УДК 004.8

АНАЛИЗ ВЛИЯНИЯ ПАРАМЕТРОВ КЛАСТЕРОВ НА КАЧЕСТВО РАБОТЫ НЕЙРОННОЙ СЕТИ

Д.С. ОСТАПОВ, В.А. ЧАСТИКОВА

Кубанский государственный технологический университет,
350072, Российская Федерация, г. Краснодар, ул. Московская, 2;
электронная почта:, krasnodar93@mail.ru

В статье проведён анализ и исследование основных параметров, которые влияют на качество кластеризации при работе нейронной сети. Проведен анализ эффективности обучения и точности работы нейронной сети, состоящей из 2 скрытых слоёв. Благодаря разбиению данных на отдельные группы появилась возможность выполнять анализ каждого кластера по отдельности. Нейронная сеть разбита на подсети, которые работают с элементами своего кластера независимо друг от друга. В результате исследования была выявлена зависимость параметров, влияющих на качество кластеризации данных для дальнейшего нейросетевого анализа. В работе проведён анализ статистических данных количества ошибок нейронной сети при разном числе кластеров. Был разработан алгоритм кластеризации с учетом параметров, оказывающих влияние на качество работы и приведена его блок-схема. В результате разбиения нейронной сети на кластеры количество ошибок значительно сократилось.
Ключевые слова: нейронная сеть, кластеризация, подсеть, обучение нейронной сети, обучение с учителем

В статье[1] автором был разработан алгоритм разбиения нейронной сети (НС) на подсети с целью повышения однородности данных, обрабатываемых каждой НС. Процесс обучения происходит на примерах, имеющих более близкую структуру и свойства, и анализирует только элементы группы, на которой она была обучена, поэтому точность работы нейронной сети повышается. Исходя из данных, представленных в [1], присутствует достаточно большое значение среднеквадратического отклонения (СКО) числа ошибок, допускаемых нейронной сетью. Целью этой работы является определение ключевых параметров кластеризации, влияющих на качество функционирования нейронных подсетей (НПС), и усовершенствование алгоритма деления обучающей выборки на кластеры.

Задачей нейронной сети является идентификация, к какому множеству относится подаваемый на её вход пример. В случае её обучения с учителем на примерах множества $I \in \{J_1, J_2, \ldots, J_n\}$ (n – число типов примеров, на которых

обучается НС) при тестировании она сможет распознавать только элементы множества J[4,5]. Число ошибок, которые НС допускает в процессе тестирования зависит от начальных весов нейронной сети[2-6]. Чтобы минимизировать зависимость следует использовать статические начальные веса: для этого необходимо создать их случайным образом и сохранить в отдельном файле, в результате для каждой подсети начальные весовые коэффициенты будут браться из этого файла и будут одинаковыми для каждой НПС.

С заданными статическими весами НС без использования кластеризации осуществляет 1190 ошибок из 283891 испытаний (0,42%).

Данные по числу неверной идентификации входящих примеров при работе нейронной сети с использованием механизма кластеризации приведены в таблице 1.

Таблица 1 – Число ошибок нейронной сети

Число кластеров	Минимальное число ошибок	Среднее число ошибок	Максимальное число ошибок	σ
1	1190	1190	1190	0
2	112	762,06	2030	555,8
3	86	563,58	2058	518,74
4	90	385,35	1585	452,26
5	65	241,07	1114	263,18
6	55	257,68	1034	282,1
7	53	200,55	1025	239,35
8	53	216,30	1282	275,4
9	54	134,55	754	146,48
10	58	150,92	873	159,54
11	58	150,92	873	159,54

σ – СКО.

$$\sigma = \sqrt{DX} = M[X - MX]^2 = \sqrt{\sum_{i=1}^{n}(x_i - MX)^2 p_i}$$ [7,8]

$$MX = \sum_{i=1}^{m} x_i p_i$$

где X – множество числа ошибок нейронной сети.

$X = \{x_1, x_2, \ldots, x_n\}$ – множество числа ошибок нейронной сети

x_i – число ошибок нейронной сети,

DX – дисперсия дискретной величины x,

MX – математическое отклонение

p_i – вероятность получения нейронной сетью числа ошибок x_i

Как видно из таблицы 1, каждый эксперимент имеет достаточно большое σ и большую разницу между максимальным и минимальным числом ошибок. Это происходит из-за того, что в одних случаях механизм кластеризации работает лучше, а в других хуже.

Алгоритм кластеризации k-средних направлен на уменьшение суммы расстояний от каждого объекта до центра соответствующего кластера. Однако для того чтобы каждая подсеть имела возможность выполнить качественную настройку весовых коэффициентов своих нейронов, этого недостаточно. Эффективность работы НПС зависит от числа примеров, на которых она обучается, и однородности данных. Чем больше число элементов обучающей выборки, тем меньше данная подсеть будет ошибаться; чем более однородные данные она анализирует, тем число неверно идентифицированных элементов будет меньше.

Таким образом, эффективность обучения нейронной сети, состоящей из m НПС определяется формулой $z(x,y) = f(x) \otimes g(y)$, где f(x) – показатель равномерности распределения примеров на подсети, g(y) – показатель их однородности.

Функции f(x) и g(y) определены на интервале $x, y \in (0; +\infty)$.

Чтобы f(x) и g(y) равнозначно влияли на функцию z(x,y), следует использовать j(t), которая будет определена на $t \in (0; +\infty)$; j(t) должна монотонно возрастать на интервале $t \in (0; +\infty)$ и иметь верхнюю и нижнюю

границу. Одной из таких функций является функция j(t) = arctg(t), график которой приведён на рисунке 1.

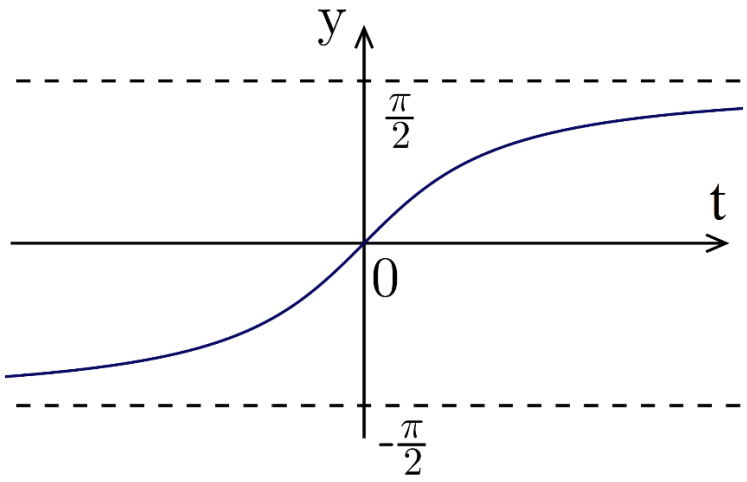

Рисунок 1 – График функции j(t) = arctg(t)

Таким образом, функция z(x,y) примет вид:

$$z(x,y) = arctg\left(f(x)\right) \otimes arct(g(y))$$

Равномерность разбиения обучающего множества на кластеры можно представить в виде следующей функции:

$f(N) = \frac{N}{N + \prod_{i=1}^{r}(1 + N_i)}$, где N – число элементов в обучающей выборке, N_i – число элементов в каждом кластере при обучении нейронной подсети, r – число кластеров. Исходя из формулы, видно, что 0 < f(N) < 1.

Каждый обучающий пример представлен характеристикой Y = {y_1, y_2, …, y_m}, где m – размерность пространства. Значения среднеквадратических отклонений по каждой характеристике обратно пропорциональны показателю однородности примеров внутри кластера. Вектор среднеквадратических отклонений кластера i выглядит следующим образом: $\sigma_i = \{\sigma_{i_1}, \sigma_{i_2}, \ldots, \sigma_{i_m}\}$. σ_{i_j} прямо пропорциональны показателю однородности кластера i. Вариант кластеризации, при выполнении которого результат z(x,y) оказался минимальным, будет наилучшим.

Значения среднеквадратических отклонений по характеристикам X_j и X_q не являются равноценными в связи с тем, что значения σ_j могут быть во много

раз больше σ_q ($\sigma_i j \gg \sigma_i q$). По этой причине при проведении K испытаний необходимо определить максимальные значения каждой характеристики из всех кластеров и всех испытаний.

Множество максимальных характеристик выглядит следующим образом:

$$\sigma_{max} = \{\sigma_{max_1}, \sigma_{max_2}, \ldots, \sigma_{max_m}\}$$

Множество нормализованных среднеквадратических отклонений характеристик кластера i выглядит следующим образом:

$\sigma_{normtitze_i} = \{\sigma_{normtitze_{i_1}}, \sigma_{normtitze_{i_2}}, \ldots, \sigma_{normtitze_{i_m}}\}$, где $\sigma_{normtitze_{i_j}}$ - нормализованное среднеквадратическое отклонение кластера i характеристики j.

$$\sigma_{normtitze_{i_j}} = \begin{cases} 0, \text{если } \sigma_{max_j} = 0 \\ \dfrac{\sigma_{i_j}}{\sigma_{max_j}}, \text{если } \sigma_{max_j} \neq 0 \end{cases}$$

Таким образом, функция, определяющая однородность кластеризации примет вид:

$$g = \sum_{i=1}^{r} \sum_{j=1}^{m} \sigma_{normtitze_{i_j}}$$

$$z = arctg\left(\frac{N}{N + \prod_{i=1}^{r}(1 + N_i)}\right) \cdot arctg\left(\left(\sum_{i=1}^{r} \sum_{j=1}^{m} \sigma_{normtitze_{i_j}}\right)\right)$$

Чем меньшее значение принимает z, тем с большей вероятностью можно утверждать, что НС, состоящая из r НПС, лучше обучится, и число допускаемых ей ошибок уменьшится.

Каждый кластер i характеризуется величиной $p_i = \{p_{i_1}, p_{i_2}, \ldots, p_{i_n}\}$, где p_{i_j} – процент содержания примеров внутри него, которые принадлежат множеству j.

Очевидно, что $\sum_{j=1}^{n} p_{i_j} * 100\% = 100\%.$, где N_i – количество элементов обучающей выборки в кластере i, – число примеров в нём, которые являются элементами подмножества $j_y \in J$.

Равномерность настройки каждого кластера на примерах, принадлежащим разным обучаемым множествам, можно представить в виде среднеквадратического отклонения по показателю p_i:

$$\sigma_{p_i} = \sqrt{\frac{1}{2} \cdot \sum_{j=1}^{2} \left(\overline{P_i} - p_{i_j}\right)^2}$$, где $\overline{P_i}$ – медиана элементов p_{i_j}. [7,8]

Чем меньшее значение будет принимать σ_{p_i}, тем лучше будет данная подсеть распознавать данные, являющиеся элементами разных подмножеств множеств $j_y \in J$.

Чтобы в формуле функции z составляющие оказывали равнозначное влияние на результат, необходимо выполнить нормализацию показателя σ_{p_i}:

$$\sigma_{normtlize_{p_i}} = \begin{cases} 0, \text{если } \sigma_{max_j} = 0 \\ \dfrac{\sigma_{p_i}}{\sigma_{max_i}}, \text{если } \sigma_{max_i} \neq 0 \end{cases}$$

Функция h характеризует равномерность обучения

$$h = \sum_{i=1}^{r} \sigma_{normtlize_{p_i}}$$

Таким образом, функция z, определяющая эффективность кластеризации, примет вид:

$$z = arctg\left(\frac{N}{N + \prod_{i=1}^{r}(1 + N_i)}\right) \cdot arctg\left(\left(\sum_{i=1}^{r}\sum_{j=1}^{m} \sigma_{normtlize_{i_j}}\right)\right) \cdot arctg\left(\sum_{i=1}^{r} \sigma_{normtlize_{p_i}}\right)$$

Используя определение минимального z, алгоритм кластеризации будет выглядеть:

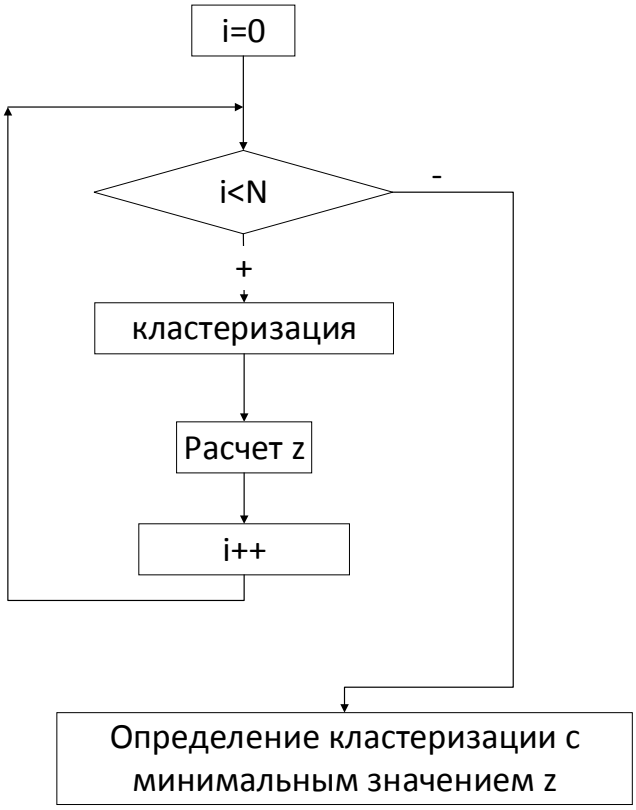

Рисунок 2 – Схема работы алгоритма

Данные о количестве ошибок при числе повторений N = 10 представлены в таблице 2.

Таблица 2 – Данные о количестве ошибок при числе повторений N = 10

Число кластеров	Минимальное число ошибок	Среднее число ошибок	Максимальное число ошибок	σ
2	100	395,9	2068	407,24
3	79	244,46	1294	263,7
4	60	223,34	1059	228,18
5	54	175,96	1210	226,24
6	58	129,04	823	109,36
7	46	127,9	1079	140,23
8	32	114,02	445	56,02
9	85	117,08	284	34,5
10	32	119,38	196	28,38
11	55	120,9	215	33,23

Сокращение количества ошибок после совершенствования кластеризации представлено в таблице 3.

Таблица 3 – Сокращение количества ошибок при кластеризации

Количество кластеров	Уменьшение минимального количества ошибок,%	Уменьшение среднего количества ошибок,%	Уменьшение максимального количества ошибок,%	Умень-шение σ, %
2	10,71	48,05	-1,87	26,73
3	8,14	56,62	37,12	49,17
4	33,33	42,04	33,19	49,55
5	16,92	27,01	-8,62	14,04
6	-5,45	49,92	20,41	61,23
7	13,21	36,23	-5,27	41,41
8	39,62	47,29	65,29	79,66
9	-57,41	12,98	62,33	76,45
10	44,83	20,90	77,55	82,21
11	5,17	19,89	75,37	79,17
Средний итог (среднее арифметическое по колонке)	10,91	36,09	35,55	55,96

Исходя из таблицы 3, видно, что среднее количество ошибок и СКО числа ошибок уменьшилось во всех экспериментах. Итог по всем показателям является положительным, что говорит об успешности применения анализа качества кластеризации на основе функции z. Сокращение среднего числа ошибок составило 36,09%, сокращение СКО составило 55,96%.

ЛИТЕРАТУРА

1. В.А. Частикова Применение методов кластеризации для повышения точности работы нейронных сетей / Частикова В.А., Остапов Д.С. // Современные проблемы науки и образования [Электронный ресурс]. - 2015. - № 57. - Режим доступа: http://www.science-education.ru/121

2. Хайкин С. Нейронные сети. Полный курс. –2-е изд. – М.: Издательский дом «Вильямс», 2006. – 1104 с

3. Г.Э. Яхъяева Нечеткие множества и нейронные сети – М.: Бином, 2008 – 315 с.

4. Н.Г. Ярушкина Основы теории нечётких и гибридных систем. Учебное пособие. - М.: Финансы и статистика, 2004. – 32

5. А.И. Галушкин Нейронные сети: основы теории – М.:Горячая линия-Телеком, 2010, 496 с.

6. В.Е. Гмурман Теория вероятностей и математическая статистика. - М.– ИД Юрайт, 2012, 479 с.

7. Г.А. Соколов Математическая статистика. – М. – Экзамен.,2007,432 с.

REFERENCES

1. V.A. Chastikova Primenenie metodov klasterizacii dlya povysheniya tochnosti raboty nejronnyh setej / Chastikova V.A., Ostapov D.S. // Sovremennye problemy nauki i obrazovaniya [EHlektronnyj resurs]. - 2015. - № 57. - Rezhim dostupa: http://www.science-education.ru/121

2. Hajkin S. Nejronnye seti. Polnyj kurs. –2-e izd. – M.: Izdatel'skij dom «Vil'yams», 2006. – 1104 p

3. G.E. Yahyaeva Nechetkie mnozhestva i nejronnye seti – M.: Binom, 2008 – 315 p.

4. N.G. Yarushkina Osnovy teorii nechyotkih i gibridnyh sistem. Uchebnoe posobie. - M.: Finansy i statistika, 2004. – 32

5. A.I. Galushkin Nejronnye seti: osnovy teorii – M.:Goryachaya liniya-Telekom, 2010, 496 p.

6. V.E. Gmurman Teoriya veroyatnostej i matematicheskaya statistika. - M.– ID YUrajt, 2012, 479 p.

7. G.A. Sokolov Matematicheskaya statistika. – M. – EHkzamen., 2007, 432 p.

ANALYSIS OF INFLUENCE CLUSTER PARAMETERS ON THE PERFORMANCE OF NEURAL NETWORK

D.S. OSTAPOV, V.A. CHASTIKOVA

Kuban State Technological University,
2, Moskovskaya st., Krasnodar, Russian Federation, 350072;
e-mail: krasnodar93@mail.ru

The article provides an analysis and study of the main parameters that affect the quality of the clustering neural network. The analysis of the effectiveness of training and the accuracy of the

neural network consisting of two hidden layers. Due to partition the data into separate groups the opportunity to perform an analysis of each cluster separately. A neural network is divided into subnetworks that operate with elements of its cluster independently. The study has revealed the dependence of the parameters affecting the quality of the data clustering neural network for further analysis. The work carried out analysis of statistical data errors of the neural network with different numbers of clusters. Clustering algorithm has been developed taking into account the parameters that influence the quality of work and shows its block diagram. As a result of the decomposition of the neural network into clusters of errors decreased significantly.

Key words: neural network, clustering, subnetworks, study of neural network.