Isaykina Anastasia Mikhailovna, master student, angel12vat@gmail.com, Russia, Tula, Tula State University

УДК 004.021
DOI: 10.24412/2071-6168-2022-7-209-218

ИССЛЕДОВАНИЕ ВЗАИМОСВЯЗЕЙ МЕЖДУ ПАРАМЕТРАМИ, ХАРАКТЕРИЗУЮЩИХ УНИВЕРСИТЕТЫ МИРА

Ю.А. Леонов, А.С. Сазонова, Л.Б. Филиппова, В.В. Гришина

Статья посвящена актуальной проблеме прогнозирования и определения взаимосвязей между параметрами массива данных для дальнейшего анализа. Приведено решение задачи кластеризации списка университетов на основе использования самоорганизующиеся карты Кохонена с автоматическим определением количества кластеров. Особое внимание уделено нормализации исходных данных и алгоритму обучения самоорганизующейся карты Кохонена, а также способу визуализации таких карт. Для решения поставленной задачи кластеризации было разработано программное обеспечение, функционал которого описан в статье. Проведен анализ полученных кластеров с целью выявления взаимосвязей между параметрами исходных данных.

Ключевые слова: самоорганизующиеся карты Кохонена, машинное обучение, кластеризация, интеллектуальный анализ данных, обучение без учителя, визуализация.

На сегодняшний день решение таких задач, как прогнозирование, поиск различного рода закономерностей в больших массивах данных, а также выявление наборов как зависимых, так и независимых признаков объектов является важной в связи с постоянным увеличением объема анализируемой информации и необходимой автоматизацией аналитических процессов. Данные задачи относятся к классу задач кластеризации – поиску независимых групп и их характеристик во всем множестве данных.

В рамках исследования в качестве исходных данных был взят список мировых университетов [1], которые необходимо объединить в группы по общим признакам – кластера.

Задача анализа заключается в необходимости кластеризации списка университетов, при этом каждый из кластеров будет обозначать сходство включенных в него университетов по характеризующим параметрам. Каждый из университетов имеет значение параметра в пределах от 0 до 100 по шести характеристикам, а также общее количество студентов университета. Часть исходных данных приведена в таблице 1.

Для решения задач кластеризации существует множество алгоритмов, которые позволяют объединить все схожие исходные объекты по описываемым параметрам. Среди них можно выделить алгоритм k-means (k - средних), EM-алгоритм, алгоритмы семейства FOREL и т.д.

Методы и материалы исследования. Для решения поставленной задачи был выбран алгоритм самоорганизующейся карты Кохонена (SOM – self-organizing map), который предназначен для визуального представления всех свойств объектов на двумерной карте [2, 3]. Такие карты помогают отображать входные данные высокой размерности в виде массива малой размерности.

Согласно алгоритму самоорганизующейся карты Кохонена для решения поставленной задачи были проведены следующие этапы [4]:

1. Задание структуры (архитектуры) нейронной сети.
2. Нормализация входных данных.
3. Инициализация весовых коэффициентов.
4. Поиск BMU для примера из обучающей выборки.
5. Вычисление радиуса окрестности BMU.
6. Коррекция вектора весов нейронов.

На этапе задания структуры нейронной сети необходимо обозначить количество нейронов выходного слоя K. Помимо этого каждый пример из обучающей выборки представляет собой n-мерный вектор $V = (v_1, v_2, \ldots, v_n)$, а каждый нейрон содержит соответствующий n-мерный вектор весов $W = (w_1, w_2, \ldots, w_n)$. Также каждый нейрон имеет свои координаты в двумерной сети x и y.

Фрагмент списка мировых университетов с характеристиками

Название университета	Качество обучения	Международный рейтинг	Оценка за исследования	Процент цитирования	Оценка по доходу	Общий рейтинг	Количество студентов
Harvard University	99.7	72.4	98.7	98.8	34.5	96.1	20152
California Institute of Technology	97.7	54.6	98.0	99.9	83.7	96.0	2243
Massachusetts Institute of Technology	97.8	82.3	91.4	99.9	87.5	95.6	11074
Stanford University	98.3	29.5	98.1	99.2	64.3	94.3	15596
University of Cambridge	90.5	77.7	94.1	94.0	57.0	91.2	18812
University of Oxford	88.2	77.2	93.9	95.1	73.5	91.2	19919

Примечание: Источник: [1].

Этап нормализации входных данных подразумевает приведение всех входных значений к промежутку [0, 1], реже к значениям в промежутке [-1, 1]. Для того чтобы нормализовать данные в пределах [0, 1], необходимо воспользоваться следующей формулой:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}, \quad (1)$$

где x – значение параметра университета, заданное изначально; x_{max} – максимальное значение параметра; x_{min} – минимальное значение параметра.

Инициализировать весовые коэффициенты можно различными способами, однако при нормализации исходных данных в пределах [0, 1] веса w_{ij} можно инициализировать случайным образом как:

$$0.5 - \frac{1}{\sqrt{M}} \le w_{ij} \le 0.5 + \frac{1}{\sqrt{M}}, \quad (3)$$

где M – количество входных переменных сети (характеристических признаков университета).

После инициализации весов необходимо провести поиск нейрона BMU для примера из обучающей выборки.

BMU (Best Matching Unit) – нейрон, компоненты вектора весов которого наиболее близки к компонентам вектора входных сигналов. Для нахождения BMU необходимо вычислить расстояние между входным вектором и вектором весов для каждого из нейронов сети по формуле:

$$D = \sqrt{\sum_{i=0}^{n} (v_i - w_i)^2}, \quad (4)$$

где v_i – i-ый компонент вектора V (пример из обучающей выборки); w_i – i-ый компонент вектора W.

Нейрон, для которого данное расстояние будет наименьшим, помечается как BMU. После нахождения BMU производится поиск нейронов, которые находятся в окрестности данного BMU. На этапе кластерного анализа BMU определяет принадлежность входного примера к соответствующему кластеру. В дальнейшем, для нейронов, которые входят в радиус окрестности BMU, корректируются значения их весов. При этом чем ближе нейрон к BMU, тем больше изменяется вес.

Радиус окрестности вычисляется по следующей формуле:

$$\delta = \delta_0 * e^{(-\frac{t}{\lambda})}, \quad (5)$$

где δ_0 – радиус окружности на первой итерации, t – номер итерации, λ – постоянная времени

При этом радиус окрестности в процессе обучения постоянно сокращается. Радиус окрестности на первой итерации δ_0 определяется по формуле:

$$\delta_0 = \frac{\max\{w,h\}}{2}, \quad (6)$$

где w – ширина сетки нейронов; h – высота сетки нейронов.

Постоянная времени λ вычисляется по формуле:

$$\lambda = \frac{T}{\ln \delta_0}, \quad (7)$$

где T – общее число итераций; δ_0 – радиус окрестности на первой итерации.

Прежде чем приступить к корректировке вектора весов, необходимо найти расстояние d от каждого нейрона до BMU и сравнить его с радиусом окружности δ:

$$d = \sqrt{(x_i - x_{BMU})^2 + (y_i - y_{BMU})^2} < \delta\,, \quad (8)$$

где x_i и y_i – координаты нейрона, который сравнивается с координатами BMU, x_{BMU} и y_{BMU} – координаты BMU.

Если для i-ого нейрона выполняется данное соотношение, то данный нейрон лежит в окрестности BMU и, следовательно, необходимо корректировать вектор весов для этого нейрона по формуле:

$$W' = W + \theta L * (V - W), \quad (9)$$

где W' – вектор весов после коррекции; W – вектор весов до коррекции; Q - влияние удаленности нейрона от BMU; L – скорость обучения; V – вектор входных значений, соответствующий BMU.

Влияние удаленности Q нейрона от BMU вычистятся по формуле:

$$\theta = e^{\left(-\frac{d^2}{2\delta^2(t)}\right)}\,, \quad (10)$$

где d – расстояние от нейрона (узла) до BMU; δ – радиус окрестности (в зависимости от итерации).

Скорость обучения L определяется формулой:

$$L = L_0 * e^{\left(-\frac{t}{\lambda}\right)}\,, \quad (11)$$

где L_0 – скорость обучения на первой итерации (~ 0.3), t – номер итерации, λ – число оставшихся итераций.

После корректировки вектора весов необходимо вернуться к шагу нахождения BMU до тех пор, пока все примеры из выборки не будут использованы в обучении карты Кохонена или скорость обучения не упадет до установленного минимального значения [5].

Как и любой другой метод, SOM удобно визуализировать для наглядного отображения работы алгоритма и конечного результата работы. Результирующие карты Кохонена можно отображать несколькими способами, а именно линейно, двухмерно и трехмерно. При этом двухмерное отображение самоорганизующейся карты Кохонена является самым распространенным и более наглядным способом визуализации (рис. 1).

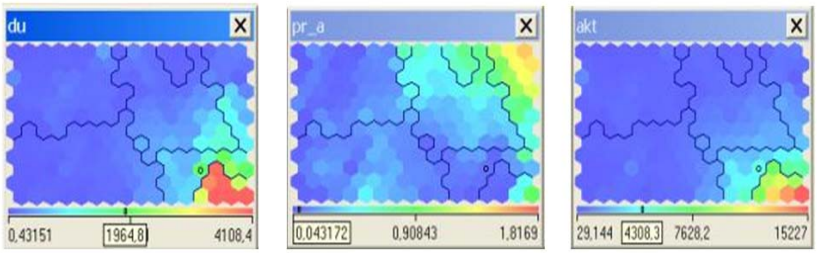

Рис. 1. Визуализация результата работы SOM

Прежде чем перейти к построению карты Кохонена, необходимо задать основные характеристики сетки, к которым относятся количество нейронов и конфигурация сетки нейронов. От заданного количества нейронов зависит степень детализации карты Кохонена, то есть чем больше будет число нейронов в сетке, тем более детально будет происходить отображение карты. Но следует отметить тот факт, что при работе с большим количеством нейронов потребуется больше времени для обучения.

В свою очередь, конфигурация сетки также важна при корректном отображении результатов работы. Часто при визуализации карты Кохонена нейроны представляют в виде прямоугольных или шестиугольных ячеек. В случае использования шестиугольных ячеек (рис. 2) происходит наиболее корректное отображение расстояния между объектами карты по сравнению с прямоугольными ячейками, так как расстояние между центрами смежных шестиугольных ячеек одинаково.

В результате работы алгоритма самоорганизующейся карты Кохонена можно получить такие карты, как карта входов и выходов нейронов, а также так называемые специализированные карты, к которым относят карты кластеров, матрицу расстояний и другие карты, характеризующие кластеры, которые получены в результате обучения сети.

Самоорганизующиеся карты Кохонена позволяют анализировать в первую очередь объекты, которые характеризуются множеством признаков или параметров. Двумерная карта выходов нейронов позволяет отображать на плоскости близость многомерных векторов при-

знаков, так как объекты, у которых векторы признаков близки относительно друг друга, попадают либо в одну ячейку, либо в смежные ячейки. Таким образом, для анализа объектов полезно знать, сколько векторов входных данных связано с каждой ячейкой карты.

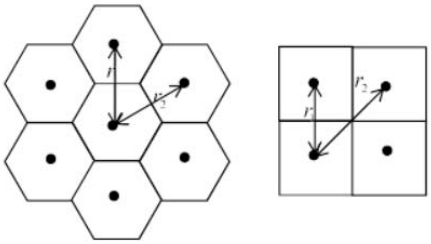

Рис. 2. Шестиугольные и прямоугольные ячейки

Помимо анализа сходства множества объектов, часто требуется провести анализ конкретных параметров, по которым проявляется сходство этих объектов. Для выполнения данной задачи необходимо построить и раскрасить такое количество карт входов нейронов, сколько параметров содержат анализируемые объекты. Другими словами, количество карт определяется количеством компонентов входных векторов университета. Таким образом, каждая построенная карта будет соответствовать конкретному параметру университета [6,7].

Между всеми вариациями карт Кохонена существует некоторая взаимосвязь, а именно все они являются различными раскрасками одних и тех же нейронов.

Стоит отметить, что ключевым моментом в использовании карт Кохонена является настройка гиперпараметров SOM, а именно размер карты, количество итераций и скорость обучения. Данные настройки напрямую влияют на конечный результат кластеризации и, соответственно, на автоматическое выделение кластеров [8].

Результаты исследований и их обсуждение. Для решения поставленной задачи была разработана программа, которая позволяет проводить автоматическую кластеризацию списка мировых университетов для последующего выявления взаимосвязей между характеризующими параметрами. Интерфейс разработанной программы представлен на рис. 3-7.

На рис. 3 представлено отображение карты кластеров в результате работы алгоритма SOM с последующим объединением ячеек непосредственно в кластеры. Также в интерфейсе программы представлено описание каждого кластера, а именно средние значения каждого параметра соответствующего кластера.

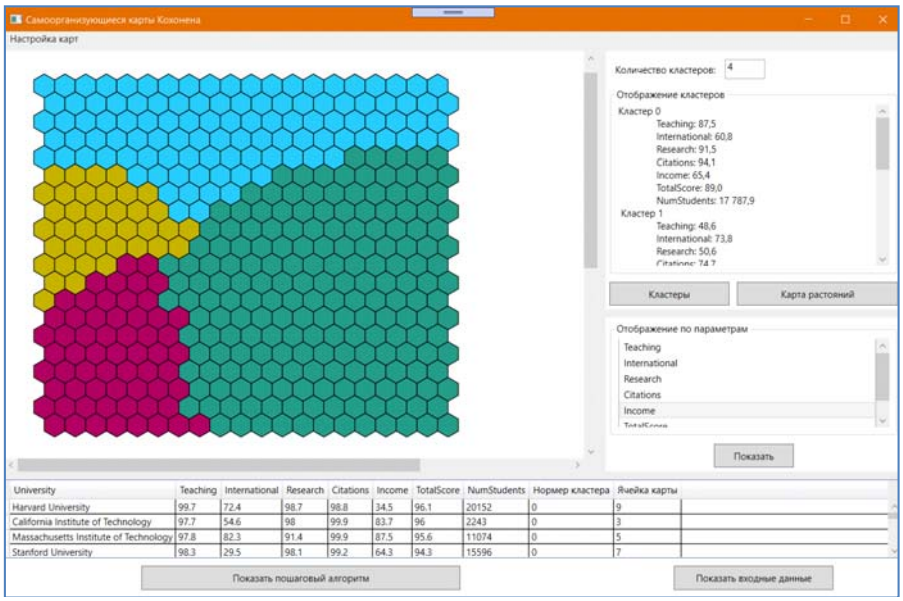

Рис. 3. Отображение карты кластеров

Карта расстояний представлена на рис. 4 и отображает расстояние между полученными кластерами.

Карта расстояний представлена в черно-белом варианте, при этом, чем темнее цвет ячейки, тем ближе в векторном пространстве она находится относительно своих соседей. Именно поэтому, данная карта практически повторяет очертание карты кластеров (рис. 3).

Рис. 5 содержит интерфейс программы, в котором происходит отображение карт характеристик. Как было сказано ранее, количество карт характеристик напрямую зависит от количества самих характеристик, описывающих исходные данные. Раскраска данных карт представлена в желто-красных цветах, при этом, чем темнее цвет, тем больше величина параметра, соответствующая данной характеристики.

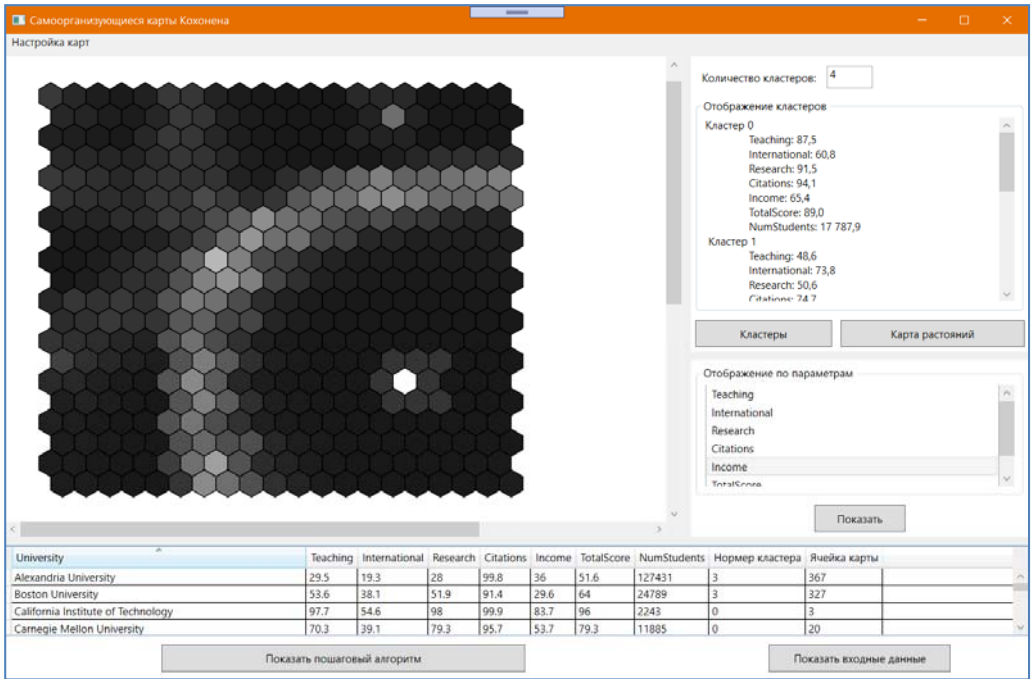

Рис. 4. Отображение карты расстояний

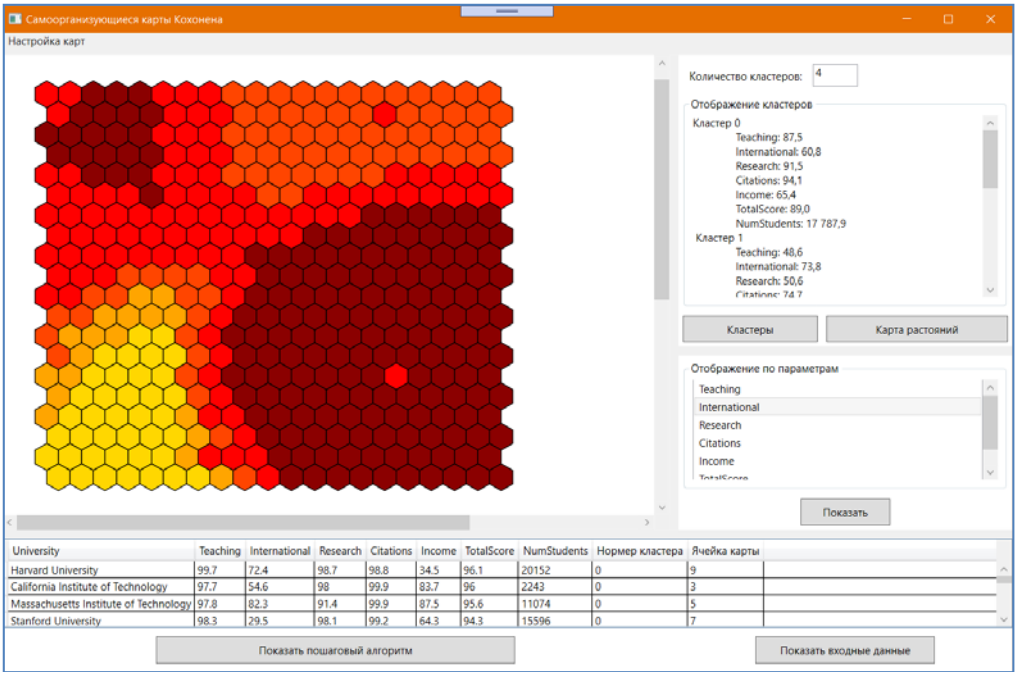

Рис. 5. Отображение карты характеристик

Помимо интерфейса программы для отображения различных вариаций карт Кохонена, разработанная программа позволяет пошагово просмотреть работу алгоритма самоорганизующейся карты Кохонена (рис. 6, 7).

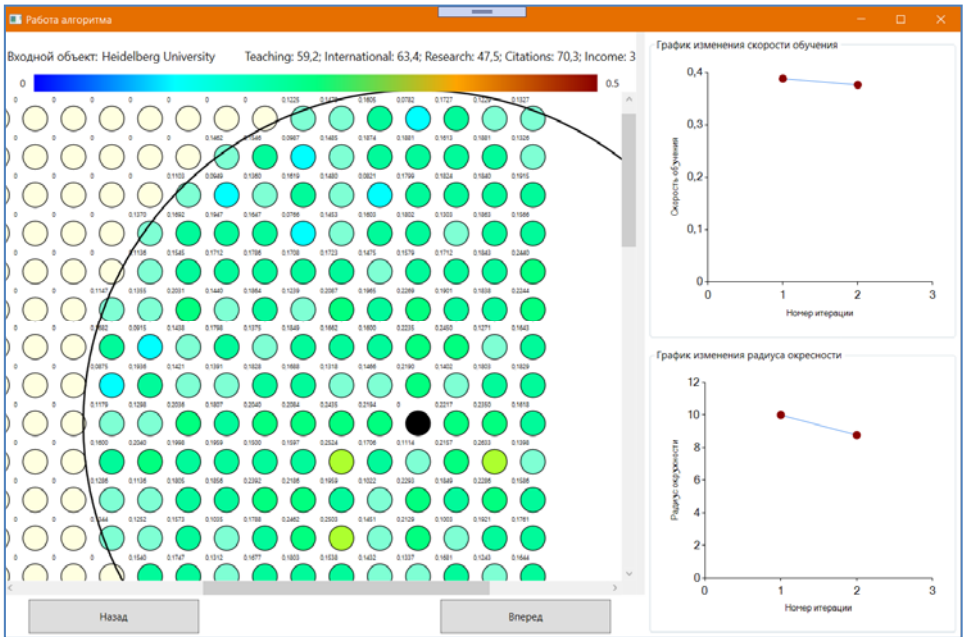

Рис. 6. Пошаговая демонстрация работы алгоритма обучения на 2 шаге обучения

В левой части экрана представлена сетка, состоящая из выходных нейронов, описывающих карту Кохонена.

Каждый нейрон сетки окрашен в различные цвета: от синего до темно красного (цветовая шкала представлена в верхней части). Цвет нейрона зависит от того, как сильно изменяется его положение в векторном пространстве относительно предыдущего шага, при этом синие оттенки говорят о том, что расстояние изменилось незначительно, красные оттенки – значительное изменение расстояния. Помимо этого, в данной сетке отображается нейрон BMU и радиус окрестности.

В правой части (рис. 6, 7) представлены графики изменения скорости обучения и радиуса окрестности. Это позволяет проследить тенденцию изменения главных параметров самоорганизующихся карт Кохонена для наилучшего понимания работы алгоритма.

Если сравнить левые части рис. 6 и 7, можно сделать вывод о том, что при одном BMU при снижении скорости обучения и уменьшении радиуса окрестности, изменение векторного расстояния каждого нейрона с каждой итерацией становится все меньше. При этом на последних итерациях данное расстояние практически не будет изменяться.

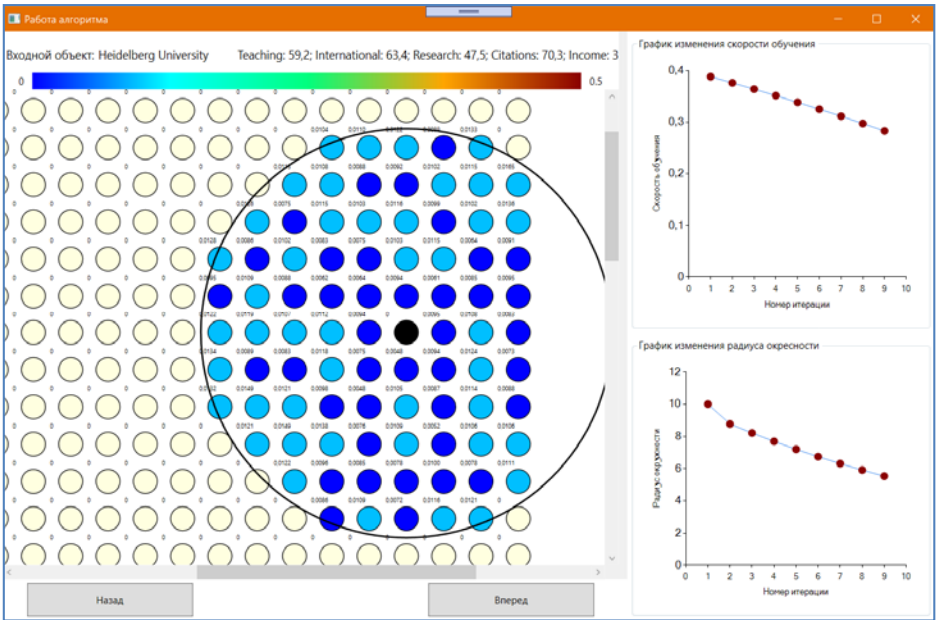

Рис. 7. Пошаговая демонстрация работы алгоритма обучения на 9 шаге обучения

В рамках оценки работы программы было проведено тестирование работы класса, предназначенного для реализации работы алгоритма обучения самоорганизующейся карты Кохонена. Данное тестирование проводилось с целью отслеживания времени работы алгоритма обучения самоорганизующейся карты Кохонена в зависимости от количества входных данных (максимальное количество – 48000 объектов). Оборудование, на котором проводилось тестирование, имеет процессор Intel Core i3-5005U (2.0 GHz), 2 ядра (4 логических процессора). Оценка работы определялась при работе с двумерной самоорганизующейся картой Кохонена размерности 20x20, с количеством итераций 25 и скоростью обучения равной 0,4. Результат тестирования представлен на рис. 8.

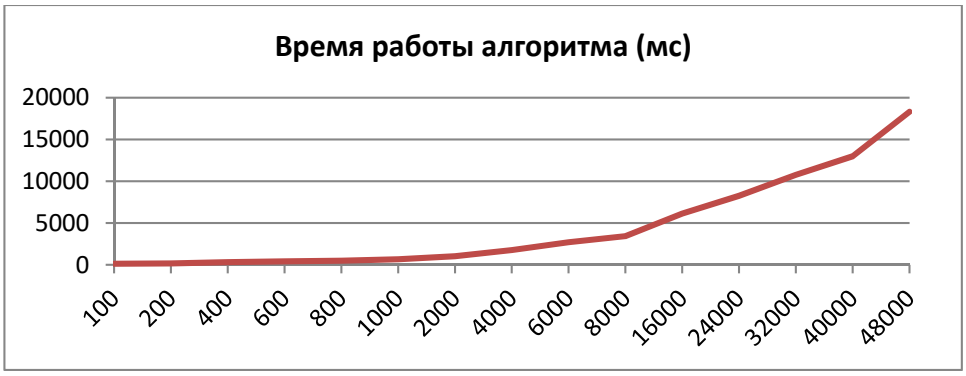

Рис. 8. Результаты тестирования работы класса SOM

Помимо тестирования работы класса SOM также проводилось тестирование скорости отображения списка университетов со значениями параметров по всем характеристикам и номером кластера и ячейки карты. Результаты данного тестирования представлены на рис. 9.

Рис. 9. Результаты тестирования отображения данных

Таким образом, можно сделать вывод о том, что при количестве входных объектов больше 8000, время работы алгоритма SOM и время отображения всех объектов начинает резко возрастать.

В результате решения поставленной задачи была проведена кластеризация списка университетов мира с автоматическим определением количества кластеров. Все университеты были объединены в 4 кластера, в каждом из которых содержатся университеты, схожие по описывающим характеристикам.

При анализе полученных кластеров можно сделать вывод о том, что:

первый кластер имеет высокое качество обучения, оценку за исследования, процент цитирования и общий рейтинг;

второй кластер включает в себя университеты с международным рейтингом и процентом цитирования выше среднего, но с низкой оценкой по доходу;

третий кластер объединяет университеты с высокой оценкой по доходу, когда как все остальные характеристики имеют средние значения показателей (55-70);

четвертый кластер имеет низкий международный рейтинг и средние оценки по остальным характеристикам (50-70), но при этом самое большое количество студентов, по сравнению с другими кластерами.

Заключение. Использование самоорганизующихся карт Кохонена в решении задачи кластеризации мировых университетов позволяет достичь поставленных задач, а именно выделить группы объектов, схожих по своим характеристикам, и визуально представить все свойства объектов на двумерной карте для проведения детального анализа и выявления закономерностей в данных.

Исходя из анализа полученных кластеров, характеристиками, которые имели наибольшее влияние на составление полученных кластеров, являются качество обучения и рейтинг университета. Однако решение данной задачи не является полностью законченной, так как решение любой задачи кластеризации во многом является подготовительным этапом для дальнейшего анализа.

Список литература

1. Рейтинг университетов мира QS. [Электронный ресурс]. URL: https://www.educationindex.ru/articles/university-rankings/qs (дата обращения: 10.05.2022).

2. Гордополов Ю.В., Лукашевич Н.С. Кластеризация регионов по уровню социально-экономического развития на основе самоорганизующихся карт Кохонена // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Экономические науки. 2010. С. 27-33.

3. Лочмеле Р.Р. Современные количественные методы экономического анализа: самоорганизующиеся карты Кохонена (СОК) // Государственное управление. Электронный вестник. 2003. №3. С. 1-5.

4. Дьяченко В.А., Михаль О.Ф. Адаптивная параллельная процедура обучения самоорганизующейся модифицированной карты Кохонена // Восточно-Европейский журнал передовых технологий. 2012. 2/4 (56). С. 11-14.

5. Кохонен Т. Самоорганизующиеся карты. Москва: БИНОМ, 2014. 656 с.

6. Анисимова Э.С. Самоорганизующиеся карты Кохонена в задачах кластеризации // Актуальные проблемы гуманитарных и естественных наук. 2014. С. 1-2

7. Головачев С.С. Кластеризация данных и роевые методы обучения искусственных нейронных сетей в прогнозировании рынка ценных бумаг // Финансовый журнал. 2013. №2. С. 85-96

8. Сеньковская И.С., Сараев П.В. Автоматическая кластеризация в анализе данных на основе самоорганизующихся карт Кохонена // Вестник Магнитогорского государственного технического университета им. Г.И. Носова. 2011. № 2. С. 78-79.

9. Kuzmenko A.A., Filippova L.B., Sazonova A.S., Filippov R.A. Intelligent System of Classification and Clusterization of Environmental Media for Economic Systems // Proceedings of the International Conference on Economics, Management and Technologies 2020 (ICEMT 2020). - Advances in Economics, Business and Management Research, 2020. Volume 139. P. 583-586. DOI 10.2991/aebmr.k.200509.103.

10. Leonov YU.A., Leonov E.A., Kuzmenko A.A., Martynenko A.A., Averchenkova E.E., Filippov R.A. Selection of rational schemes automation based on working synthesis instruments for technological processes. Yelm, WA, USA: Science Book Publishing House LLC, 2019. 192 p.

11. Кузьменко А.А., Кондратенко С.В., Сазонова А.С., Аверченков А.В., Филиппов Р.А. Разработка структуры WEB-ресурса на основе потребностей конечного пользователя // Новые информационные технологии в научных исследованиях Материалы XXIII Всероссийской научно-технической конференции студентов, молодых ученых и специалистов. 2018. Т. 2 Рязань: Рязанский государственный радиотехнический университет. С. 183-185.

12. Филиппов Р.А., Филиппова Л.Б., Сазонова А.С. Интернет вещей: основные понятия: учебно-методическое пособие. Брянск: БГТУ, 2016. 112 с.

13. Leonov E.A., Intellectual subsystems for collecting information from the internet to create knowledge bases for self-learning systems / E.A. Leonov, Y.A. Leonov, Y.M. Kazakov, L.B. Filippova/ In: Abraham A., Kovalev S., Tarassov V., Snasel V., Vasileva M., Sukhanov A. (eds) — Text : electronic // Proceedings of the Second International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’17). IITI 2017. Advances in Intelligent Systems and Computing. 2017. Vol. 679. Springer, Cham. P. 95-103. DOI:10.1007/978-3-319-68321-8_10.

14. Казаков Ю.М., Тищенко А.А., Кузьменко А.А. Оценка научной деятельности аспирантов и молодых ученых с использованием когнитивного моделирования // VIII Международной научно-практической конференция «Современные технологии в российской и зарубежных системах образования» сборник статей. Пенза, ПГАУ, 2019. С. 46-49.

15. Аверченкова Е.Э., Сазонова А.С., Аверченков А.В., Кузьменко А.А., Тищенко А.А., Филиппов Р.А. Основы инновационной деятельности предприятия: учебное пособие. М.: ООО «Флинта», 2019. 162 с.

Леонов Юрий Алексеевич, канд. техн. наук. доцент, yorleon@yandex.ru, Россия, Брянск, Брянский государственный технический университет,

Сазонова Анна Сергеевна, канд. техн. наук. доцент, asazonova@list.ru, Россия, Брянск, Брянский государственный технический университет,

Филиппова Людмила Борисовна, канд. техн. наук, доцент, libv88@mail.ru, Россия, Брянск, Брянский государственный технический университет,

Гришина Валерия Викторовна, студент, libv88@yandex.ru, Россия, Брянск, Брянский государственный технический университет

RESEARCH OF THE INTERRELATION BETWEEN PARAMETERS THAT CHARACTERIZE UNIVERSITIES AROUND THE WORLD, BASED ON THE USE OF SELF-ORGANIZING MAPS

Yu.A. Leonov, A.S. Sazonova, L.B. Filippova, V.V. Grishina

The article is devoted to the relevant problem of prediction and determination the interrelation between the parameters of the data array for further analysis. The solution of the problem of clustering a list of universities based on the use of self-organizing maps with automatic determination of the number of clusters is presented. Specific attention is devoted to the normalization of the initial data, the learning algorithm of the self-organizing map and the method of visualizing such maps. To solve the task of clustering, software was developed, the functionality of which is described in the article. The obtained clusters are analyzed in order to identify interrelation between the parameters of the initial data.

Key words: self-organizing maps, machine learning, clustering, data mining, unsupervised learning, visualization.

Leonov Yuriy Alekseyevich, candidate of technical sciences, docent, yorleon@yandex.ru, Russia, Bryansk, Bryansk state technical University,

Sazonova Anna Sergeyevna, candidate of technical sciences, docent, asazonova@list.ru, Russia, Bryansk, Bryansk state technical University,

Filippova Lyudmila Borisovna, candidate of technical sciences, docent, libv88@mail.ru, Russia, Bryansk, Bryansk state technical University,

Grishina Valeriya Viktorovna, student, libv88@yandex.ru, Russia, Bryansk, Bryansk State Technical University