Кластеризация данных методом растущего нейронного газа

А.В. Чернов[1], *М.А. Бутакова*[1], *П.С. Шевчук*[1,2]

[1]*Ростовский государственный университет путей сообщения, Ростов-на-Дону*
[2]*Российская таможенная академия, Ростовский филиал, Ростов-на-Дону*

Аннотация: В статье рассматриваются задачи, возникающие при распознавании образов, связанные с кластеризацией и абстракцией данных. Детализированы типовые варианты кластеризации данных. Приведена задача преобразования данных методом векторного квантования с наименьшей ошибкой. Описана система конкурентного обучения искусственной нейронной сети на основе растущего нейронного газа. С использованием метода растущего нейронного газа предложен улучшенный алгоритм самообучающейся искусственной нейронной сети конкурентного обучения. Определены критерии завершения кластеризации с использованием критерия адаптации в качестве критерия остановки. Приведены примеры кластеризации данных искусственной нейронной сетью методом растущего нейронного газа.
Ключевые слова: кластеризация, искусственная нейронная сеть, компьютерное моделирование, распознавание образов, самообучающиеся интеллектуальные системы

Распознавание образов – это уже давно стандартная задача определения объекта или нахождения каких-либо его отличительных свойств по его графическому изображению, аудиозаписи и/или другим характеристикам. Распознавание образов напрямую использует искусственный интеллект и машинное обучение, анализ данных и поиск отличительных характеристик в «сырых» данных [1]. Анализ данных состоит из нескольких основных этапов [2]: подготовки входных данных, выделения значимых признаков, очистки входных данных от потенциально лишней информации, применения методов обнаружения в исходных данных полезных и доступных к дальнейшему использованию знаний, а также постобработки данных и интерпретации полученных результатов. Один из вариантов распознавания образов в информационных системах – это кластеризация входных данных [3].

Задача кластеризации данных

Кластеризация – это классификация моделей (наблюдений, элементов данных или признаков) на группы, называемые кластерами. Каждый кластер

формируется на основе некоторой сходной характеристики данных, сортируя данные таким образом, что похожие объекты оказываются в одном кластере данных. Задачи классификации [4] разрозненной информации постоянно возникают во множестве повседневных задач, что доказывает её полезность как одного из этапов анализа данных. Однако кластеризация [5,6] является сложной проблемой со стороны комбинаторики, и различия в допущениях и контекстах данных замедлили развитие общих концепций и методологий анализа данных.

Проблема анализа данных концептуально может быть разделена на аналитический и подтверждающий анализ (формирование гипотезы и принятие решений). Основной особенностью обоих видов анализа является классификация характеристик входных данных либо на основе соответствия каких-либо характеристик предполагаемой модели данных, либо с помощью группировки этих данных на основе характеристик, выявленных в результате анализа (кластеризация данных). Кластеризация – это типизация похожих объектов на разные группы по какому-либо признаку – разбиение набора входных данных на некоторые подмножества (кластеры), так, что данные в каждом подмножестве имеют какую-либо общую характеристику.

Кластерный анализ данных – это организация набора входных данных в кластеры на основе близкого, либо полностью идентичного признака. Таким образом, итоговые данные в рассматриваемом кластере похожи между собой намного сильнее, чем на данные из другого кластера. Пример кластеризации входных данных показан на рис. 1: на рис. 1А приведено случайное расположение некоторых входных данных, результатом процесса кластеризации данных является рис. 1Б, на котором четко видны сформированные кластеры данных – данные, принадлежащие одному общему кластеру, имеют одинаковую метку, в данном случае цифру [7].

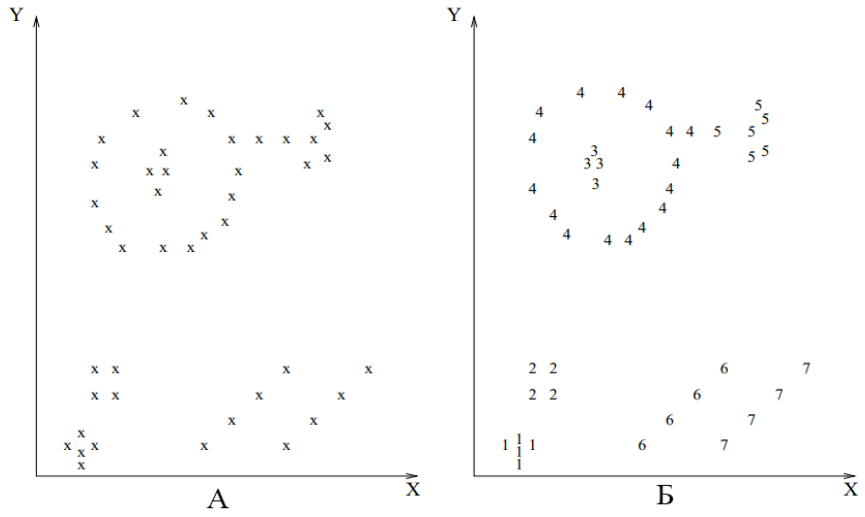

Рис. 1. – Типовой вариант кластеризации данных

В результате процесса абстракции данных возникает проблема обратного соответствия кластеризованных представлений данных от итоговых данных. Для решения данной проблемы применяются оценки достоверности соответствия представлений данных реальным данным. В контексте кластеризации абстракция данных представляет собой компактное описание каждого кластера, обычно в терминах кластерных прототипов или репрезентативных паттернов.

Задачу преобразования данных в представления с наименьшей ошибкой с изначально заданной скоростью, или с наименьшей скоростью при изначально заданном параметре ошибки называют задачей квантования данных. Основной проблемой преобразования с наименьшей ошибкой является итоговая сложность решения – в некоторых случаях требуемый параметр ошибки оказывается достижим лишь в случае полного представления данных. При этом подразумевается не только сложность нахождения хороших квантователей для заданного класса источников, но и сложность непосредственно алгоритма квантования [8].

Квантование, используемое при обработке изображений, представляет собой метод сжатия с потерями, которые возникают в процессе сжатия

диапазона значений используемой цветовой палитры до единого квантового значения. Когда количество дискретных символов в данном потоке уменьшается, поток становится более сжимаемым. Например, уменьшение количества цветов, необходимых для представления цифрового изображения, позволяет уменьшить размер его файла. Независимое квантование каждого значения сигнала или параметра называется скалярным квантованием, в то время как совместное квантование блока параметров называется блочным или векторным квантованием (*VQ*). Ключом к сжатию данных векторного квантования является хорошая кодовая книга. Каждый вектор сигнала, подлежащего сжатию, сравнивается с записями кодовой книги, содержащей репрезентативные векторы. Адрес записи кодовой книги, наиболее похожий на входной вектор сигнала, передается в приемник. В приемнике адрес обращается к записи из идентичной кодовой книги, таким образом восстанавливая сжатое представление данных к исходному сигналу - рис. 2 [8].

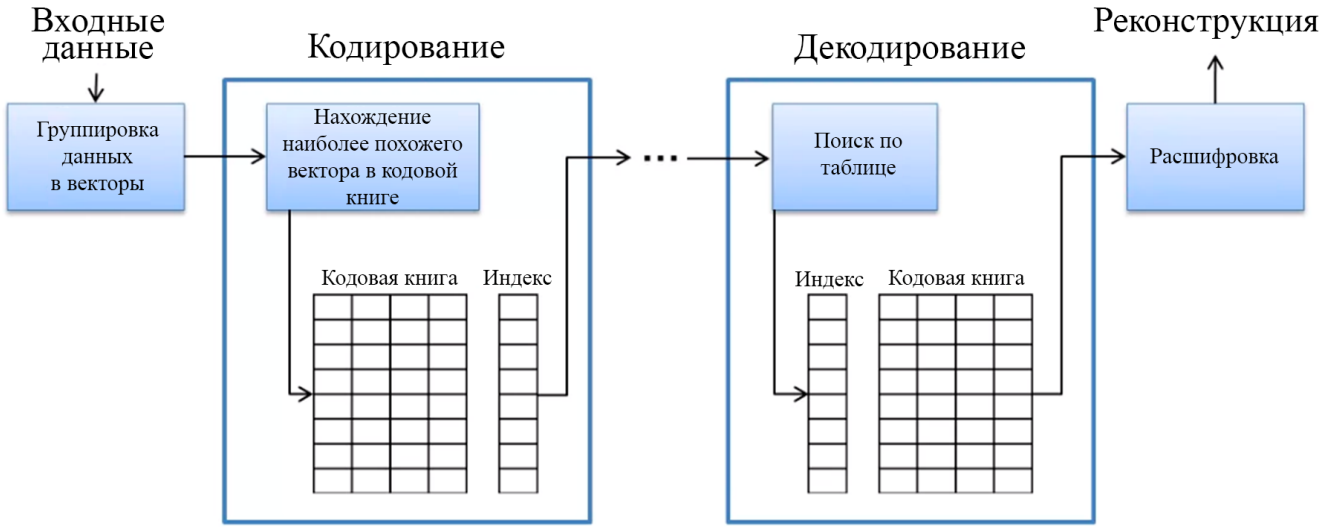

Рис. 2. – Принцип векторного квантования

Для того чтобы сравнивать данные в процессе кластеризации, необходимо иметь некоторый критерий, на основании которого будет происходить сравнение. Обычно таким критерием является расстояние

между объектами на плоскости. При выборе различных метрик оценки расположения данных результаты кластеризации могут существенно отличаться.

Одной из возможных целей абстракции данных искусственной нейронной сети (ИНС), обучаемой без учителя, когда доступны только входные данные, но нет никакой информации о желаемых выходных данных, является уменьшение размерности входных данных: поиск низкоразмерного подпространства входного векторного пространства, содержащего большинство или все входные данные. Линейные подпространства с этим свойством могут быть вычислены непосредственно с помощью анализа главных компонентов или итеративно при помощи известных сетевых моделей нейронных сетей. Самоорганизующиеся карты Кохонена и растущие клеточные структуры Фритцке и Вильке [9] (в том числе и растущий нейронный газ) в свою очередь являются адаптивным алгоритмом, направленным на оценку плотности распределения данных. Таким образом, следует помнить, что в зависимости от отношения между собственной размерностью данных и размерностью целевого пространства, некоторая информация о топологическом расположении входных данных может быть потеряна в процессе, потому что обратимое отображение из многомерных данных в низкоразмерные пространства (или структуры) не существует.

Модифицированный алгоритм растущего нейронного газа

Системы конкурентного обучения ИНС, такие как нейронный газ, предполагают наличие некоторого количество нейронов в исследуемой области R^n и последовательное соединение их между собой топологическими связями на основе оценки входных сигналов, получаемых из базового распределения входных данных $P(\xi)$. Принцип такого подхода: для каждого входного сигнала x соединить два ближайших (в евклидовом

пространстве) нейрона синапсом. Таким образом, только нейроны, лежащие на подмножестве входных данных или в его окрестностях, фактически развивают синапсы между собой. Остальные создающиеся нейроны бесполезны в процессе кластеризации данных, называются мертвыми нейронами и удаляются в процессе работы ИНС. Таким образом, для того, чтобы ИНС использовала созданные нейроны, они должны быть размещены в тех областях R^n, где $P(\xi)$ отличается от нуля, а ξ – входной вектор данных.

Это может быть сделано при помощи любой процедуры векторного квантования. Мартинец и Шультен ещё в 1991 году предложили особый вид метода векторного квантования, названный нейронным газом. Основной принцип нейронного газа заключается в следующем: для каждого входного сигнала x адаптировать k нейронов, в результате чего коэффициент адаптации уменьшается от большого начального значения до небольшого конечного значения. Чем больше значение коэффициента адаптации, тем больше нейронов будет сдвинуто к входному сигналу x. Затем коэффициент адаптации уменьшается до тех пор, пока не будет адаптирован только ближайший нейрон для каждого входного сигнала. Алгоритм уменьшения коэффициента адаптации заранее определяется начальными параметрами для нейронного газа – критерием остановки [10].

Рассмотрим подробнее алгоритм улучшенного нейронного газа.

Основная идея метода заключается в последовательном добавлении новых нейронов в первоначально небольшую ИНС путем оценки локальных статистических показателей, собранных на предыдущих этапах адаптации сети. Процесс обучения ИНС обычно начинается с двух случайных точек, количество которых последовательно увеличивается, вследствие чего нейронный газ расширяется в пространстве.

Полный алгоритм построения и обучения ИНС опишем следующим образом:

1 ИНС имеет два нейрона a и b создаваемых в случайных точках пространства, обозначенных соответственно W_a и W_b в R^n.

2 ИНС получает некоторый входной сигнал ξ в соответствии с $P(\xi)$: выбирается некоторая необходимая группа входных данных (в данном случае набор заранее подготовленных входных данных представляет красная точка на рис. 3).

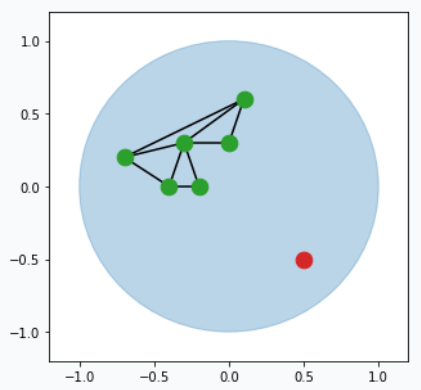

Рис. 3. – Набор входных данных

Синяя область на рис. 3 представляет собой общий набор данных, в рамках которого работает алгоритм нейронного газа. Зеленые точки, связанные между собой черными линиями, представляют сеть растущего нейронного газа, где зеленые точки – это нейроны, а черные линии представляют связи между нейронами (синапсы).

3 ИНС находит ближайшие нейроны к искомой группе входных данных (нейроны, обозначенные на графике синими точками s_1. и s_2.) и соединяет эти нейроны соответствующими связями – синапсами. В случае, если они уже соединены, происходит обновление синапса, рис. 4.

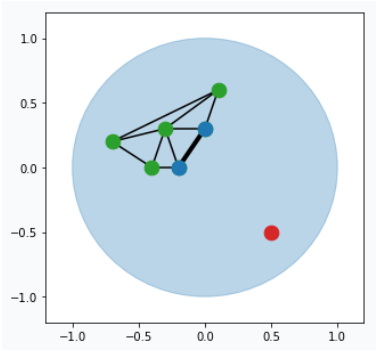

Рис. 4. – Подготовка ближайших нейронов

4 ИНС увеличивает вес всех синапсов, начиная с s_1.

5 Каждый из нейронов ИНС имеет параметр ошибки, накапливаемый в течение времени. Для каждого обновленного или только что созданного нейрона также пересчитывается параметр ошибки, который высчитывается как расстояние от нейрона до искомой группы входных данных. Чем оно больше – тем больше параметр ошибки:

$$\Delta error\,(s_1) = \left\|w_{s_1} - \xi\right\|^2$$

6 Нейронный газ перемещает ближайший нейрон s_1 и его прямых топологических соседей к ξ на доли ε_b и ε_n соответственно от общего расстояния:

$$\Delta w_{s_1} = \varepsilon_b(\xi - w_{s_1})$$

$\Delta w_n = \varepsilon_n(\xi - w_n)$ для всех прямых соседей n из s_1,

где $\varepsilon_b, \varepsilon_n$ – скорость обучения обновленного или созданного нейрона и соседних ему нейронов соответственно.

Кроме того, также в сторону тех же данных перемещаются и нейроны, соединенные синапсами с перемещаемым нейроном (на рис. 5 силуэтами отражена базовая позиция нейронов).

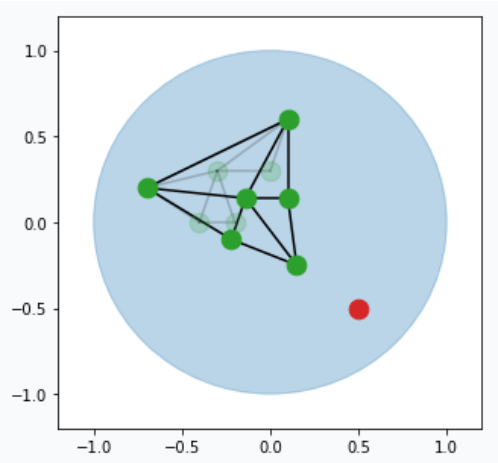

Рис. 5. – Перемещение ближайшего нейрона к искомым данным

7 В случае, если s_1 и s_2 соединены синапсом, ИНС устанавливает вес этого синапса как число ноль. Если такого синапса не существует, сеть создаёт его.

8 Следующим этапом находятся синапсы, которые не обновлялись длительное время (параметр a_{max} регулируемый и может составлять как 50, так и 100 и 200 итераций и представляет собой максимальный возраст синапсов) и удаление их из сети (рис. 6). Также на этом этапе производится проверка на наличие нейронов, которые не соединены синапсами вовсе, и также удаление их из сети.

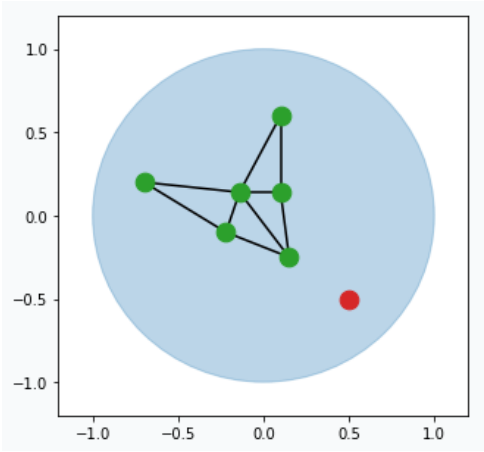

Рис. 6. – Очищение нейронной сети от неиспользуемых данных (синапсов, которые не обновлялись длительное время)

9 Если количество сгенерированных входных сигналов слишком велико и является целым кратным параметру λ – периоду между итерациями порождения новых нейронов, а также в случае нахождения нейрона с наибольшей накопленной ошибкой (приблизительно каждые 100-200 итераций находится нейрон, который имеет наибольшую накопленную ошибку) в нейросеть добавляется новый нейрон, но только в случае, если количество нейронов не достигло максимального значения, заданного параметром n_{max}.

10 Определяется нейрон q с максимальной накопленной ошибкой. Для этого нейрона находится соседний ему нейрон f с наибольшей накопленной ошибкой. Посредине между ними создается новый нейрон r (синяя точка данных на рис. 7), который будет автоматически соединен с двумя исходными нейронами новым синапсом с разрушением начального синапса между этими нейронами (рис. 7). Положение нейрона определяется по формуле:

$$w_r = 0.5(w_q + w_f)$$

Новый нейрон помогает начальному нейрону с максимальной накопленной ошибкой в ИНС уменьшить уровень этой ошибки, умножив их на константу α – параметр затухания накопленных ошибок при создании новых нейронов, передавая, таким образом, новорожденному нейрону r часть ошибок соседних нейронов.

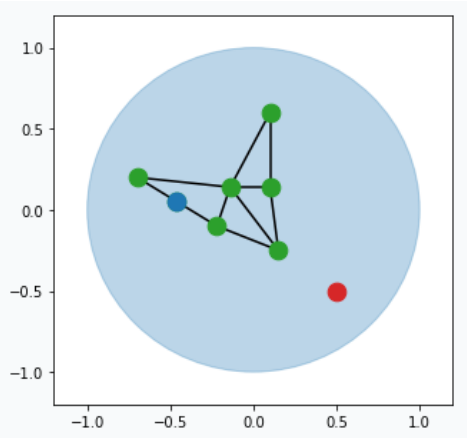

Рис. 7. – Создание нового нейрона на основе нейрона с максимальным значением накопленной ошибки

11 Нейронная сеть уменьшает общий уровень ошибки нейронов, умножив их на константу d. Уменьшение уровня ошибки нейронов означает улучшение структуры данных сети.

12 Если критерий остановки (например, размер нейросети или какой-либо иной показатель эффективности) еще не выполнен, нейронный газ возвращается к шагу 1 [10].

Экспериментальная проверка предложенного алгоритма

Далее выполним проверку предложенного алгоритма на некотором наборе экспериментальных данных. Продемонстрируем работу алгоритма сети растущего нейронного газа, которая динамически адаптируется к распределению сигнала в пространстве. На рис. 8 представлена начальная стадия работы алгоритма, в исследуемой области R^n, которая на рисунке обозначена внешним кубом. В качестве распределения входных данных $P(\xi)$ на рисунке представлен массив внутренних кубов с множеством черных меток внутри, каждая из которых представляет собой вектор входных данных ξ. Каждый из нейронов ИНС обозначен на рисунках зеленой меткой, а синапсы между ними – черными соединительными линиями. Рис. 9, 10, 11 демонстрируют работу нейронной сети в течение времени, увеличивая

количество нейронов сети и корректируя их местоположения в процессе кластеризации данных. Приведем заданные ограничения для нейронной сети: $n_{max} = 100$, $\varepsilon_b = 0.02$, $\varepsilon_n = 0.006$, $\lambda = 500$, $a_{max} = 100$.

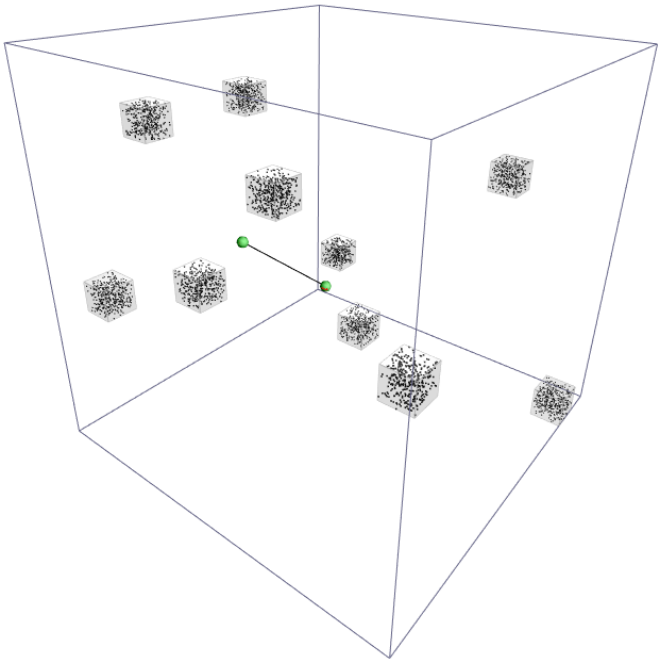

Рис. 8. – Начальный этап работы ИНС растущего нейронного газа

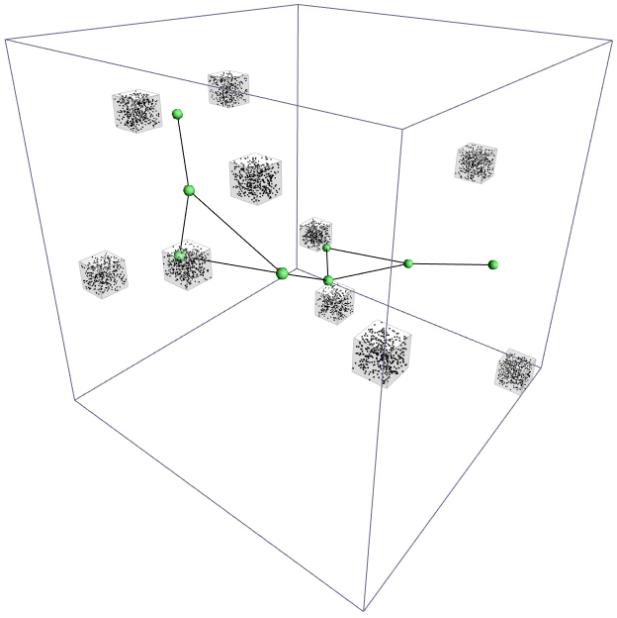

Рис. 9. – Этап 1 работы алгоритма растущего нейронного газа

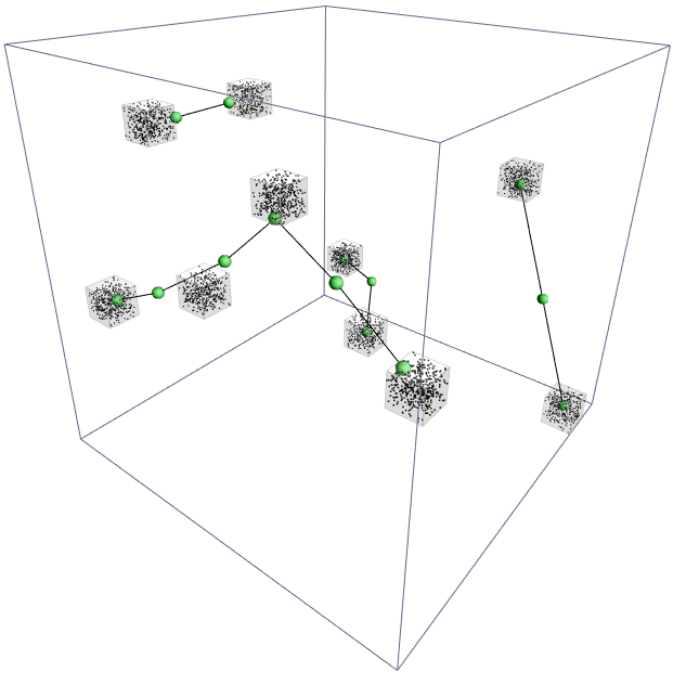

Рис. 10 – Этап 2 работы алгоритма растущего нейронного газа.

На рис. 11 видно, как на итоговом этапе работы алгоритма разрываются лишние синаптические связи и уже найдены кластеры данных.

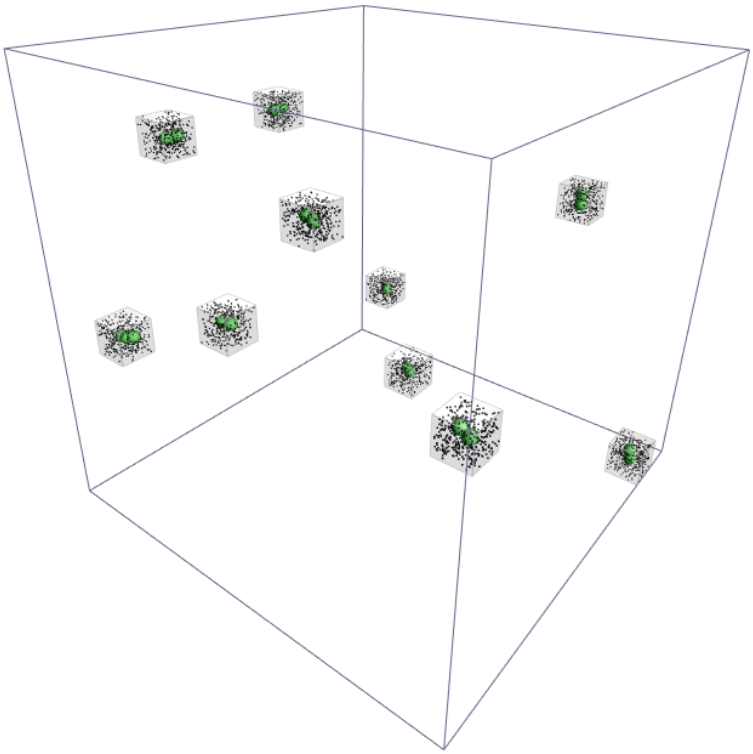

Рис. 11. – Результат работы алгоритма растущего нейронного газа

Несмотря на заданные ограничения, алгоритм завершил свою работу по кластеризации данных через $t = 12504$ входных сигналов, создав всего 23 нейрона и 13 синапсов между ними и обнаружив 10 кластеров данных.

На рис. 11 можно увидеть, как ИНС успешно кластеризовала входные данные. Отсутствие синапсов между выходными кластерами данных также указывает на успешность выполненного процесса кластеризации.

Заключение

В статье предложен модифицированный алгоритм кластеризации, основанный на выявлении топологии входных группы данных при помощи растущего нейронного газа и последующей их кластеризации. Нейронный газ позволяет быстро построить ИНС, отображающую топологию исходных данных. Ограничивая размер построенной сети, можно эффективно управлять вычислительной сложностью алгоритма. Алгоритм кластеризации представляет полученную ИНС как граф и выполняет кластеризацию, автоматически подбирая нужное число кластеров. Нейроны сети становятся центрами кластеров (при этом каждый кластер описывается несколькими центрами).

К плюсам алгоритма относится скорость его работы (при этом она может контролироваться исследователем), а также способность находить кластеры любой формы (за счёт того, что кластер описывается несколькими центрами). Однако приведенный алгоритм обладает и недостатками. Неудачный выбор настраиваемых параметров сети приводит к тому, что алгоритм не может корректно определить число кластеров и работает неэффективно. При этом большое количество настраиваемых параметров невозможно подобрать при помощи кросс-валидации [11] (из-за вычислительной сложности) или простого метода подбора, поэтому подбирать параметры приходится на основе эвристики.

Работа выполнена при финансовой поддержке РФФИ, проекты 18-01-00402-а, 19-01-00246-аб 19-07-00329-а.

Литература

1. Карташов О.О., Бутакова М.А., Чернов А.В., Костюков А.В., Жарков Ю.И. Средства представления знаний и извлечения данных для интеллектуального анализа ситуаций // Инженерный вестник Дона, 2018, №4. URL: ivdon.ru/ru/magazine/archive/n4y2018/5421

2. Кондратьева Т.Н., Эксузян К.А. Анализ данных на базе технологии частного облака // Инженерный вестник Дона, 2018, №3. URL: ivdon.ru/ru/magazine/archive/n3y2018/5165

3. Семенов А. М., Соловьева Р.А. Основы искусственного интеллекта: электронный курс лекций. М-во науки и высш. образования Рос. Федерации, Федер. гос. бюджет. образоват. учреждение высшего образования «Оренбург. гос. ун-т». – Оренбург: ОГУ. – 2019. – 930 с.

4. Крашенинников А.М., Гданский Н.И., Рысин М.Л. Построение сложных классификаторов для объектов в многомерных пространствах // Инженерный вестник Дона, 2013, №2. URL: ivdon.ru/ru/magazine/archive/n2y2013/1611

5. Al-Chalabi, H. Al-Douri Y., Zhang L. Data clustering // Cogent Engineering. – 2018. – №5. – pp. 1 – 16.

6. Savvas, I.K., Michos C., Chernov A., Butakova M. High Performance Clustering Techniques: A Survey. // Advances in Intelligent Systems and Computing, vol 1156. Springer, Cham. – 2020. – pp. 252-259.

7. Воронцов К. В. Методы кластеризации: курс лекций. URL: machinelearning.ru/wiki (дата обращения 10.07.20).

8. Иванов А.П., Гилевский С.В. Применение векторного квантования для увеличения устойчивости к ошибкам в канале связи алгоритма сжатия JPEG2000 // Электроника инфо. – 2016. – №9. – С. 54-55.

9. Fritzke B. A. Growing Neural Gas Network Learns Topologies. Advances in Neural Information Processing Systems, 7. MIT Press, Cambridge MA. – 1995. pp. 625–632.

10. Riveiro M., Ventocilla E. Visual Growing Neural Gas for Exploratory Data Analysis // 10th International Conference on Information Visualization Theory and Applications. – 2019. – pp. 58 – 71.

11. Красоткина О.В., Моттль В.В., Разин Н.А., Черноусова Е.О. Бесперебойная кросс-валидация отбора признаков в линейной регрессионной модели // Известия Тульского государственного университета. Технические науки. – 2013. – №7-2. – С. 88-98.

References

1. Kartashov O.O., Butakova M.A., Chernov A.V., Kostjukov A.V., Zharkov Ju.I. Inzhenernyj vestnik Dona, 2018, №4. URL:ivdon.ru/ru/magazine/archive/n4y2018/5421

2. Kondrat'eva T.N., Jeksuzjan K.A. Inzhenernyj vestnik Dona, 2018, №3. URL:ivdon.ru/ru/magazine/archive/n3y2018/5165

3. Semenov A. M., Solov'eva R.A. Osnovy iskusstvennogo intellekta: jelektronnyj kurs lekcij. [The basics of artificial intelligence: an electronic course of lectures]. M-vo nauki i vyssh. obrazovanija Ros. Federacii, Feder. gos. bjudzhet. obrazovat. uchrezhdenie vyssh. obrazovanija «Orenburg. gos. un-t». Orenburg: OGU. 2019. 930 p.

4. Krasheninnikov A.M., Gdanskij N.I., Rysin M.L. Inzhenernyj vestnik Dona, 2013, №2. URL:ivdon.ru/ru/magazine/archive/n2y2013/1611

5. Al-Chalabi, H. Al-Douri Y., Zhang L. Cogent Engineering. 2018. №5. pp. 1 – 16.

6. Savvas, I.K., Michos C., Chernov A., Butakova M. Advances in Intelligent Systems and Computing, vol 1156. Springer, Cham. 2020. pp. 252-259.

7. Voroncov K. V. Metody klasterizacii: kurs lekcij. [Clustering methods: a course of lectures]. URL: machinelearning.ru/wiki (Date of access 10.07.20).

8. Ivanov A.P., Gilevskij S.V. Jelektronika info. 2016. №9. pp. 54-55.

9. Fritzke B. A. Advances in Neural Information Processing Systems, 7. MIT Press, Cambridge MA. 1995. pp. 625–632.

10. Riveiro M., Ventocilla E. 10th International Conference on Information Visualization Theory and Applications. 2019. pp. 58 – 71.

11. Krasotkina O.V., Mottl' V.V., Razin N.A., Chernousova E.O. Izvestija Tul'skogo gosudarstvennogo universiteta. Tehnicheskie nauki. 2013. №7-2. pp. 88-98.