УДК 004.8

ПРИМЕНЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ ДЛЯ ПОВЫШЕНИЯ ТОЧНОСТИ РАБОТЫ НЕЙРОННЫХ СЕТЕЙ

Частикова В.А., Остапов Д.С.

ФГБОУ ВПО «Кубанский государственный технологический университет», Краснодар, e-mail:chastikova_va@mail.ru

В работе рассмотрены некоторые методы кластеризации для разбиения множества данных на кластеры. В качестве метода кластеризации предложено использовать алгоритм k-means++. 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 случае тестирования на конечном множестве примеров при необходимости происходит переобучение нейронной сети (обратная связь). В работе приведён анализ статистических данных количества ошибок нейронной сети при разном числе кластеров. В результате разбиения нейронной сети на кластеры количество ошибок значительно сократилось; работающие независимо друг от друга подсети позволяют реализовать механизм параллельных вычислений.

Ключевые слова: нейронные сети, кластеризация, k-means, k-means++.

USING CLUSTERING METHODS TO INCREASE NEURAL NETWORKS ACCURACY

Chastikova V.A., Ostapov D.S.

Kuban State Technological University, Krasnodar, e-mail:chastikova_va@mail.ru

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Some clustering methods to share data to clusters were analyzed in this article. There are some clustering methods for partitioning the data set into clusters.As a method proposed clustering algorithm to use k-means ++. The effectiveness of trainingand the accuracy of the neural network consisting of two hidden layers was analyzed. By partition the data into separate groups there is the opportunity to carry out an analysis of each cluster separately. A neural network is divided into subnetworks that work with elements of its cluster independently. In the case of testing on a finite set of examples, if necessary, retraining occurs neural network (feedback).The statistical errors of the neural network with different numbers of clusters were analyzed in this article.As a result of the decomposition of the neural network into clusters the number of errors decreased significantly; working independently of one another subnet can realize the mechanism of parallel computing.

Keywords: neural networks, clustering, k-means, k-means++.

Искусственная нейронная сеть (ИНС) – математическая модель, моделирующая работу нейронов головного мозга человека. Одной из основных задач ИНС является анализ информации и определение свойств поступающих новых данных на основе проведённого анализа. В связи с высокой скоростью развития информационных технологий объём и разнородность анализируемых данных постоянно возрастает, в результате чего ИНС работает с большей погрешностью [4,5]. Основной задачей, решаемой в ходе данного исследования, является повышение однородности данных, обрабатываемых искусственной нейронной сетью. Одним из способов решения данной проблемы является кластеризация.

Кластерный анализ – методика разделения множества данных на подмножества таким образом, чтобы каждый кластер состоял из «схожих» объектов, а объекты разных кластеров

имели сильные отличия [1]. Методы кластеризации имеют наибольшую актуальность в следующих случаях:

- обрабатываемые данные имеют большой объём;
- отсутствует заранее классифицированная обучающая выборка;
- нет предварительной информации об описании, границах и количестве классов обрабатываемых данных [2].

Целью данного исследования является снижение числа ошибок при работе нейронной сети.

Задачи:

1. Выбор архитектуры нейронной сети.
2. Повышение однородности анализируемых искусственной нейронной сетью данных.
3. Выбор метода кластеризации.
4. Разбиение нейронной сети на подсети.

Материалы и методы исследования

Кластер, один из основных терминов кластерного анализа, – группа из элементов со схожими свойствами. Слово «кластер» образовано от английского “cluster”, что означает пучок, группа, скопление, гроздь, куст. Не менее важное понятие центроид – центр кластера. Впервые задача кластеризации была поставлена в 1930-х годах [3].

Задача кластеризации заключается в разбиении множества обучающих примеров K на заранее заданное непересекающееся число подмножеств K_i таким образом, что

$K = \{K_1, K_2, K_3, \ldots, K_n\}$

$K_i \cap K_j = 0, i \neq j,$

где $\{K_1, K_2, K_3, \ldots, K_n\}$ – кластеры множества K.

Элементы кластера K_i имеют схожие между собой свойства и несхожие свойства по отношению к элементам кластера K_j. Однако в ряде случаев используется менее жесткая методика кластеризации с использованием нечетких множеств. В данном случае кластеры характеризуются функциями членства, которые определяют степень принадлежности каждого элемента соответствующему кластеру [2].

Для разбиения большого объёма данных на кластеры используются неиерархические методы кластеризации. Одними из наиболее популярных являются алгоритмы k-meansи k-means++.

Метод k-средних (k-means) заключается в минимизации суммарного квадратического отклонения точек кластеров от их центров. Данный метод состоит из следующих этапов:

1. Число кластеров, на которые будет разбито множество элементов K, заранее известно.

2. Случайным образом определяются центры кластеров.

3. Вычисляется $V = \sum_{i=1}^{k} \sum_{x_i \in S_i} (x_j - \mu_i)^2$, где k – число кластеров, S_i – полученные кластеры, i = 1,2,3,4,…,k, μ_i – центры кластеров.

4. Вычисляется центр масс μ для каждого кластера

$\mu = \{\mu_1, \mu_2, \dots, \mu_i, \dots, \mu_n\} = \{\frac{1}{S_1}\sum_{m=1}^{S_1} x_1, \frac{1}{S_2}\sum_{m=1}^{S_2} x_2, \dots, \frac{1}{S_1}\sum_{m=1}^{S_i} x_i, \dots, \frac{1}{S_1}\sum_{m=1}^{S_n} x_n\}$, где

S_i – число элементов в кластере i.

5. Если центр масс изменился, то происходит переход к шагу 2, и продолжается работа алгоритма. Если не изменился, итерации завершаются.

Одной из модификаций алгоритма k-menas является алгоритм k-means++. Данный алгоритм был предложен в 2007 году Дэвидом Артуром и Сергеем Вассильвитским [6]. K-means++ является модификацией алгоритма k-means и направлен на поиск оптимальных начальных центров кластеров. Алгоритм состоит из следующих шагов:

1. Выбрать случайным образом центроид из всех имеющихся объектов множества K.

2. Для каждой точки определить значение квадрата расстояния до ближайшего центроида $D(x)^2$.

3. Для каждой точки определить новый центр. Суммирование $D(x)^2$ необходимо выполнять до тех пор, пока сумма не превысит $Rand(0;1)\cdot D(x)^2$. Как только это произойдет, суммирование можно остановить и определить текущую точку в качестве центроида. При выборе каждого следующего центроида нужно контролировать, чтобы он не совпал с одним из уже выбранных в качестве центроидов элементов.

Одним из критериев оценки эффективности кластеризации является сумма расстояний до центров кластеров. Для осуществления сравнения качества работы алгоритмов k-means и k-means++ был разработан программный комплекс; на его основе проведены исследования, результаты которых представлены в таблице 1, где V – сумма расстояний от точек до центров кластеров.

Из табл.1 видно, что алгоритм k-means++ лучше справляется с задачей кластеризации, чем k-means, поэтому для повышения точности работы нейронной сети предлагается использовать метод k-means++.

Таблица 1

Сравнение алгоритмов k-means и k-means++

Алгоритм кластеризации	Число кластеров	V	Прирост точности (по сравнению с k-means), %
k-means	10	44713577494947,2	-
k-means++	10	29870985390182,4	33.19

k-means	8	44856741214940,31	-
k-means++	8	30951151438308,81	31
k-means	4	44999504532950,7	-
k-means++	4	29422271430446,5	34

В качестве практической задачи для исследования эффективности работы кластеризованной нейронной сети была выбрана задача анализа сетевого трафика с целью определения наличия DDoS-атак. Для обучения была использована база [7], состоящая из 141946 обучающих примеров, 78,59 % из которых являются примерами DDoS-атак. В качестве базы для тестирования применялась база [7], состоящая из 283891 тестовых примеров, 78,66 % из которых являются атаками. Обучение происходит с использованием метода обратного распространения ошибки.

Так как нейронная сеть обрабатывает примеры разной структуры с различными свойствами, подобрать весовые коэффициенты каждого нейрона таким образом, чтобы анализ разнородных примеров происходил с большой точностью, достаточно сложно. Именно поэтому предлагается использовать кластеризованную нейронную сеть. Структура алгоритма обучения нейронной сети представлена на рис. 1.

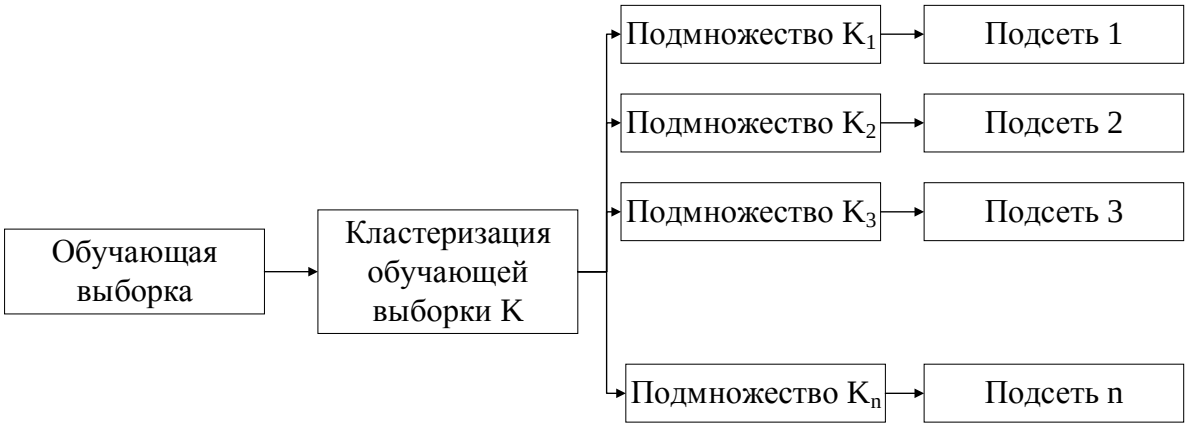

Рис. 1. Структура алгоритма обучения кластеризованной нейронной сети

Алгоритм обучения кластеризованной нейронной сети выглядит следующим образом:

1. Происходит разбиение обучающей выборки K на кластеры K_1, K_2, …, K_n таким образом, что $K = \{K_1, K_2, K_3, \ldots, K_n\}$, $K_i \cap K_j = 0, i \neq j$. Нейронная сеть представляет собой множество подсетей $N = \{N_1, N_2, \ldots, N_n\}$.

2. Элемент K_{ij} подается на нейронную сеть N_i, соответствующую кластеру K_i (элементы каждого кластера обрабатываются только на соответствующей им подсети).

Алгоритм анализа примеров при тестировании нейронной сети представлен на рис.2.

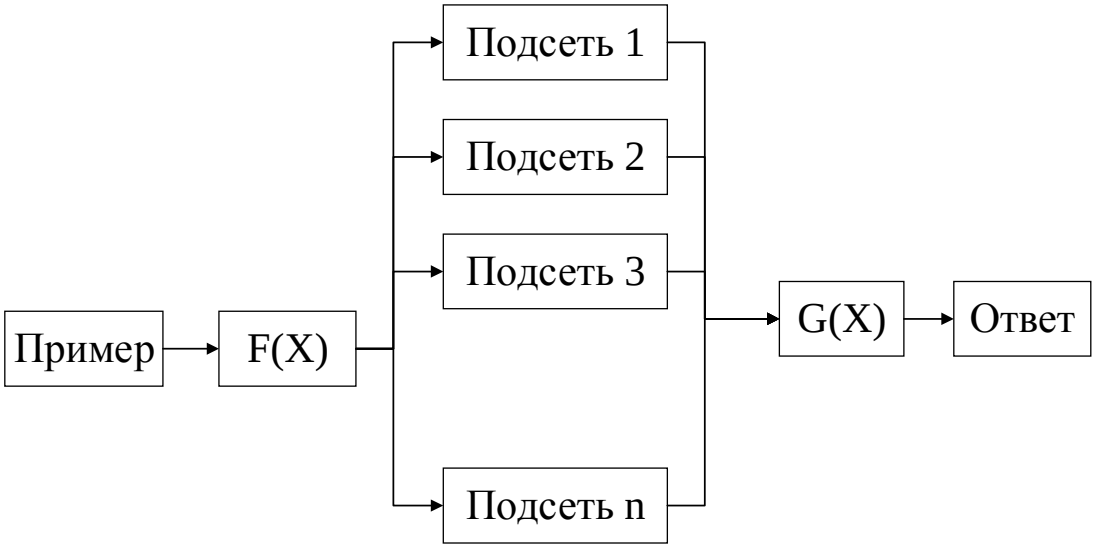

Рис. 2. Алгоритм анализа примеров при тестировании нейронной сети

Алгоритм состоит из следующих этапов:

1. Подаётся пример на нейронную сеть.
2. Функцией F(X) определяется, какому кластеру принадлежит данный пример и какая подсеть должна его обрабатывать.
3. Подсеть, обрабатывавшая пример, передаёт свой ответ функции G(X).

Выводы

Если множество тестовых примеров K` является конечным, можно определить, какая часть элементов принадлежит каждому кластеру.

$P_i = K_i/K \cdot 100\ \%$ – принадлежность элементов кластеру K_i

$P`_i = K`_i/K` \cdot 100\ \%$ – принадлежность элементов кластеру $K`_i$

Если $P`_i >> P_i$, необходимо произвести процесс обучения нейронной сети сначала. Подобное переобучение позволит избежать ошибок, связанных с недостаточным обучением подсетей. Например, если нейронную сеть обучать на 100 000 примеров, а на подсеть подать только 10 примеров для обучения (0.01 % от общего числа примеров), подсеть может недостаточно обучиться. Если же нейронную сеть тестировать на 200 000 примеров, а на подсеть подать 10000 (5 % от общего числа примеров), может возникнуть ситуация, когда данная подсеть будет часто ошибаться.

Среднее число ошибок в зависимости от числа кластеров представлено в табл. 2.

Таблица 2

Среднее число ошибок нейронной сети

Число кластеров	Среднее число ошибок	% ошибок нейронной сети
Без разбиения	1856	0,654
2	1314	0,463

3	1233	0,434
4	923	0,325
5	741	0,261
6	511	0,180
7	431	0,152
8	403	0,142
9	**297**	**0,105**
10	411	0,145
11	343	0,121
12	497	0,175

Как видно из табл. 2, для рассматриваемой задачи оптимальным является разбиение базы примеров на 9 кластеров, в результате чего число ошибок снижается более, чем в 6 раз.

К преимуществам кластеризованной нейронной сети с использованием алгоритма k-means++ можно отнести:

- повышение точности работы нейронной сети;
- работающие независимо друг от друга подсети позволяют реализовать механизм параллельных вычислений.

Список литературы

1. Бирюков А.С. Методы построения коллективных решений кластерного анализа: дис. … канд. техн. наук. – М., 2005. – С. 3-5.
2. Бояркин М.И. Синтез информационной системы группировки многомерных данных с использованием кластерного анализа: дис. … канд. техн. наук. – М., 2008. – С. 12-33.
3. Киреев В.С. Методы двухэтапной и многокритериальной кластеризации данных выборок больших объёмов: дис. … канд. техн. наук. – Самара, 2008. – С. 11-18.
4. Малыхина М.П., Бегман Ю.В. Гибридные нейроэкспертные системы в образовании // Материалы XIV Всероссийской научно-практической конференции «Инновационные процессы в высшей школе», 2008. – С. 193-194.
5. Частиков А.П., Тотухов К.Е., Урвачев П.М. Теоретические основы интеллектуальной диагностики виртуального робота // Современные проблемы науки и образования, 2013. – № 1; URL: www.science-education.ru/107-8310.
6. K-means++: The Advantages of Careful Seeding. David Arthur and Sergei Vassilvitskii.
7. База данных университета MIT http://kdd.ics.uci.edu/databases/kddcup99/.

Рецензенты:

Ключко В.И., д.т.н., профессор, профессор кафедры ИСП Кубанского государственного технологического университета, г. Краснодар;

Пиотровский Д.Л., д.т.н., профессор, зав. кафедрой АПП Кубанского государственного технологическогоуниверситета, г. Краснодар.