ТЕХНИЧЕСКИЕ НАУКИ

УДК 004.8; 519.7
DOI: 10.35330/1991-6639-2021-1-99-15-19
MSC: 68T07

ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ ПРИ КЛАСТЕРИЗАЦИИ ДАННЫХ

Р.А. ЖИЛОВ

Институт прикладной математики и автоматизации –
филиал ФГБНУ «Федеральный научный центр
«Кабардино-Балкарский научный центр Российской академии наук»
360000, КБР, г. Нальчик, ул. Шортанова, 89 А
E-mail: ipma@niipma.ru

В работе рассматривается использование нейронных сетей различной структуры для решения задачи разделения объектов по схожим признакам на заранее неизвестное количество классов. Кластерный анализ применяется для анализа больших объемов данных на наличие скрытых закономерностей и для визуализации структуры данных для большей наглядности. Использование искусственных нейронных сетей для решения задачи кластеризации является обоснованным в силу их специфики работы. В зависимости от типа данных, принадлежащих кластеризации, и от специфики задачи уместно использование различных структур нейронных сетей.

Ключевые слова: кластеризация данных, нейронные сети, сигма-пи нейрон, кластерный анализ, сеть Кохонена, многослойный перцептрон, обучающая выборка.

Поступила в редакцию 09.02.2021 г.

Для цитирования. Жилов Р.А. Применение нейронных сетей при кластеризации данных // Известия Кабардино-Балкарского научного центра РАН. 2021. № 1(99). С. 15-19.

Введение

Кластеризация (или кластерный анализ) – это задача объединения схожих объектов в группы по одинаковым или схожим признакам. Эти группы в дальнейшем называются кластерами. Внутри каждой группы оказываются объекты со схожими признаками. Кластеризация от классификации отличается тем, что при кластеризации заранее не известно количество групп, в которые нужно объединить объекты, и данное количество определяется в процессе кластеризации. Кластеризация является одной из самых важных задач в области анализа больших массивов данных на наличие скрытых и неоднозначных закономерностей. Список областей применения достаточно широк, начиная от разбиения изображений, анализа текстов и заканчивая маркетингом.

Кластерный анализ

В анализе информации важное место занимает выявление общих принципов формирования однородных или похожих данных. В соответствии с этими принципами все данные можно представить в виде различных групп. Дальше на основании исследования части данных из одной группы можно делать выводы о группе данных в целом. Такой процесс анализа данных называется кластерным анализом данных. Естественно, для облегчения кластерного анализа необходима автоматизация данного процесса. Применение кластерного анализа в общем виде сводится к следующим этапам [1]:

1. Отбор выборки объектов для кластеризации.
2. Определение множества переменных, по которым будут оцениваться объекты в выборке.
3. Вычисление значения меры сходства между объектами.
4. Применение метода кластерного анализа для создания групп сходных объектов.
5. Представление результатов анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

Для того чтобы объединить объекты в отдельный кластер по схожим признакам, необходимо определить, насколько же они схожи. Для решения данной проблемы вводится понятие меры близости или функции расстояния. Существует несколько мер близости, но самыми основными являются евклидово расстояние, квадрат евклидова расстояния, манхэттенское расстояние, расстояние Чебышева и расстояние Минковского. Данные расстояния применяют, когда объекты можно представить как точки в n-мерном пространстве. В случаях, когда такое представление не является возможным, применяется метрика I – коэффициент корреляции Пирсона.

Существует достаточно много алгоритмов кластерного анализа, они делятся на иерархические и неиерархические.

Иерархические алгоритмы – наиболее распространенные алгоритмы, использующиеся при кластеризации данных. Они делятся на алгомеративные и итеративные по принципу действия. При использовании алгомеративных алгоритмов на начальном этапе все объекты определяются в отдельные кластеры, и в процессе работы алгоритма наиболее схожие по признакам кластеры объединяются в один. А при использовании иерархических алгоритмов на начальном этапе все объекты объединены в один кластер, и в процессе работы происходит разделение по их отдаленности по признаку. При использовании иерархических алгоритмов на каждом шаге приходится работать с матрицей расстояния, что при больших количествах объектов, нуждающихся в кластеризации, приводит к большой затрате вычислительного времени и ресурсов.

В методе древовидной кластеризации предусмотрены различные правила иерархического объединения в кластеры [2]:

1. Правило одиночной связи. На начальном этапе происходит объединение между двумя самыми близкими объектами, т.е. теми объектами, которые максимально похожи по определенному признаку. В дальнейшем на каждом этапе к ним присоединяется объект, максимально похожий по признаку с одним из объектов кластера, т.е. для его включения в кластер требуется, чтобы данный объект был похож на один из первых двух кластеров. Другим названием данного метода является метод ближайшего соседа, потому что дистанция от одного кластера до другого вычисляется как дистанция между двумя наиболее близкими объектами в различных кластерах. Это правило «нанизывает» объекты для формирования кластеров. Самым большим недостатком такого метода является то, что получившиеся кластеры имеют вытянутую форму.

2. Правило полных связей. Данный метод помогает убрать основной недостаток предыдущего метода. Важным дополнением является то, что объекты, принадлежащие одному кластеру, имеют коэффициент сходства, который больше некоторого порогового значения S. В терминах евклидова расстояния можно сказать, что расстояние между двумя точками (объектами) кластера не должно превышать некоторого порогового значения d. Таким образом, d определяет максимально допустимый диаметр под-

множества, образующего кластер. Этот метод называют еще методом наиболее удаленных соседей, так как при достаточно большом пороговом значении d расстояние между кластерами определяется наибольшим расстоянием между любыми двумя объектами в различных кластерах.

3. Правило невзвешенного попарного среднего. Близость двух кластеров находим как среднее расстояние между всеми парами объектов в них. Данный метод хорошо работает, когда объекты в действительности формируют различные группы, однако он работает одинаково хорошо и в случаях протяженных (цепочного типа) кластеров.

4. Правило взвешенного попарного среднего. Данный метод отличается от предыдущего тем, что в качестве коэффициента при вычислении используется размер соответствующих кластеров. Данный метод хорошо работает, когда размеры кластеров, в которые нужно объединить объекты, имеют неравные размеры.

5. Невзвешенный центройдный метод. В данном методе дистанция между кластерами вычисляется как дистанция между их центрами тяжести.

6. Взвешенный центройдный метод. В данном методе к предыдущему методу добавляется весовой коэффициент, который отражает разницу между размерами кластеров. Исходя из этого данный метод более предпочтителен, когда есть большое отличие между размерами кластеров.

7. Правило Уорда. В данном методе в качестве целевой функции применяется сумма квадратов отклонений внутри группы, которая является суммой квадратов расстояний между каждым объектом и средним значением по кластеру, которая содержит в себе данный объект. На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов отклонений. Данный метод объединяет близко расположенные по признаку кластеры. Получающиеся методом Уорда кластеры представляют собой гиперсферы, примерно равные по размеру.

Применение нейронной сети для решения задачи кластеризации

Для того чтобы классифицировать данные, могут быть использованы такие структуры нейронных сетей, как многослойный перцептрон, сети с самоорганизацией на основе конкуренции и сети с сигма-пи структурой. При использовании сетей с самоорганизацией применяются алгоритмы обучения без учителя и не требуется аппаратное задание классов, что позволяет подобрать необходимое количество в процессе их обучения и функционирования [3]. Контролируемое обучение возможно не только на начальном этапе при подготовке к работе, но и непосредственно в процессе работы. В процессе работы, когда появляются новые кластеры, самоорганизующимся сетям не требуется переобучение, и они обладают способностью адаптироваться.

Многослойный перцептрон. На входы нейронной сети подается входной вектор признаков; в соответствии с алгоритмом обратного распространения ошибки происходит корректировка весов согласно заранее предопределенному классу входного вектора. После обучения на определенном количестве входных образов происходит проверка сети на образах, не участвовавших в обучении.

Сеть Кохонена. На входы нейронной сети подаются значения признаков выбранного объекта. Нейросеть обрабатывает эти сигналы, после чего в выходном слое определяется нейрон-победитель. Нейрон-победитель выходного слоя определяет класс объекта, признаки которого были поданы на входы нейросети. Так как каждому классу в процессе

обучения сети был присвоен классификационный код, то при подаче на входы нейронной сети вектора признаков неизвестного объекта сеть способна определить его код. Если нейрон-победитель не определяет класс объекта, то для него создается новый класс.

$\sum\prod$-*нейрон*. Один логико-арифметический $\sum\prod$-нейрон позволяет эффективно разделять конечные множества $X \sqsubseteq \{0,1\}^n$ на два класса [3]. Для разделения на q непересекающихся классов $(q > 2)$ строили сеть из q нейронов. Однако это возможно далеко не всегда. Модель $\sum\prod$-нейромодуля с конкурирующим функционированием по правилу WTA

$$y_i = h\big(s_i - \max\{s_1, \dots, s_q\}\big),$$

$$s_i = sp_i(x_1, \dots, x_n)$$

позволяет разделить любое подмножество X на q непересекающихся классов. $\sum\prod$-нейромодуль с конкурирующим функционированием по правилу WTA обучается так же легко, как и единственный логико-арифметический $\sum\prod$-нейрон.

Для обучения $\sum\prod$-нейрона требуется предварительное упорядочение обучающей выборки, но обучение такого вида нейронная сеть проходит за один проход обучающей выборки, что существенно снижает затраты как вычислительной мощности, так и машинного времени.

Заключение

Таким образом, становится однозначно понятно, что использование искусственных нейронных сетей для решения задачи разделения объектов на заранее не определенное количество непересекающихся классов по схожим признакам является обоснованным. Такой подход к кластеризации особенно необходим при работе с большими объемами данных, требующими больших затрат вычислительной мощности и машинного времени. Нейронные сети также актуальны при поиске скрытых закономерностей в больших таблицах данных и являются инструментом, который совершенствуется быстрыми темпами.

ЛИТЕРАТУРА

1. *Мандель И.Д.* Кластерный анализ. М: Финансы и статистика, 1988. 176 с.

2. *Dimitrichenko D.* Algorithm for constructing logical neural networks based on logical various-valued functions // Advances in Intelligent Systems and Computing (AISC), 2020. Vol. 1310. Pp. 91-96.

3. *Шибзухов З.М.* Конструктивные методы обучения нейронных сетей. М: Наука, 2006. 159 с.

REFERENCES

1. Mandel I. D. *Klasternyy analiz* [Cluster analysis]. M: Finance and statistics, 1988, 176 p.

2. Dimitrichenko D. Algorithm for constructing logical neural networks based on logical various-valued functions // Advances in Intelligent Systems and Computing (AISC), 2020, vol. 1310. Pp. 91-96.

3. Shibzukhov Z.M. *Konstruktivnyye metody obucheniya neyronnykh setey* [Constructive methods for training neural networks]. M: Nauka, 2006, 159 p.

APPLICATION OF NEURAL NETWORKS FOR DATA CLUSTERING

R.A. ZHILOV

Institute of Applied Mathematics and Automation –
branch of the FSBSE «Federal Scientific Center
«Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences»
360000, KBR, Nalchik, 89 A Shortanov str.
E-mail: ipma@niipma.ru

In this paper, we consider the use of neural networks of various structures to solve the problem of dividing objects according to similar features into an unknown number of classes in advance. Cluster analysis is used to analyze large amounts of data for hidden patterns and to visualize the data structure for greater clarity. The use of artificial neural networks to solve the clustering problem is justified due to their specifics of work. Depending on the type of data belonging to clustering and on the specifics of the task, it is appropriate to use various structures of neural networks.

Keywords: data clustering, neural networks, sigma-pi neural networks, cluster analysis, Kohonen networks, multilayer perceptron, training set.

Received by the editors 09.02.2021 г.

For citation. Zhilov R.A. Application of neural networks for data clustering // News of the Kabardino-Balkarian Scientific Center of RAS. 2021. No. 1 (99). Pp. 15-19.

Сведения об авторе:

Жилов Руслан Альбердович, стажер-исследователь отдела нейроинформатики и машинного обучения Института прикладной математики и автоматизации – филиала Кабардино-Балкарского научного центра РАН.
360000, КБР, г. Нальчик, ул. Шортанова, 89 А.
E-mail: zhilov91@gmail.com

Information about the author:

Zhilov Ruslan Alberdovich, Trainee Researcher, Neuroinformatics and Machine Learning Department, Institute of Applied Mathematics and Automation – branch of the Federal State Budget Scientific Establishment "Federal Scientific Center "Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences".
360000, KBR, Nalchik, 89 A, Shortanov street.
E-mail: zhilov91@gmail.com