УДК 519.254

ПРЕДОБРАБОТКА ДАННЫХ ДЛЯ ОБУЧЕНИЯ НЕЙРОННОЙ СЕТИ

Качановский Ю.П., Коротков Е.А.

ГОУ ВПО «Липецкий государственный технический университет», Липецк;
ОАО «Новолипецкий металлургический комбинат», Липецк, e-mail: evg.korotkov@mail.ru

В данных, используемых нейронной сетью для обучения, могут присутствовать противоречивые сведения. Чтобы исключить такие данные предлагается использовать профиль компактности, который изначально применялся в задаче классификации. Для сведения задачи предобработки данных в моделировании к задаче предобработки в классификации выполняется процедура кластерного анализа над выходными параметрами нейросетевой модели. Эффективность данного решения подтверждается результатами проведенного эксперимента.

Ключевые слова: нейронные сети, предобработка данных, профиль компактности

PREPROCESSING DATA FOR TRAINING OF NEURAL NETWORKS

Kachanovskij J.P., Korotkov E.A.

Lipetsk state technical university, Lipetsk;
Novolipetsk iron and steel corporation, e-mail: evg.korotkov@mail.ru

Inconsistent data can be present In the data used by neural networks for training. To exclude such data it is offered to use a profile of compactness which was initially applied in a classification problem. The procedure of cluster analysis is performed over target parameters of neural models to transform task of preprocessing data in classification to task of preprocessing data in modelling. Efficiency of the given decision is confirmed by results of the experiment.

Keywords: neural networks, preprocessing data, profile of compactness

В работах [5, 6] для решения задачи моделировании сквозной технологии производства холоднокатаного проката (автолистовой стали), характеризующейся большим числом учитываемых параметров, предложено использовать нейронную сеть. Обучение нейронной сети на первичных экспериментальных производственных данных часто не приводит к ожидаемому результату. Это связано с тем, что в данных для обучения могут присутствовать противоречивые экспериментальные значения, т.е. одинаковые значения векторов независимых парамеров имеют разные значения векторов зависимых параметров [8]. К противоречивости приводят погрешности измерений и субъективные ошибки (часть данных снимается вручную). Противоречивость будет также наблюдаться, если в модель не включены значимые входные параметры (ошибка выбора параметров модели). Такие строки данных следует исключать из обучающей выборки, так как они влияют на скорость обучения, величину ошибок обучения и обобщения и могут привести, в конечном счете, к неработоспособности построенной нейронной модели. Пример подобных данных приведен в таблице.

Данные таблицы представляют собой фрагмент экспериментального массива данных, используемого для обучения нейронной сети, которая предназначена для моделирования зависимости показателей качества готовой продукции (механических свойств автолистового проката – выходные параметры модели) от управляемых технологических характеристик производства (параметров горячей, холодной прокатки и непрерывного отжига – входные параметры модели). Из 19 входных параметров только в 2-х случаях отклонение между значениями превышает 1%, при этом у 4-х выходных параметров из 5 отклонение выше 1%, т.е. входные значения отличаются незначительно, а значения выходных параметров отличаются существенно.

Повышение качества обучения нейронной сети в этом случае возможно только при использовании эффективных методов предобработки данных. Для обработки данных перед обучением нейронной сети предлагается использовать метод, основанный на понятии профиля компактности и комбинаторных формулах для эффективного вычисления функционала скользящего контроля. Метод изначально применялся для подготовки данных в задачах классификации [2]. Чтобы свести свою задачу предобработки данных в моделировании к задаче предобработки данных при классификации данных (которая имеет решение, использующее профиль компактности), предлагается провести кластерный анализ на выходных параметрах нейронной сети. Таким образом, будем иметь задачу классификации, для которой известно решение предобработки данных.

Пример противоречивых данных

Тип параметра	Название параметра	Значение 1-й строки данных	Значение 2-й строки данных	Относительное отклонение значений, %
Входные параметры	Химсостав. C	0,087	0,087	0
	Химсостав. Mn	0,44	0,44	0
	Химсостав. P	0,01	0,01	0
	Химсостав. S	0,013	0,013	0
	Химсостав. Ni	0,02	0,02	0
	Химсостав. Al	0,034	0,034	0
	Химсостав. N2	0,003	0,003	0
	Стан 2000. Температура проката после клети 5	1000	1002	0,2
	Стан 2000. Температура конца прокатки	829	830	0,12
	Стан 2000. Температура смотки	656	656	0
	Стан 2000. Степень обжатия на клети 12	0,149	0,142	4,93
	Стан 2000. Скорость прокатки клети 12	634	616	2,92
	Стан 2030. Скорость прокатки клети 5	14,68	14,59	0,71
	Стан 2030. Степень обжатия	0,8	0,8	0
	АНО. Температура нагрева	678,1	678,1	0
	АНО. Температура выдержки	679,3	679,8	0,07
	АНО. Температура охлаждения	299,3	299,7	0,13
	АНО. Температура предвар. охлажд.	348,8	348,5	0,09
	АНО. Температура повторного нагрева	385	384,6	0,1
Выходные параметры	Предел прочности	365	360	1,39
	Относительное удлинение	36	34	5,88
	Глубина сферической лунки по эриксену	57	55	3,64
	Твердость по роквеллу	9,2	9,3	1,08
	Предел текучести	290	290	0

По методу происходит разделение обучающих объектов на 3 категории: шумовые выбросы, неинформативные (периферийные) объекты и опорные объекты. Исключение шумовых и периферийных объектов из обучающей выборки повышает обобщающую способность метода обучения.

Данный метод, начиная с полной выборки, последовательно исключает объекты. На каждом шаге выбирается тот объект, исключение которого минимизирует функционал. Оказывается, что процесс отсева объектов разбивается на две стадии. Сначала исключаются шумовые выбросы, затем исключаются неинформативные периферийные объекты. Процесс останавливается, когда остаются объекты, исключение которых заметно увеличивает функционал, тогда в массиве данных остаются столпы, или опорные объекты.

Основным результатом применения комбинаторной формулы для оценки функционала полного скользящего контроля является то, что она одинаково хорошо подходит как для исключения шумовых объектов, так и для сокращения множества прецедентов, являясь при этом эффективно вычислимым, точным значением функционала [3].

Метод опирается на предположение, которое называется гипотезой компактности: схожие объекты гораздо чаще лежат в одном классе, чем в разных. В этом случае граница между классами имеет достаточно простую форму, а классы образуют компактно локализованные области в пространстве объектов (в математическом анализе компактными называются ограниченные замкнутые множества, гипотеза компактности не имеет ничего общего с этим понятием).

Как правило, объекты обучения не являются равноценными. Среди них могут находиться типичные представители классов – эталоны. Если классифицируемый объект близок к эталону, то, скорее всего, он принадлежит тому же классу. Еще одна категория объектов – неинформативные, или периферийные. Они плотно окружены другими объектами того же класса. Если их удалить из выборки, это практически не отразится на качестве обучения. Наконец, в выборку может попасть некоторое количество шумовых выбросов – объектов, находящихся в чужом классе. Обычно их удаление только улучшает качество классификации.

Исключение из выборки шумовых и неинформативных объектов дает несколько преимуществ одновременно: повышается качество классификации, сокращается объем хранимых данных и уменьшается время

классификации, затрачиваемое на поиск ближайших эталонов [2].

Перейдем к рассмотрению минимизируемого функционала выборки. Пусть имеется множество объектов X и множество имен классов Y. Задана обучающая выборка пар «объект-ответ»

$$X^m = \{(x_1, y_1),...,(x_m, y_m)\} \in X \times Y.$$

Пусть на множестве объектов задана функция расстояния $\rho(x, x')$. Эта функция должна быть достаточно адекватной моделью сходства объектов. Чем меньше значение этой функции, тем более схожи объекты x, x'.

Для произвольного объекта u расположим объекты обучающей выборки x_i в порядке возрастания расстояний до u:

$$\rho(u, x_{1u}) \le \rho(u, x_{2u}) \le ... \le \rho(u, x_{mu}), \quad (1)$$

где через x_{iu} обозначается элемент обучающей выборки, который является i-м соседом объекта u. Аналогичное обозначение введем и для ответа на i-м соседе – y_{iu}. Каждый объект $u \in X$ порождает свою перенумерацию выборки.

Рассматривается метод ближайшего соседа, который относит классифицируемый объект u к тому классу, которому принадлежит ближайший к u объект обучающей выборки:

$$a(u, X^m) = y_{1u}. \quad (2)$$

Профиль компактности выборки X^m есть функция:

$$R(j, X^m) = \frac{1}{m}\sum_{i=1}^{m}[y_i \ne y_{jx_i}]. \quad (3)$$

Иными словами, профиль компактности $R(j)$ – это доля объектов выборки, для которых j-й сосед лежит в другом классе.

Профиль компактности является формальным выражением гипотезы компактности – предположения о том, что схожие объекты гораздо чаще лежат в одном классе, чем в разных.

Выборка X^L разбивается всевозможными $N = C_L^k$ способами на две непересекающиеся подвыборки:

$$X^L = X_n^m \cup X_n^k,$$

где X_n^m – обучающая подвыборка длины m, X_n^k – контрольная подвыборка длины $k = L - m$, $n = 1, ..., N$ – номер разбиения.

Для каждого разбиения n строится алгоритм $a_n(u, X)_n^m$. Функционал полного скользящего контроля (complete cross-validation, *CCV*) определяется как средняя (по всем разбиениям) ошибка на контроле:

$$CCV(X^L) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{k}\sum_{x_i \in X_n^k}[a_n(x_i, X_n^m) \ne y_i]. \quad (4)$$

Функционал полного скользящего контроля характеризует обобщающую способность метода ближайшего соседа.

Справедлива формула для эффективного вычисления *CCV* через профиль компактности:

$$CCV(X^L) = \sum_{j=1}^{k} R(j, X^L)\,\Gamma(j), \quad (5)$$

где $\Gamma(j) = \frac{C_{L-1-j}^{m-1}}{C_{L-1}^{m}}$.

Комбинаторный множитель $\Gamma(j)$ быстро убывает с ростом j. Для минимизации функционала *CCV* достаточно, чтобы при малых j профиль $R(j, X^L)$ принимал значения, близкие к нулю. Это означает, что близкие объекты должны лежать преимущественно в одном классе. Таким образом, профиль действительно является формальным выражением гипотезы компактности [2, 3].

Предлагается использовать кластерный анализ для разделения значений выходов сети на группы, чтобы свести задачу предобработки данных при моделировании с помощью нейронной сети к задаче предобработки данных при классификации, в которой используется теория профиля компактности.

В качестве характеристики близости выходных значений нейронной сети взято евклидово расстояние между точками $\|y_i - y_j\|$. Для произвольного вектора v с числом элементов n евклидова норма находится следующим образом [4]:

$$\|v\| = \sqrt{\sum_{l=1}^{m}|v_l|^2}. \quad (6)$$

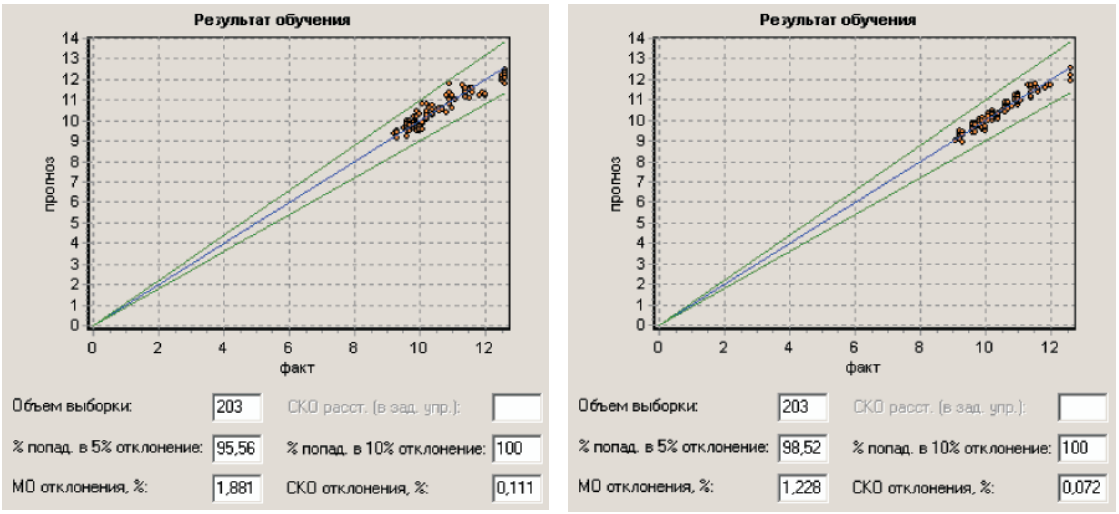
Евклидово расстояние является самой популярной метрикой в кластерном анализе: оно отвечает интуитивным представлениям о близости и, кроме того, очень удачно вписывается своей квадратичной формой в традиционно статистические конструкции. Геометрически оно лучше всего объединяет объекты в шарообразных скоплениях, которые весьма типичны для слабо коррелированных совокупностей [7].

На первом шаге кластерного анализа каждый объект считается отдельным кластером. На следующем шаге объединяются два ближайших объекта, которые образуют новый класс, определяются расстояния от этого класса до всех остальных объектов, и размерность матрицы расстояний сокра-

щается на единицу. Процедура повторяется на текущей матрице расстояний, пока не будет достигнуто некоторое число кластеров [1]. Здесь в работе число кластеров берется равным $\sqrt{n}$, где n – число объектов.

Рассмотрим эффективность методики на примере. Возьмем массив данных и удалим из него некомплектные данные – строки, в которых по какому-либо параметру отсутствуют значения. Затем удалим противоречивые и неинформативные объекты, используя данный метод, и проведем обучение. Сравним результаты этого обучения с обучением по данным того же объема без предобработки (рисунок). Необходимо отметить, что анализ данных выполнялся по отнормированным значениям массива данных, используемом для обучения нейронной сети.

С предобработкой данных сеть обучается лучше: математическое ожидание отклонения, СКО отклонения уменьшаются, вследствие этого увеличивается процент попадания в 5% область отклонения (95,56 и 98,52%). Показатель по 10% области одинаковый – 100%.



а б

Результат обучения нейронной сети:
а – без предобработки данных; б – с предобработкой данных

Таким образом, предлагаемый метод предобработки данных дает более качественное обучение нейронной сети. Предобработка заключается в удалении из массива противоречивых примеров. Поиск таких примеров основан на теории профиля компактности, которая применяется также для предобработки данных, но в задаче классификации. Чтобы использовать известное решение (теорию профиля компактности) в задаче моделирования, необходимо с помощью кластерного анализа выделить группы над значениями выходных параметров нейронной сети.

Список литературы

1. Айвазян С.А. Прикладная статистика: Классификация и снижение размерности: справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков. – М.: Финансы и статистика, 1989. – 607 с.

2. Воронцов К.В. Комбинаторные оценки качества обучения по прецедентам // Докл. РАН. – 2004. – Т. 394, №2. – С. 175–178.

3. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. – 2004. – №13. – С. 5–36.

4. Данилина Н.И. Численные методы / Н.И. Данилина, Н.С. Дубровская, О.П. Кваша. – М.: Высшая школа, 1976. – 368 с.

5. Качановский Ю.П. Выбор архитектуры нейронной сети для моделирования и управления технологическими процессами металлургического производства / Ю.П. Качановский, Е.А. Коротков // Информационно-вычислительные технологии и их приложения: сборник статей X Международной научно-технической конференции. – Пенза: РИО ПГСХА, 2009. – С. 116–119.

6. Коротков Е.А. Анализ возможности применения нейросетевого моделирования в задачах управления качеством металлургического производства // Управление большими системами: сборник трудов V Всероссийской школы-семинара молодых ученых. – Т. 1. – Липецк: ЛГТУ, 2008. – С. 313–320.

7. Мандель И.Д. Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.

8. Царегородцев В.Г. Оптимизация предобработки данных: константа Липшица обучающей выборки и свойства обученных нейронных сетей // Нейрокомпьютеры: разработка, применение. – 2003. – №7. – С. 3–8.

Рецензенты:

Кудинов Ю.И., д.т.н., профессор, зав. кафедрой информатики ГОУ ВПО «Липецкий государственный технический университет» Министерства образования и науки РФ, г. Липецк;

Блюмин С.Л., д.ф.-м.н., профессор, профессор кафедры прикладной математики ГОУ ВПО «Липецкий государственный технический университет» Министерства образования и науки РФ, г. Липецк.

Работа поступила в редакцию 03.08.2011.