

## Оптимизация предобработки признаков выборки данных: критерии оптимальности

Царегородцев В.Г.

www.neuropro.ru

tsar@neuropro.ru

Рассматриваются способы предобработки количественных признаков обучающей выборки, индивидуальные для признака и интегральные для выборки критерии оптимальности предобработки. Эксперименты подтверждают ускорение обучения backprop-нейросети при смене способа предобработки, показывают возможность быстрой оценки изменения степени оптимальности предобработки и достижимости ускорения обучения.

Scaling and transformation schemes for continuous variables are discussed with some ideas of optimality criteria selection for a sole variable and for a whole set of variables. Experiments shows that its possible to achieve a great speedup of back propagation network training time as a result of preprocessing scheme change, and the movement to a bad preprocessing scheme can be detected by analysis of changes of optimality criteria values.

### Задача оптимальной предобработки выборки данных

Для искусственных нейронных сетей, обучаемых с учителем градиентными методами на основе метода обратного распространения ошибки, скорость (время) обучения зависит от способа предобработки значений признаков [1,2]. В [1,2] и в этой работе рассматривается оптимизация предобработки количественных независимых признаков (входных сигналов нейросети), поскольку для булевых и номинальных независимых и зависимых признаков схемы предобработки однозначны и вариаций не допускают [3], а для предобработки количественных зависимых признаков необходимо заранее знать диапазоны сигналов, которые могут выдавать выходные нейроны сети.

В качестве индикатора оптимальности предобработки в [1,2] была взята выборочная оценка константы Липшица (КЛ): для выборки  $\{\mathbf{x}_i, \mathbf{y}_i\}, i=1, \dots, N$ , где  $N$  – число примеров выборки, а  $\mathbf{x}_i \in R^n$ ,  $\mathbf{y}_i \in R^m$  – вектора входных сигналов и требуемых выходных сигналов нейросети, оценка КЛ равна

$$L = \max_{i \neq j, \mathbf{x}_i \neq \mathbf{x}_j} \frac{\|\mathbf{y}_i - \mathbf{y}_j\|}{\|\mathbf{x}_i - \mathbf{x}_j\|}.$$

Даже при сохранении неизменными кросс-энтропий между независимыми и зависимыми и величинами, т.е. при линейных масштабированиях количественных признаков, уменьшение КЛ выборки ускоряет обучение нейросети [1,2].

Оценка КЛ требует порядка  $N^2$  вычислений расстояний между парами примеров, поэтому в [1,2] были рассмотрены способы снижения трудозатрат: для задачи классификации и при ограничении вариантов предобработки только монотонными  $R \rightarrow R$  преобразованиями отдельных переменных возможен отбор примеров, лежащих вдоль разделяющей классы поверхности, и оптимизация предобработки (и вычисление КЛ как индикатора её оптимальности) на этом меньшем числе отобранных примеров.

Оценка КЛ выборки и определение пары примеров, дающей максимум КЛ, необходимы, фактически, только для целенаправленной смены схем предобработки отдельных признаков на иные схемы. Так, для дающей максимум КЛ пары примеров можно определить признаки, значения которых различаются на этой паре примеров, и преобразовать эти признаки так, чтобы увеличить различие между этими примерами и

снизить КЛ. Возможные варианты: масштабирование признака в более широкий диапазон значений, порождение из признака двух или более сигналов [3], нелинейная трансформация признака с целью понизить его выборочную плотность распределения в области конфликта и повысить в потенциально менее конфликтных областях [1,2].

Перебор пар примеров выборки необходим также и при поиске дублирующих (одинаковых) и противоречивых (имеющих одинаковые вектора независимых признаков и разные вектора зависимых признаков) примеров. На основе степеней противоречивости и повторяемости данных можно оценивать свойства выборок и целенаправленно управлять данными или переформулировать задачу [4,5]. Но при оптимизации предобработки выборки часто требуется не однократный, а многократный расчет КЛ, что на практике представляется достаточно затруднительным.

Поэтому основными целями настоящей работы являются следующие:

- проверка возможности использования КЛ как индикатора оптимальности предобработки, анализ необходимости применения других или дополнительных индикаторов,
- для схем линейного масштабирования – выстраивание их по гипотетической степени оптимальности и проверка действенности такого упорядочения,
- изучение возможных индикаторов оптимальности предобработки, затраты на вычисление которых линейны по числу примеров выборки, как альтернатив КЛ.

#### Линейное масштабирование значений отдельных признаков

При линейном масштабировании можно предположить, что схема, которая распределит предобработанные значения признака на большем интервале по сравнению с другими способами масштабирования признака, и будет лучшей, поскольку максимальным образом (среди набора используемых схем шкалирования) снизит выборочную КЛ. При этом не требуется точный расчет КЛ выборки до и после каждой нормировки – можно сравнивать интервалы значений переменных по итогам действий схем шкалирования, а интервалы получать трансформацией выборочных минимума и максимума значений признака.

Формула линейного масштабирования значения признака  $x$  для  $i$ -го примера выборки в интервал  $[a,b]$  такова:  $\tilde{x}_i = \frac{(x_i - x_{\min})(b - a)}{(x_{\max} - x_{\min})} + a$ , где  $x_{\min}, x_{\max}$  – минимальное и максимальное выборочные значения признака. Схему для интервала  $[-1,1]$  и возьмем в качестве первой рассматриваемой и базы для сравнения:

$$\tilde{x}_i = \frac{2 \cdot (x_i - x_{\min})}{(x_{\max} - x_{\min})} - 1. \quad (1)$$

Если закон распределения признака  $x$  имеет длинные хвосты, то можно допускать выход значений  $\tilde{x}$  за интервал  $[-1,1]$ : главное, чтобы новый интервал и начальная генерация весов значений синапсов нейросети с самого начала не приводили нейроны к насыщению, что затруднит обучение. Поэтому альтернативой (1) можно взять масштабирование, сдвигающее выборочное среднее  $M(x)$  в 0 и помещающее ближайшее к  $M(x)$  граничное (минимальное или максимальное) выборочное значение в -1 или 1, а другую границу выводящее за интервал  $[-1,1]$ :

$$\tilde{x}_i = \frac{x_i - M(x)}{\min\{M(x) - x_{\min}, x_{\max} - M(x)\}}, \quad (2)$$

что больше подходит для немного асимметричного унимодального закона распределения  $x$ . Если же асимметрия довольно мала и/или закон распределения ближе к нормальному, то увеличить интервал  $\tilde{x}$  по сравнению с (1) и (2) можно нормировкой

$$\tilde{x}_i = \frac{x_i - M(x)}{\sigma(x)}, \quad (3)$$

где  $\sigma(x)$  – выборочное среднее квадратичное отклонение признака  $x$ .

Схемы (2), (3) увеличивают диапазон значений преобразованного признака по сравнению с (1) и потенциально снижают КЛ выборки, если значения именно этого признака различны у пары примеров, задающей КЛ.

### **Редукция исходных интервалов значений признаков для задачи классификации**

В задаче классификации с учителем законы распределения примеров разных классов могут иметь различающиеся средние значения количественного признака и/или разные значения его дисперсий. Проверка статистических гипотез о различиях средних и дисперсий может служить формальным индикатором для возможности выполнения излагаемых далее процедур редукции; дополнительно можно привлекать экспертные мнения о проблемной области и типичности экстремальных выборочных значений.

Если для признака  $x$  можно упорядочить его минимальные значения для каждого класса в ряд  $\min(x|_{\text{класс}K}) < \min(x|_{\text{класс}L}) \leq \dots \leq \min(x|_{\text{класс}M})$ , где между первыми двумя членами ряда выполняется строгое неравенство, то в качестве нового выборочного минимального значения можно взять величину  $x'_{\min}$ :  $\min(x|_{\text{класс}K}) < x'_{\min} < \min(x|_{\text{класс}L})$  и перед преобработкой по формулам (1)-(3) или иным ограничивать диапазон значений признака этим новым граничным значением: для  $i$ -го примера брать  $x'_i = \max\{x'_{\min}, x_i\}$ . Такая редукция оставляет возможность разделения классов  $K, L$  как по исходным, так и по преобработанным значениям этого признака.

Подобная же схема строится и для правой границы интервала значений признака – для редукции максимального выборочного значения.

Фактически, это наиболее простой способ коррекции возможных выбросов в независимых переменных. Конечно, желательно предварительно применять процедуры коррекции выбросов, одновременно рассматривающие все множество независимых переменных (например, [6]), как более строгие и точные, а затем выполнять редукцию для отдельных признаков описанным способом.

Далее схемы масштабирования в виде пары «редукция – одна из формул (1)-(3)» будут соответственно называться (1р), (2р), (3р).

### **Индикаторы свойств выборок**

Укажем свойства выборок, которые можно использовать наряду с оценкой КЛ или взамен её, и затраты на оценку которых линейны по числу примеров выборки.

Геометрию "облака" данных можно описать через доли общей дисперсии данных, соответствующие главным компонентам: можно оценить вытянутость эллипсоида рассеяния в главном направлении и возможность масштабирования данных вдоль осей эллипсоида рассеяния, а не вдоль осей координат – например, для декорреляции признаков и максимизации их совместной энтропии.

Естественно, не всегда применение такого индикатора корректно (например, в случае наличия булевых и номинальных признаков); для повышения надежности желательно применять робастные методы оценивания главных компонент. Понимая это, всё же используем в работе простейший вариант решения – спектральный анализ ковариационной матрицы независимых признаков.

КЛ и доли дисперсии, выбираемые главными компонентами, характеризуют только выборку, а взгляд на выборку со стороны нейромодели отражают, например, свойства матрицы Гессе целевой функции по адаптивным параметрам нейросети.

Скорость оптимизации (обучения) методом наискорейшего спуска очень чувствительна к обусловленности матрицы Гессе, да и методы сопряженных градиентов и квазиньютоновские методы оптимизации используют наискорейший спуск на своей первой итерации или после рестарта – поэтому-то такой критерий и интересен.

Для оценки собственных векторов и чисел матрицы Гессе нейросети в [7] предложена итерационная процедура, не требующая непосредственного вычисления гессиана. Пусть  $\psi$  – вектор с размерностью, равной числу адаптивных параметров нейросети, изначально выбираемый случайно,  $N(\psi) = \psi / \|\psi\|$  означает нормализацию вектора,  $w$  – вектор значений адаптивных параметров сети,  $E(w)$  – значение целевой функции для  $i$ -го примера выборки или среднее на выборке. Действие матрицы Гессе

$H$  на вектор  $\psi$  можно аппроксимировать как  $H\psi \approx \frac{\nabla E(w + a\psi) - \nabla E(w)}{a}$ , а итерации  $\psi \leftarrow HN(\psi)$  сходятся к максимальному собственному вектору  $H$ . От требуемого для каждой итерации двукратного вычисления суммарного градиента выборки можно избавиться, используя градиенты отдельных примеров и экспоненциальное сглаживание:  $\psi \leftarrow (1 - \gamma)\psi + \gamma \frac{\nabla E(w + aN(\psi)) - \nabla E(w)}{a} \Big|_{x_i, y_i}$ , где  $\gamma \in (0, 1)$  – малая

константа. Как указано в [7], оценка при последнем варианте стабилизируется после просмотра нескольких сотен примеров, что быстрее, чем одна итерация на основе суммарных градиентов. Последующие собственные вектора оцениваются аналогично, но на каждой итерации оценку  $\psi$  вдобавок нужно будет ортогонализировать по отношению к ранее найденным собственным векторам процедурой Грама-Шмидта.

### Экспериментальная проверка

Для экспериментов были взяты несколько баз данных из [8], критериями выбора являлись следующие:

- наличие количественных признаков для возможности применения формул (1)-(3);
- задача классификации с учителем (для возможности редукции исходных интервалов значений признаков);
- значительная сложность задачи (число примеров в обучающей выборке);
- простота трансляции данных в формат, воспринимаемый нейропрограммой автора.

Требование большого объема выборки основывалось на результатах [9] о сходимости с ростом объема выборки ошибок обучения и обобщения к асимптотическому значению, характеризующему предел предсказуемости (из-за шума, неинформативности признаков или противоречивости примеров) для задачи – чтобы предполагать, что качество обобщения для обученной сети будет примерно соответствовать качеству решения обучающей выборки, и не использовать никаких дополнительных приемов повышения качества обобщения, которые могли бы повлиять на результаты.

Было отобрано 12 задач (12 баз данных), свойства которых приведены в Таблице 1. Там же указаны размеры нейросетей (число нейронов в скрытом слое и общее число синапсов в сети) для каждой задачи и необходимость ранней остановки обучения (если 100%-ное распознавание при обучении невозможно при разумном объеме нейросети или из-за наличия конфликтных примеров в выборке, то останавливаем обучение сети не при попадании в первый локальный минимум, а при достижении заданного уровня точности решения обучающей выборки).

Признаки для предобработок (2), (3) выбраны эмпирически: взяты признаки с унимодальным неодносторонним законом распределения. Остальные признаки всегда предобрабатываются по схеме (1), как и отобранные признаки в стартовом случае

отсутствия специфически подобранной нормировки. Возможность редукции интервалов значений определялась для всех признаков, а не только для специально отобранных для предобработок (2), (3).

Таблица 1. Свойства исследованных баз данных.

Имя базы данных	Число примеров обучающей / тестовой выборки	Число классов	Число независимых признаков (входных сигналов сети)	Число предобработываемых по (2)-(3) признаков	Число признаков с возможной редукцией выборочного интервала значений		Размер сети (нейронов в скрытом слое / синапсов сети)	Ранний останов обучения
					из всех независимых признаков	из предобработываемых по (2)-(3) признаков		
<b>AnnThyroid</b>	3772 / 3428	3	21	5	5	5	10 / 253	+
<b>HypoThyroid</b>	3163	2	19	5	4	4	10 / 222	+
<b>Letter</b>	20000	26	16	16	4	4	30 / 1316	+
<b>MUSK Clean2</b>	6598	2	166	99	108	88	10 / 1692	-
<b>Opt digits</b>	3823 / 1797	10	62	0	13	0	10 / 740	-
<b>Page blocks</b>	5473	5	10	2	7	0	10 / 165	+
<b>Pen digits</b>	7494 / 3498	10	16	1	0	0	10 / 280	+
<b>Pima</b>	968	2	8	5	7	4	25 / 277	+
<b>Satellite</b>	4435 / 2000	6	36	36	36	36	15 / 651	+
<b>Statlog Shuttle</b>	43500 / 14500	7	9	9	9	9	15 / 262	+
<b>Spambase</b>	4601	2	57	0	54	0	20 / 1202	+
<b>Yeast</b>	1484	10	8	6	5	5	40 / 770	+

Начальные веса синапсов генерируются в интервале  $[-0.1, 0.1]$  и затем при обучении не ограничиваются. Сети имеют один скрытый слой нейронов с сигмоидной нелинейностью  $f(z) = z / (0.1 + |z|)$ , нейроны выходного слоя без нелинейностей. Обучение идет методом сопряженных градиентов с оптимизацией шага на каждой итерации обучения. Использована специальная классификаторная целевая функция из [3], веса классов обратно пропорциональны числу примеров класса. Для каждой ситуации (задача и способ предобработки) обучено 25 нейросетей.

Достигнутые скорости обучения (среднее число итераций метода сопряженных градиентов) даны в Таблице 2 для предобработок (1), (2), (3), (1p), (2p), (3p) на фоне соответствующих значений КЛ. Для каждой задачи способы предобработки упорядочены по уменьшению КЛ выборки. В таблице также указаны доли общей дисперсии набора независимых переменных, выбираемой первой главной компонентой в каждом случае, среднее значение максимальных собственных чисел  $\lambda_{\max}$  матриц Гессе нейросетей перед их обучением, средние значения ошибок обобщения.

Жирным шрифтом в Таблице 2 выделены ситуации со статистически достоверным увеличением среднего числа шагов обучения по сравнению с достигнутым на предыдущих шагах минимумом, т.е. когда новая нормировка замедлила скорость обучения, а курсивом – показатели отличных от КЛ интегральных индикаторов оптимальности выборки, которые показали причину замедления обучения.

Видно, что для ускорения обучения в разы необходимо снижать КЛ тоже в разы, а для ускорения на порядок – снижать КЛ на порядок как сменой способа масштабирования, так редукцией начальных интервалов значений признаков.

Рядам (1)-(2)-(3), (1p)-(2p)-(3p) предобработок соответствует снижение времени обучения вдоль таких рядов. Исключения объясняются в следующем разделе.

**Таблица 2.** Свойства выборок данных и нейросетей для вариантов предобработки.

База данных	Результаты экспериментов, упорядоченные по убыванию КП. <i>N</i> – вариант предобработки, <i>Итоги</i> – результаты: первая строка – оценка КП выборки, вторая – число шагов обучения, третья – среднее значение оценки $\lambda_{\max}$ для необученных сетей, четвертая – доля общей дисперсии независимых признаков, описываемая первой главной компонентой, пятая (у отдельных задач) – % неправильно решенных тестовых примеров											
	1		2		3		4		5		6	
	<i>N</i>	<i>Итоги</i>	<i>N</i>	<i>Итоги</i>	<i>N</i>	<i>Итоги</i>	<i>N</i>	<i>Итоги</i>	<i>N</i>	<i>Итоги</i>	<i>N</i>	<i>Итоги</i>
<b>AnnThyroid</b>	1	102.60 1052.3 5.10 0.151 2.98	1p	63.63 740.4 5.07 0.135 2.98	2	63.38 773.0 5.03 0.138 2.99	2p	60.51 678.2 5.16 0.130 3.06	3	41.93 601.4 4.91 0.156 3.15	3p	41.18 <b>684.3</b> 5.23 0.171 3.19
<b>HypoThyroid</b>	1	40.22 1031.6 6.52 0.144	2	25.51 548.4 6.28 0.129	1p	24.5 539.0 6.36 0.126	2p	18.48 460.1 6.21 0.169	3	15.31 420.0 5.97 0.145	3p	12.82 356.6 6.03 0.195
<b>Letter</b>	1	21.21 307.6 1.92 0.154	1p	21.21 307.7 1.90 0.151	2	15.20 250.6 1.37 0.180	2p	15.20 244.4 1.37 0.180	3	7.26 <b>338.4</b> 1.93 0.146	3p	7.26 <b>331.9</b> 1.27 0.146
<b>MUSK Clean2</b>	1	3.84 191.04 7.15 0.048	1p	3.11 141.84 6.23 0.044	2	3.08 136.36 5.47 0.045	2p	2.99 121.44 5.17 0.045	3	2.44 122.16 5.62 0.045	3p	2.26 122.92 5.29 0.044
<b>OptDigits</b>	1	1.66 223.3 2.64 0.067 4.70	1p	1.66 218.8 2.66 0.066 4.66								
<b>PageBlocks</b>	1	4165 242.1 22.82 0.343	2	4165 <b>262.1</b> 24.84 0.425	3	4165 <b>338.3</b> 30.1 0.459	1p	540.3 138.2 21.58 0.237	2p	540.3 128.6 24.14 0.308	3p	540.3 <b>143.2</b> 30.04 0.355
<b>PenDigits</b>	1	12.50 179.3 1.40 0.168 4.75	2	12.41 176.1 1.40 0.167 4.79	3	11.08 180.6 1.39 0.162 4.72						
<b>Pima</b>	1	26.80 1260.4 1.82 0.208	1p	24.0 1005.4 1.75 0.211	2	17.33 827.4 1.69 0.252	2p	16.90 816.7 1.65 0.239	3	11.40 464.7 1.64 0.228	3p	8.76 <b>2285.6</b> 9.79 0.763
<b>Satellite</b>	1	9.24 762.6 2.44 0.195 14.16	2	8.61 556.4 2.0 0.211 14.13	1p	7.78 595.6 2.55 0.193 14.4	2p	7.13 510.2 2.09 0.218 14.35	3	3.59 <b>563.1</b> 3.21 0.230 14.54	3p	3.57 <b>564.9</b> 3.19 0.228 14.59
<b>Statlog Shuttle</b>	1	2333 856.4 4.27 0.464 0.184	2	2272.6 603.2 7.57 0.589 0.186	1p	1461.3 <b>692.2</b> 4.72 0.490 0.191	2p	990.2 445.7 10.79 0.600 0.181	3	36.80 106.3 13.03 0.244 0.213	3p	28.97 108.9 13.61 0.225 0.204
<b>Spambase</b>	1	22405 299.8 15.03 0.0577	1p	9898.4 87.04 13.57 0.168								
<b>Yeast</b>	1	75.82	1p	61.87	2	50.99	2p	50.96	3	15.21	3p	15.11

	1024.9	815.1	501.9	466.2	190.6	188.4
	5.64	5.44	5.09	5.11	6.04	6.09
	0.215	0.206	0.219	0.212	0.215	0.215

Таким образом, можно при сравнении схем предобработки оценить порядок возможного выигрыша или проигрыша в скорости обучения как коэффициент, равный отношению длины нового интервала значений предобработанного признака к длине интервала при предыдущем варианте предобработки.

Наличие всего пяти задач с тестовыми выборками не позволяет однозначно сделать выводы про влияние на качество обобщения. Тем более, что к двум задачам (OptDigits, PenDigits) были применены только отдельные варианты предобработок. Но для трех задач (AnnThyroid, Satellite, Shuttle) имеется тренд ухудшения обобщения с уменьшением КЛ выборки: это означает, что сложность выборок при предобработке действительно снижается, нейросеть становится более избыточной и может излишне настроиться на свойства, присущие только обучающей выборке, а не всей генеральной совокупности, т.е. приводить к проявлению эффекта overfitting.

Комментарий вне основной темы. В [7] для обучения с неадаптивным шагом предлагается выбирать шаг, равный  $1/\lambda_{\max}$ , поэтому оценки  $\lambda_{\max}$  из Таблицы 2 могут показаться неверными: величины шагов на основе этих оценок будут завышены на порядки и на практике дадут расхождение градиентного обучения. Причина в том, что табличные значения получены при специализированной целевой функции – для МНК-критерия они бы выросли не меньше чем в 25 ÷ 30 раз для задач с двумя классами и еще больше для задач с бóльшим числом классов, что привело бы к адекватности оценки величин шагов вдоль направления спуска по оценкам  $\lambda_{\max}$ .

#### **Анализ поведения индикаторов оптимальности предобработки**

Как видно из Таблицы 2, не всегда снижение КЛ выборки приводит к уменьшению времени обучения: задача AnnThyroid при предобработке (3p), Letter при (3) и (3p), Pima при (3p), Satellite при (3), (3p), Shuttle при (1p). Почти всем этим случаям, как и негативному тренду задачи PageBlocks, можно дать объяснение исходя из изменения значений дополнительных индикаторов.

Рост первого собственного значения (и, вероятно, обусловленности) матрицы Гессе необученной нейросети по сравнению с предыдущими случаями приводит к увеличению времени обучения: все описанные случаи, кроме случаев Shuttle (1p) и Letter (3p), обладают этим свойством.

Увеличение вытянутости облака данных в главном направлении (увеличение доли дисперсии, выбираемой первой главной компонентой) иногда не играет ухудшающей роли даже при довольно значительном увеличении вытянутости (например, NuroThyroid при (2p), (3), (3p), Shuttle при (2), (2p)), но иногда приводит к ухудшению или даже катастрофе: AnnThyroid при (3p), PageBlocks при (2), (3), (3p), Pima при (3p), Satellite при (3), (3p). Потому можно предположить, что максимизация совместной энтропии независимых переменных, т.е. увеличение доли объема, заполняемого точками выборки в параллелепипеде (координатами вершин которого по каждой оси являются минимальные и максимальные значения предобработанных признаков) не всегда будет оптимальным способом предобработки.

Т.о., эксперименты показывают, что дополнительные индикаторы, менее затратные по сравнению с оценкой КЛ, позволяют достаточно надежно сопоставлять между собой схемы предобработки и делать вывод в пользу той или иной схемы.

Покажем, почему нельзя полностью отказаться от использования КЛ. Можно сформулировать два требования к оптимальной предобработке и, соответственно, описать

индикаторы выполнения этих требований:

1. оптимальная схема предобработки должна минимизировать обусловленность матрицы Гессе нейросети в момент начала обучения;
2. оптимальная предобработка должна приводить к низкой КЛ выборки как индикатору оптимальности для завершающих шагов обучения (чем бóльший скачок поверхности отклика придется аппроксимировать, тем предположительно больше будет обусловленность гессиана целевой функции в этой области высоких чувствительностей выходного сигнала сети к изменению входных сигналов).

Второй пункт говорит, что только изучение нескольких пар примеров, порождающих максимальные и близкие к ним значения КЛ, и покажет, вдоль каких главных компонент необходимо производить масштабирование данных: вдоль тех главных компонент, которые преимущественно заданы осями координат, по которым имеются минимальные отличия у этих конфликтных пар примеров. Поэтому и нельзя связать оптимальность предобработки просто с максимизацией совместной энтропии данных – нужны локализация конфликтной области (области значений признаков для примеров, порождающих высокие значения КЛ, – через эту область и должна будет пройти разделяющая классы поверхность) и увеличение изменения объема этой области при масштабировании значений признаков.

### Заключение

Для оценки оптимальности предобработки обучающей выборки рассмотрено два интегральных для выборки критерия, вычислительно менее затратные по сравнению с ранее использовавшейся оценкой константы Липшица; экспериментально показана достаточная синхронность поведения этих индикаторов и их чувствительность к ухудшению способа предобработки, иногда не идентифицируемому только по оценке КЛ. При использовании отдельных интегральных критериев (КЛ выборки, главные направления облака данных) возможно получение рекомендаций по смене предобработки уже отдельных признаков выборки.

Эксперименты также подтверждают возможность значительного ускорения обучения нейросети при смене способа предобработки даже при использовании таких эффективных методов оптимизации, как метод сопряженных градиентов.

Трудность вопроса оптимизации предобработки данных применительно к нейронным сетям обусловлена нелинейностью обучаемой нейромодели – поэтому в работе сделана попытка сформировать требования к критерию оптимальности в разные моменты процесса обучения: оптимальная предобработка должна обеспечивать быстрое градиентное обучение на первых шагах обучения (когда мощные методы оптимизации еще не успели хорошо аппроксимировать кривизну второго порядка целевой функции, и оптимизация идет преимущественно вдоль градиентных направлений), а аппроксимируемая нейросетью функция  $R^n \rightarrow R^m$  не должна иметь резких скачков, поскольку трудно (и долго при градиентной адаптации параметров нелинейной модели, что и происходит в случае обучения нейросети) хорошо аппроксимировать область высокой чувствительности выходных сигналов сети и правильности решения задачи к малым изменениям входных сигналов.

Именно попытка рассмотреть индикаторы оптимальности и их поведение и отличает данную работу от иных. Последние (например, [10,11]) предлагают изменять алгоритмы обучения для декорреляций сигналов или их центрирования, в том числе центрирования динамически изменяющихся в процессе обучения величин невязок и сигналов нейронов внутренних слоёв сети. Хотя центрирование сигналов и действительно [2,11], изменение алгоритмов обучения не позволяет ни сопоставлять между собой трансформированные разными способами выборки, ни предлагать советов по

изменению способа трансформации (предобработки) выборки или её переменных.

### Литература

1. *Царегородцев В.Г.* Предобработка обучающей выборки, выборочная константа Липшица и свойства обученных нейронных сетей // Материалы X Всеросс. семинара "Нейроинформатика и ее приложения", Красноярск, 2002. 185с. – С.146-150.
2. *Царегородцев В.Г.* Оптимизация предобработки данных: константа Липшица обучающей выборки и свойства обученных нейронных сетей // Нейрокомпьютеры: разработка, применение. 2003, №7. – С.3-8.
3. *Миркес Е.М.* Нейрокомпьютер: проект стандарта. Новосибирск: Наука, 1999. - 337с.
4. *Tiv E., Refenes A.N.* Removal of catastrophic noise in hetero-associative training samples / Proc. IJCNN, Nagoya, Japan, 1993. Vol.3. – pp.2628-2633.
5. *Крислов Р.А., Тарасенко В.А.* Предварительная оценка качества обучающей выборки для нейронных сетей в задачах прогнозирования временных рядов / Тр. Одес. Политехн. Ун-та. Одесса, 2001. - Вып.1. – С.90-93.
6. *Hämäläinen J.J., Järvinäki I.* Input projection method for safe use of neural networks based on process data / Proc. IJCNN, Anchorage, Alaska, USA, 1998. – pp.193-198.
7. *LeCun Y., Simard P.Y., Pearlmutter B.* Automatic learning rate maximization by on-line estimation of the Hessian's eigenvectors / Advances in Neural Information Processing Systems 5 (1992). Morgan Kaufmann, 1993. – pp.156-163.
8. UCI KDD Database Repository. <http://kdd.ics.uci.edu/>
9. *Cortes C., Jackel L.D., Solla S.A., Vapnik V., Denker J.S.* Learning curves: Asymptotic values and rate of convergence / Advances in Neural Information Processing Systems 6 (1993). Morgan Kaufmann, 1994. – pp.327-334.
10. *Pérez-Illarbe M.J.* Preconditioning method to accelerate neural networks gradient training algorithms / Proc. IJCNN, Washington, DC, USA, 1999. - 5p.
11. *Schraudolph N.N.* Centering neural network gradient factors / Neural networks: Tricks of the trade (G.Orr and K-R.Müller eds). Springer Verlag. Lecture Notes in Comp. Sci., Vol.1524. 1998. – pp.207-226.