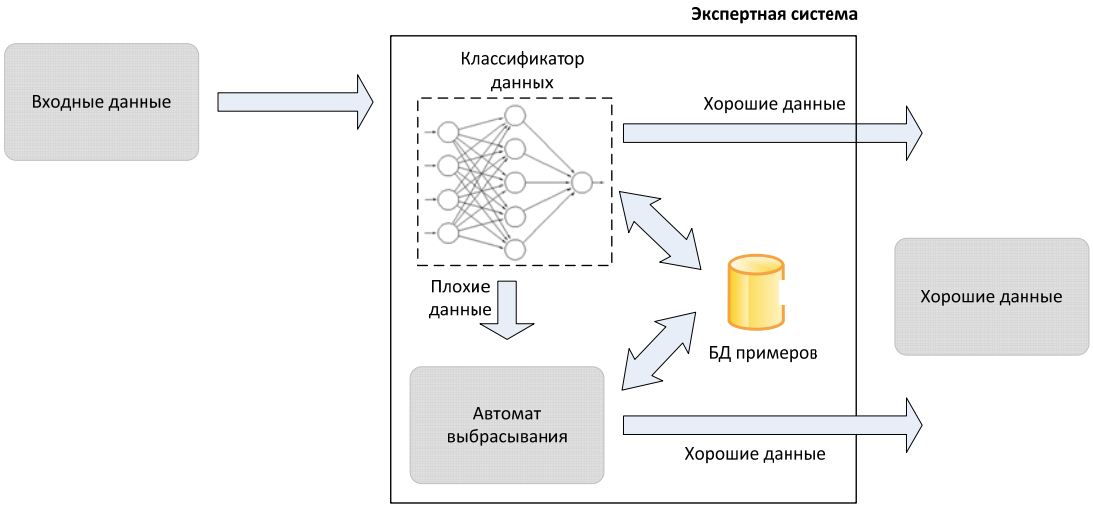

Рис. 4 Экспертная система

В результате можно получить экспертную систему, которая улучшает исходные данные, выбрасывая у части «плохих» последовательностей один наихудший пример. Предварительные оценки показывают, что экспертное цензурирование данных в малых обучающих выборках позволяет снизить относительную ошибку вычисления математического ожидания и среднеквадратического отклонения до 20% при отбрасывании одного наихудшего примера. Это эквивалентно увеличению размеров обучающей выборки с 21 до 27 примеров. Потенциальный выигрыш от создания экспертной системы цензурирования малых обучающих выборок на наличие в них грубых ошибок оказывается значительным.

Литература:

1. ГОСТ Р 52633.5-2011 «Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия-код доступа».
2. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и математических работников. – М.: ФИЗМАТЛИТ, 2006 г. -816 с.
3. Ахметов Б.С., Волчихин В.И., Иванов А.И., Малыгин А.Ю. Алгоритмы тестирования биометрико-нейросетевых механизмов защиты информации Казахстан, Алматы, КазНТУ им. Сатпаева, 2013 г.- 152 с. ISBN 978-101-228-586-4, http://portal.kaznu.kz/files/publicate/2014-01-04-11940.pdf

Материал поступил 27.01.2014 г., публикуется по положительной рецензии к.т.н. Шумкина С.Н.