Санкт-Петербургский государственный университет

Прикладная математика, программирование и искусственный интеллект

Отчет по учебной практике (научно-исследовательской работе)

Нейронные сети Кохонена для задачи кластеризации

Выполнил:

Иващенко Сергей Юрьевич

Кафедра прикладной кибернетики

Санкт-Петербург

2024

Содержание

Введение

Кластерный анализ – способ группировки многомерных объектов, основанный на представлении результатов отдельных наблюдений точками подходящего геометрического пространства с последующим выделением групп, как «сгустков» этих точек (кластеров). Иными словами, кластеризация – классификация объектов на основе их сходства друг с другом, при которой принадлежность этих объектов каким-либо классам не задаётся.

Кластерный анализ как научное направление зародился в середине 60-х годов и с тех пор бурно развивается. В современной жизни кластеризация применяется почти во всех сферах, где имеет место анализ и обработка данных: искусственный интеллект и машинное обучение, распознавание образов и обработка изображений, работа с документами и квантование [1].

Нейронная сеть – математическая модель, имитирующая собой поведение группы нервных клеток организма. По сути, это суперпозиция линейных и нелинейных функций, выстроенных в определённую структуру, аппроксимирующая неизвестную зависимость. Аппроксимация достигается выбором параметров сети, коэффициентов линейных функций, достигающих экстремум некоторого критерия.

Нейронная сеть Кохонена – вид нейронной сети, имеющий только два слоя (входной и выходной), где каждый нейрон первого слоя соединён с каждым нейроном второго слоя. Такая структура была впервые описана в 1980-х годах финским учёным в области искусственного интеллекта Теуво Кохоненым [2].

Формально такую модель можно представить как алгоритм кластеризации, где в качестве координат центров кластеров выступают коэффициенты сети. При этом число нейронов выходного слоя регулируемое, что позволяет настроить количество кластеров для соответствующей детализации. Оптимальное расположение центров кластеров происходит в результате обучения сети Кохонена, во время которой минимизируется функционал качества – среднее внутриклассовое расстояние.

В работе изложено математическое описание нейронной сети Кохонена, приведены её модификации. Также предоставлено сравнение результатов работы алгоритма с результатами метода K-means.

Постановка задачи

Цель данной работы – исследовать нейронную сеть Кохонена в качестве метода кластеризации.

Чтобы достичь этой цели необходимо решить следующие задачи:

1. Изучить архитектуру нейронной сети Кохонена и применить к задаче кластеризации.
2. Найти и рассмотреть модификации метода.
3. Сравнить результаты работы сети и результаты метода K-means.

Глава 1. Описание нейронной сети Кохонена

1.1 Базовое описание модели

Нейронная сеть Кохонена представляет собой двухслойную нейронную сеть. Каждый нейрон первого (распределительного) слоя соединен со всеми нейронами второго (выходного) слоя, которые расположены в виде двумерной решетки.

Нейроны выходного слоя называются кластерными элементами, их количество определят максимальное количество групп, на которые система может разделить входные данные.

Выходной слой сети Кохонена состоит из некоторого количества n параллельно действующих линейных элементов. Все они имеют одинаковое число входов m и получают на свои входы один и тот же вектор входных сигналов $x = (x_1,\dots, x_m)$. Значение на выходе j-го линейного элемента:

$$y_j = \sum_{i=1}^{m} w_{ji} x_i,\ j \in 1\dots n,$$

где w_{ji} – весовой коэффициент i-го входа j-го нейрона.

После прохождения слоя линейных элементов значения посылаются на обработку по правилу WTA (англ.: Winner Takes All – победитель забирает всё): среди выходных значений y_j ищется максимальный с номером $j_{max} = \operatorname{argmax}_j y_j$. Нейрон с номером j_{max} называется нейроном-победителем.

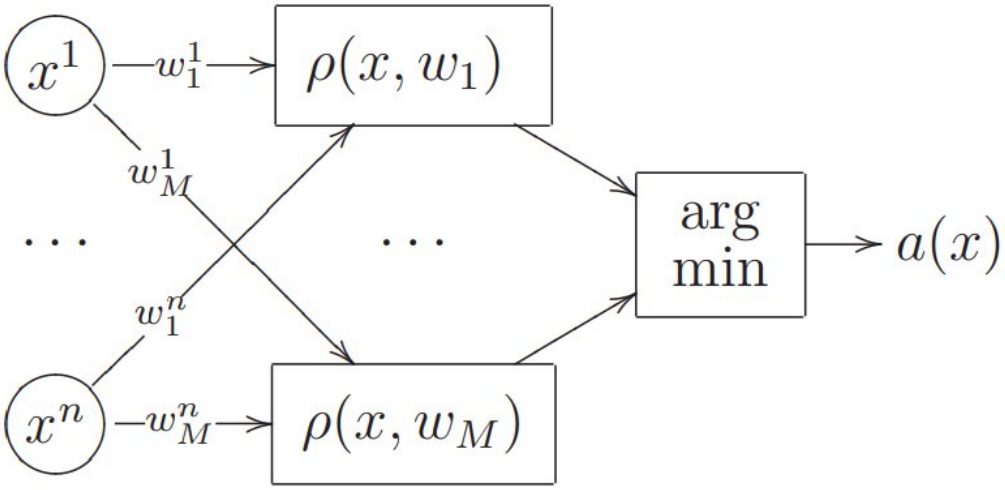
Окончательно, на выходе сигнал с номером j_{max} равен единице, остальные — нулю. Если максимум одновременно достигается для нескольких j_{max}, то либо принимают все соответствующие сигналы равными единице, либо только первый в списке (по соглашению).

1.2 Описание модели для кластеризации

При задаче кластеризации векторы $w_m \in \mathbb{R}^n,\ m = 1,\dots,M$ описывают центры кластеров. Произвольный объект $x \in X$ относится к ближайшему кластеру по следующему правилу:

$$a(x) = \operatorname*{argmin}_{m \in Y} \rho(x, w_m).$$

Данный алгоритм является модификацией базовой версии нейронной сети Кохонена, где вместо скалярных произведений вычисляются расстояния до объектов, а функция аргумента минимизации заменяет функцию аргумента максимизации.



Представление в виде схемы.

1.3 Настройка алгоритма кластеризации

Настройка алгоритма сводится к оптимизации расположения центров w_m. Для этого минимизируется функционал качества кластеризации, равный половине суммы квадратов расстояний между объектами и центрами кластеров:

$$Q(w_1,\ldots, w_M) = \frac{1}{2}\sum_{i=1}^{l} \rho^2\left(x_i, w_{a(x_i)}\right) \to \min_{\{w_m\}}.$$

Допустим, метрика евклидова: $\rho(x,w) = \|x - w\|$. Тогда градиент функционала по вектору w_m:

$$\frac{\partial Q}{\partial w_m} = \sum_{i=1}^{l} (w_m - x_i)\,[a(x_i) = m].$$

Для поиска векторов w_m можно использовать стохастический градиентный спуск с выражением на этапе обновления весов:

$$w_m := w_m + \eta\,(x_i - w_m)\,[a(x_i) = m],$$

где x_i – случайный объект выборки, η – градиентный шаг (можно выбрать как обратное к числу итераций).

Таким образом, если объект x_i относится к кластеру m, то центр этого кластера w_m немного сдвигается в направлении объекта x_i, остальные центры не изменяются.

Глава 2. Модификации сети Кохонена как алгоритма кластеризации

2.1 Правило справедливой конкуренции (conscience WTA)

Во время обучения сети Кохонена возникает необходимость в начальной инициализации весов. Как при случайной инициализации, так и при инициализации дальнейшими объектами нейрон Кохонена может попасть в такую область, где он никогда не станет нейроном-победителем. Потенциально это означает появление пустого кластера.

Во избежание этой проблемы вводится механизм штрафа за слишком частое присоединение объектов. Этот механизм называется правилом conscience WTA (англ.: conscience – справедливый).

Формула алгоритма модифицируется следующим образом:

$$a(x) = \underset{m \in Y}{\operatorname{argmin}} C_m \rho(x, w_m),$$

где C_m – количество побед m-го нейрона в ходе обучения.

2.2 Правило мягкой конкуренции (WTM)

Другим недостатком правила WTA является медленная скорость сходимости, связанная с тем, что на каждой итерации модифицируется только один нейрон-победитель. Для ускорения сходимости, особенно на начальных итерациях, можно подстраивать сразу несколько нейронов, близких к объекту x_i.

Для этого вводится ядро — неотрицательная монотонно убывающая на $[0, +\infty)$ функция расстояния, например, $K(\rho) = \exp(-\beta\rho^2)$. Здесь $\beta > 0$.

Градиентный шаг меняется следующим образом:

$$w_m := w_m + \eta (x_i - w_m) K(\rho(x_i, w_m)), \; m = 1, \ldots, M.$$

Теперь на каждой итерации центры всех кластеров смещаются в сторону x_i, но чем дальше центр находится от x_i, тем меньше величина смещения.

Модификация является обобщением базовой модели, где $K = [a(x_i) = m]$.

На начальных итерациях имеет смысл выбрать небольшое значение коэффициента β, чтобы все весовые векторы успели переместиться ближе к области входных векторов. Затем β можно увеличивать, делая конкуренцию всё более жёсткой, и постепенно переходя к коррекции только одного нейрона-победителя.

Например, можно положить $\beta = \exp\left(t^{-2}\right)$, где t – число проведённых итераций.

Глава 3. Сравнение результатов работы методов кластеризации

3.1 Сведения о данных и о реализации алгоритмов

Далее будут предложены результаты применения алгоритма Кохонена и K-means. В качестве источника данных выбран дата-сет с сервиса Kaggle под названием «Clustering Categorical Peoples Interests». Набор содержит 6340 записей, состоящих из поля группы (категория “C”, “R”, “I” или “P”), 217 вопросов о хобби (поля с числом 2, 1 или 0), а также поля с суммой этих 217 полей.

Данные были обработаны перед применением алгоритмов: проведена нормализация (стандартное отклонение 1, среднее арифметическое 0), удалены столбцы (признаки), где отсутствуют значения более чем для 2000 записей. Необходимость в последнем действии возникла в процессе применения алгоритмов. Полный учёт всех признаков сильно понижал качество кластеризации обоих алгоритмов. Вероятно, некоторые признаки вносили «шум» в набор данных. В результате процедуры осталось 11 признаков. Пропущенные значения заменены нулём. В контексте данных это означает, что пропущенное значение равносильно отрицательному ответу на вопрос о хобби (пропущенных значений в других признаках не обнаружено). Категориальные признаки заменены числовыми (категория в поле группы заменена числом от 1 до 4).

Реализация K-means взята из библиотеки для работы с данными sklearn. Начальные данные (центры кластеров) выбираются случайно. Количество кластеров (настраиваемый параметр в этих алгоритмах) равно трём, так как это число даёт лучшие значения критериев.

Программные реализации написаны на языке программирования Python с использованием библиотек numpy, pandas, а также seaborn и pyplot.

3.2 Критерии качества кластеризации

Поскольку данные многомерные, то наглядно показать разбиение объектов по кластерам проблематично. Для оценки результатов предлагается использовать некоторые популярные внутренние критерии качества кластеризации.

Среди них:

1. Компактность (меньше – лучше):
$$WSS = \frac{\sum_{i<j} [y_i = y_j] \rho(x_i, x_j)}{\sum_{i<j} [y_i = y_j]}.$$
2. Отделимость (больше – лучше):
$$BSS = \frac{\sum_{i<j} [y_i \neq y_j] \rho(x_i, x_j)}{\sum_{i<j} [y_i \neq y_j]}.$$

3. «Силуэт» (больше – лучше):

$$SIL = \frac{1}{N}\sum_{c_k \in C}\sum_{x_i \in c_k}\frac{b(x_i, c_k) - a(x_i, c_k}{\max\{a(x_i, c_k), b(x_i, c_k\}},$$

где $a(x_i, c_k) = \frac{1}{|c_k|}\sum_{x_j \in c_k}\left\|x_i - x_j\right\|$, $b(x_i, c_k) = \min_{c_l \in C \setminus c_k}\{\frac{1}{|c_l|}\sum_{x_j \in c_l}\left\|x_i - x_j\right\|\}$.

4. Индекс Данна (больше – лучше):

$$D = \frac{\min\limits_{c_k \in C}\{\min\limits_{c_l \in C}\{\|c_k^{cp} - c_l^{cp}\|\}\}}{\max\limits_{c_k \in C}\{\max\limits_{x_i, x_j \in c_k}\{\left\|x_i - x_j\right\|\}\}}.$$

5. Индекс Дэвиса-Болдуина (меньше – лучше):

$$\frac{1}{K}\sum_{c_k \in C}\max_{c_l \in C \setminus c_k}\{\frac{S(c_k) + S(c_l)}{\|c_k^{cp} - c_l^{cp}\|}\},$$

где $S(c_k) = \frac{1}{|c_k|}\sum_{x_i \in c_k}\|x_i - c_k^{cp}\|$.

3.3 Результаты работы алгоритмов

Сравнение критериев. Слева направо: индекс Данна, отделимость, силуэт, компактность, индекс Дэвиса-Болдуина. Синий – значение K-means, жёлтый – сети Кохонена.

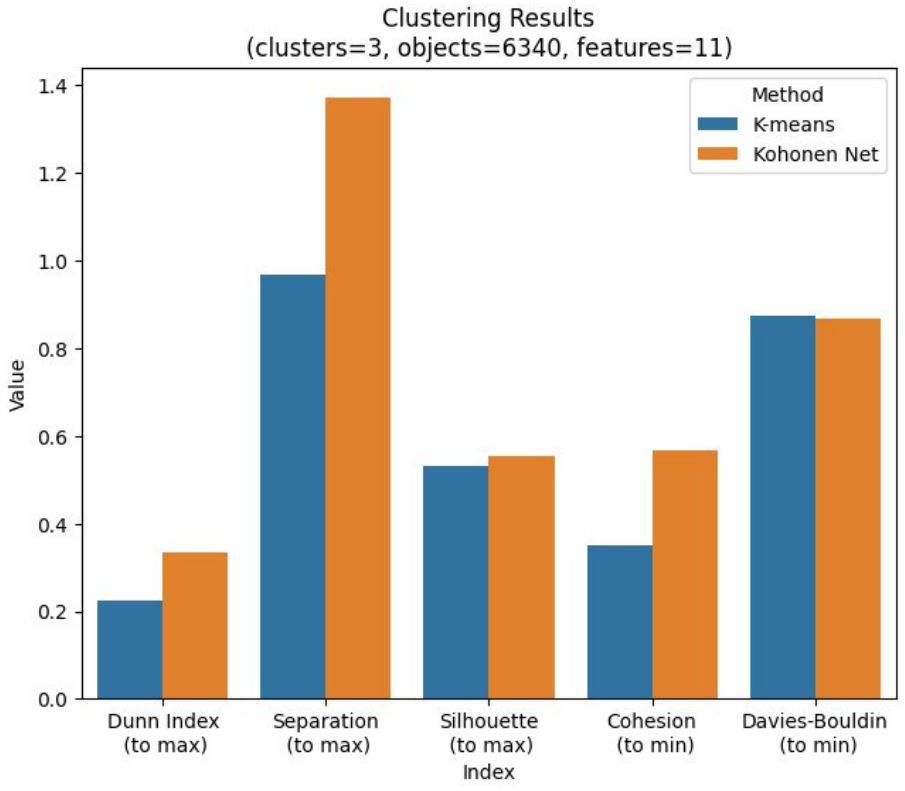


Рис. 1. Результаты сравнения алгоритмов K-means и сети Кохонена по критериям Данна, отделимости, силуэта, компактности, Дэвиса-Болдуина.

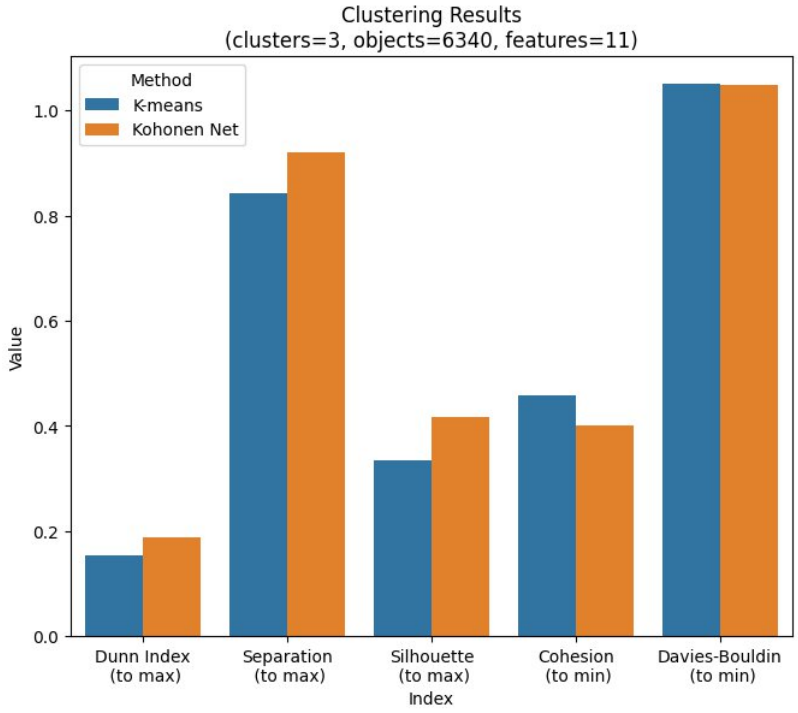


Рис. 2. Результаты сравнения алгоритмов K-means и сети Кохонена с правилом мягкой конкуренции по критериям Данна, отделимости, силуэта, компактности, Дэвиса-Болдуина.

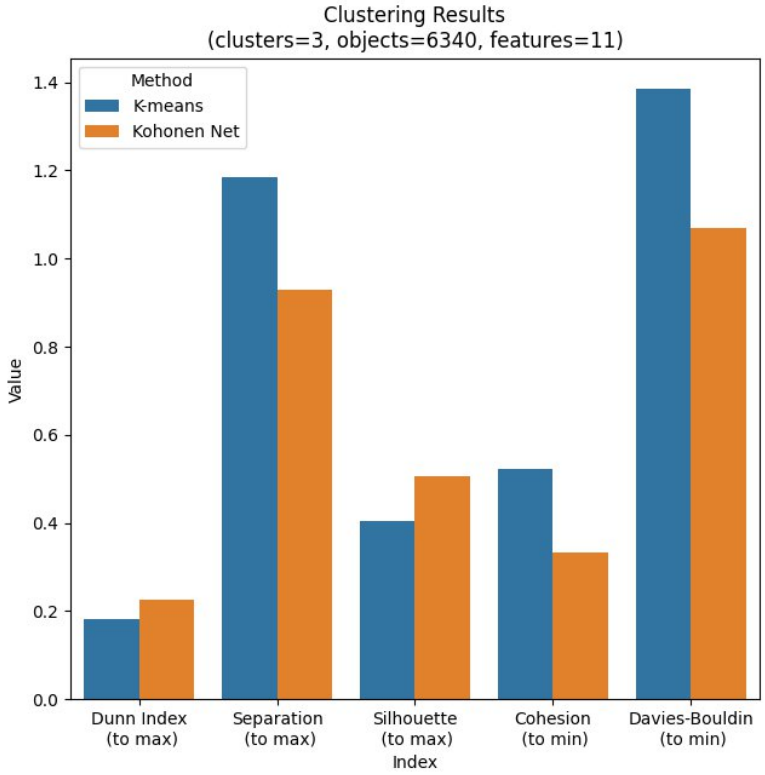


Рис 3. Результаты сравнения алгоритмов K-means и сети Кохонена с правилом справедливой конкуренции по критериям Данна, отделимости, силуэта, компактности, Дэвиса-Болдуина.

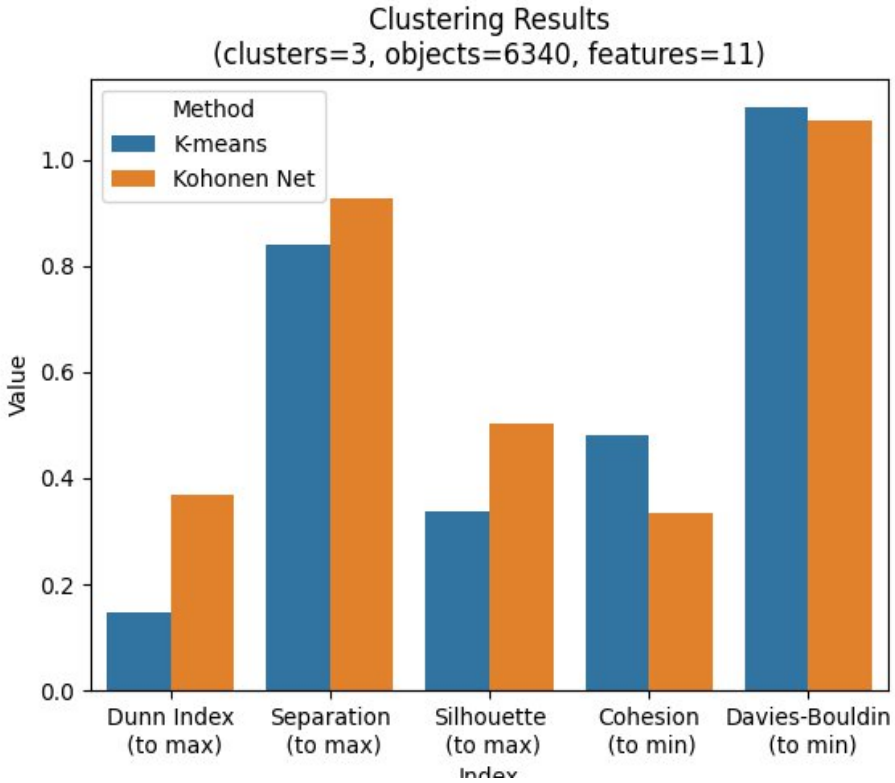


Рис. 4. Результаты сравнения алгоритмов K-means и сети Кохонена с правилами справедливой и мягкой конкуренции по критериям Данна, отделимости, силуэта, компактности, Дэвиса-Болдуина.

3.3 Анализ результатов работы алгоритмов

Для базового алгоритма (без модификаций) основные отличия имеют место в показаниях отделимости и компактности. Сеть Кохонена выигрывает по первому критерию, но кластеры получаются менее компактными. По этой же причине индексы Данна и Дэвиса-Болдуина находятся примерно на одном уровне, так как пропорциональны отношению отделимости к компактности (или обратно пропорциональны). Значения силуэта также примерно равны. В целом, можно судить о приблизительно равном качестве кластеризации.

Модификации вносят вклад в работу алгоритма. Правило мягкой конкуренции проявляет некоторый эффект сглаживания, сильно сокращая компактность кластеров (улучшает качество). Очевидным обратным эффектом является понижение отделимости. Однако пониженное значение отделимости всё равно превосходит таковое у K-means. Поэтому в целом качество кластеризации улучшается, о чём свидетельствует значение силуэта с появившейся положительной относительно сети Кохонена разницей. Модификация правилом справедливой конкуренции ещё сильнее улучшает компактность, но без перемещения всех центров несильно ухудшает итоговую отделимость. Критерий Дэвиса-Болдуина показывает заметное улучшение.

При применении обеих модификаций видны положительные эффекты. Улучшаются компактность, отделимость и силуэт. Общее качество кластеризации превосходит K-means по рассмотренным критериям.

Заключение

В ходе этой работы была описана нейронная сеть Кохонена, её применение в задачах кластеризации многомерных объектов. Помимо этого, были представлены модификации данного метода, продемонстрированы и проанализированы результаты сравнения нейронной сети с методом кластеризации K-means.

Анализ полученных результатов даёт понять, что сеть Кохонена – крайне перспективное решение для выполнения задач кластеризации. В зависимости от необходимого критерия качества возможно подобрать вариант сети так, чтобы получить выигрыш по сравнению с K-means.

В заключение следует отметить, что метод кластеризации сетью Кохонена имеет большой потенциал к дальнейшему развитию. Причиной этому является нейросетевая суть подхода. Она позволяет объединять в себе свойства метрических методик, а также нейронных моделей. Так, улучшения в структуре нейронной сети Кохонена повлекут положительные изменения в результате кластеризации. Например, обучение путём минимизации функционала позволяет получить выраженный прирост к внутрикластерному расстоянию, а обилие методов нахождения минимума функционала предоставляет возможность улучшения вычислительной стороны вопроса. К сожалению, сеть Кохонена не способна решить некоторые фундаментальные проблемы кластеризации. Одна из них – выбор начальных данных. Решения одной и той же задачи могут кардинально отличаться в зависимости от случайно выбранных начальных данных. Таким образом, направление имеет открытые вопросы, которые могут стать интересными направлениями изучения в будущих исследованиях.

Список литературы

1. Jain A., Murty M., Flynn P. Data Clustering: A Review. // ACM Computing Surveys. 1999. Vol. 31, no. 3.
2. T.Kohonen, “Self-Organizing Maps” Springer, 1995.
3. Воронцов, К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования / К. В. Воронцов. — М.: МГУ, 2007.
4. Кластеризатор на основе нейронной сети Кохонена [Электронный ресурс] URL: http://mechanoid.su/neural-net-kohonen-clusterization.html#x1-20011 (Дата обращения 21.05.2024).