DOI 10.31891/2307-5732-2022-307-2-87-91
УДК 637.5.02

ПЕТРОВ Д. Д.
https://orcid.org/0000-0001-9108-7903
e-mail: dmytro.petrov.knm.2018@lpnu.ua
БОЙЧУК А. Р.
Національний університет "Львівська Політехніка"
https://orcid.org/0000-0002-0563-5748
e-mail: andrii.r.boichuk@lpnu.ua

МЕТОД КЛАСТЕРИЗАЦІЇ ДАНИХ НА ОСНОВІ АЛГОРИТМУ ОБХОДУ ГРАФА

В роботі наведено результати досліджень методу кластеризації, побудованому на основі обходу графу в ширину та представлено результати роботи на типовому наборі даних для задачі кластеризації, проведено порівняльний аналіз застосування методу.

Ключові слова: кластеризація, граф, обхід графа в ширину.

DMYTRO PETROV, ANDRYY BOICHUK
Lviv Polytechnic National University

CLUSTERIZATION METHOD BASED ON BREADTH FIRST SEARCH OR BFS FOR A GRAPH

Clusterization is one of the types of algorithms of unsupervised learning. The idea behind it is that an algorithm learns patterns from untagged data. Such type of algorithm helps to find unseen dependencies in the untagged data itself. This paper presented algorithms based on Breadth-First Search or BFS for a Graph. The method was built based on the basic theory of clusterization. To the theory of clusterization, the calculated distance between the two farthest points in the cluster should be less than the distance between the closest two points from different clusters. By this rule, we defined that two parameters of the method should be the maximum distance between points by which these can be connected and assumed to be in one cluster. The second had to be the maximum distance in the cluster, aka the cluster's diameter. A cluster's diameter is the farthest distance between two points within a cluster. With these hyperparameters and the defined distance method, we can assume that every point is a vertex of a graph, two points within the threshold of the distance between pairs of ones are neighbours, and count the connection between counts as an edge of a graph. The group of connected vertexes or a particular vertex is a graph. The diameter hyperparameter ought to keep the data homogeneity in a cluster. We can define every graph as a cluster with defined rules based on previous assumptions. Later in this paper will be visualized the clusterization of three-dimensional data points. We took one of the most popular clusterization dataset - the iris dataset for visualizing purposes. The paper contains several examples of clusterization of the dataset with different hyperparameters. We took KMeans [3] as an example of the clusterization method. The method based on BFS is a flexible clusterization method that relies on meta-information about distancing between data points.

Keywords: clusterization, BFS, KMeans, unsupervised.

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Для якісного пошуку прихованих залежностей у даних використовують різні методи навчання без вчителя [1–4]. З метою розширення методологічної бази аналізу даних авторами роботи запропоновано метод кластеризації на основі алгоритму обходу графа в ширину. Даний метод надасть можливість для ширшого аналізу даних та за допомогою додаткових гіперпараметрів гнучко налаштовуватися під особливості конкретного набору інформації.

Аналіз досліджень та публікацій

В роботі [4] наведено перелік основних статистичних методів для пошуку прихованих залежностей даних. Пропонується замість побудови фільтру, який є необхідною передумовою звичайного навчання, відтворення даних, елементи яких є незалежними, що дає змогу утворювати асоціації з логічними функціями елементів, а не тільки з самими елементами.

В роботі [3] представлений K-mean алгоритм кластеризації, що утворюють динамічні групи даних в кількості заданих ядр, а також пропонується удосконалення, що зменшує обчислювальну складність алгоритму без значних змін у якості роботи.

В роботі [2] представляється алгоритм проходу графу в ширину та показується розкид при відборі даних за допомогою даного методу. В статті обраховується степінь розподілу ймовірності дослідження вершини для випадкового графа методом обходу в ширину і представляється як функція поділу пройдених вершин.

В роботі [1] наводяться базові поняття кластерного аналізу, його приклади застосування та методи кластеризації. Серед іншого наводяться алгоритми кластеризації K-means з варіаціями, ієрархічна кластеризація та кластеризація на основі щільності даних.

Формулювання цілей статті

Метою роботи є дослідження ефективності кластеризації методом побудованим на основі алгоритму обходу графу в ширину.

Виклад основного матеріалу

Методика побудови методу кластеризації на основі алгоритму обходу графа в ширину складається з означення загальних понять, визначення вимог до алгоритму відповідно до теорії кластеризації, опис алгоритму та експериментальне доведення роботи методу.

Кластерний аналіз – це задача розбиття множини значень на підмножини, що називаються кластерами, в межах яких значення будуть вважатися подібними, а значення між будь-якими елементами з двох кластерів будуть вважатися різними, або істотно відмінними. Діаметр кластера – це найбільша відстань між двома значеннями в межах цього кластера.

Графом [5] називають скінченну дискретну не порожню множину точок, що називають вершинами, які з'єднані між собою ребрами. Діаметром графа називається найбільша відстань між будь-якими двома його вершинами.

Згідно з кластерним аналізом, як було описано вище, необхідно контролювати відстань між точками в межах кластеру та відстань між кластерами, щоб зберігалися відповідні відношення. В такому випадку вдасться уникнути ситуації, коли до кластеру буде віднесена точка, що знаходиться на віддалі від графа більшій ніж до інших графів. Для визначення подібності використовуються різні метрики відстані.

Для побудови алгоритму припустимо, що всі точки являються вершинами. Тоді для побудови графа нам потрібно пройтися по всім парам точкам, ребра яких мають відстань меншу за задане обмеження. Ребра, які не перевищують довжину обмеження вважаються існуючими, а інші не існуючими. Якщо ребро від поточної точки до наступної існує, то при додані точки до графа перевіряється, чи діаметр графа не перевищує заданий максимальний діаметр. За умови, що при доданні точки діаметр не перевищується, точка додається в кластер, якщо діаметр виходить за обмеження, то точка не додається. Для визначення подібності точок і розрахунку довжин ребра використовуються евклідова відстань [6].

$$D(p,q) = \sqrt{(p_1 - q_1)^2 + \ldots + (p_i - q_i)^2 + \ldots + (p_n - q_n)^2}, \quad (1)$$

З метою проведення дослідження класифікації запропонованим методом було вибрано набір даних ірисів [7] із 150 записів 4 колонок характеристик. Дані описують 150 записів квітів по 4 параметрам. Серед них є 3 підвиди ірису по 50 записів на кожен. Оглядовий опис даних представлений на рис. 1.

	sepal_length	sepal_width	petal_length	petal_width
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
...	...	...	...	...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

Рис. 1. Описове представлення набору даних квітів ірису

Для візуалізації кластеризації даних було вибрано 3 характеристики з 4 можливих, щоб була змога представити утворення групи даних у просторовій діаграмі. Серед наявних колонок були використані наступні: petal_width, sepal_width та petal_length. Приклад візуалізації в просторі із використанням характеритстик petal_width, sepal_width та sepal_length наведений на Рис. 2.

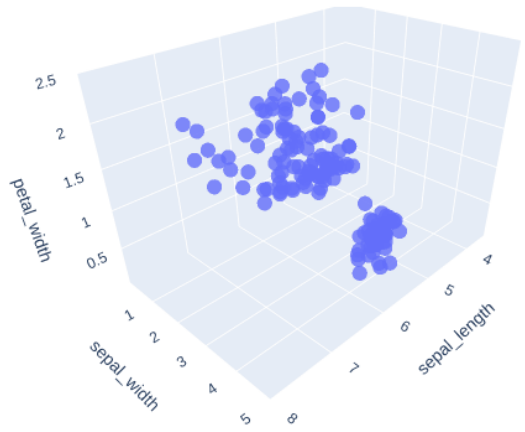

Рис. 2. Просторове представлення даних квітів ірису

Оригінальні дані містять колонку підпису виду ірису, що для візуалізації просторового розподілу було використано у візуальному аналізі з метою використання як еталон при проведенні кластеризації. Дані розмічені цілочисельними значеннями, в колонці species_id, де 1 позначає вид setosa, число 2 позначає versicolor, а значення ідентифікатора виду 3 позначає virginica. Вище згадані класи на рис. 3 диференціюються за кольором: темно-синій відповідає за вид setosa, маджента [8] відповідає за вид versicolor та жовтий позначає вид virginica. З рисунку видно, що клас 1 однозначно відділяється як група значень від інших. Класи 2 та 3 мають доволі близькі значення на межі один між одним.

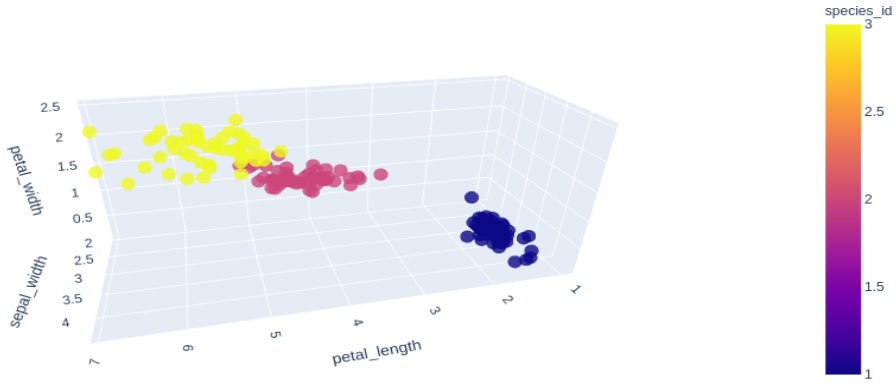

Рис. 3. Просторове представлення розподілу даних за класом запису

Проведемо експеримент задавши початкові параметри випадковим чином. Обмежимо дожину ребра в 0.5 та діаметр класу обмежимо до 2.5. Таким чином ми отримаємо розбиття даних на 6 класів. Як видно з Рис. 4 , клас 0 згідно з обмежень не включив точку, що була віднесена до класу 1 і з обмеженням ребра і відсутністю інших близьких точок була єдина віднесена до класу. Як ми бачимо однорідність даних налаштовується вказаними гіперпараметрами.

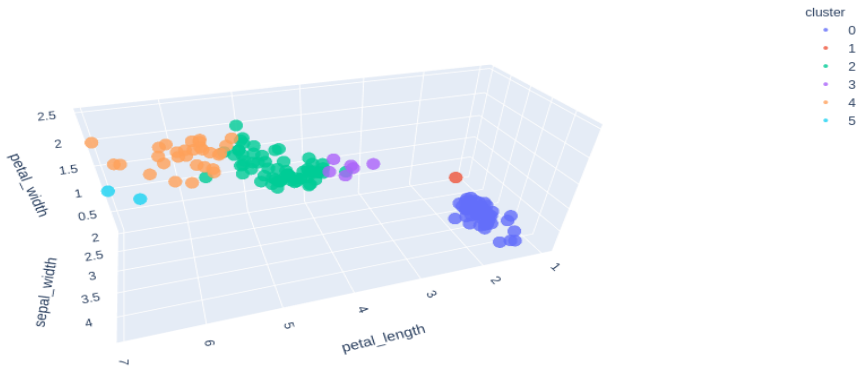

Рис. 4. Візуалізація просторової кластеризації даних за допомогою кластеризації основаної на BFS.

Виставимо показники основуючись на візуальних характеристиках даних і встановимо обмеження діаметру кластеру в 5, а максимальну відстань між даними в кластері 1. Як зображено на рис. 5, класи поділені візуально на 2 однорідні доволі різні групи завдяки підібраним гіперпараметрам.

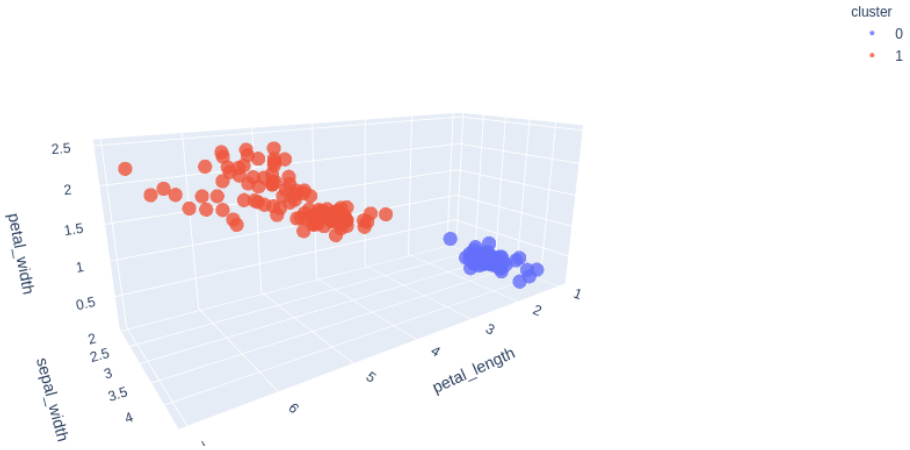

Рис. 5. Візуалізація просторової кластеризації даних за допомогою кластеризації основаної на BFS з підібраними параметрами на основі візуальних даних

Основуючись на числовому аналізі даних, встановимо обмеження діаметру на 3, згідно до відмінностей в даних між різними класами та максимальну відстань між точками в 1. Згідно з рис. 6, задані параметри дозволили наблизитися до кластеризації близької до реально лише з 5% помилково кластеризованих точок.

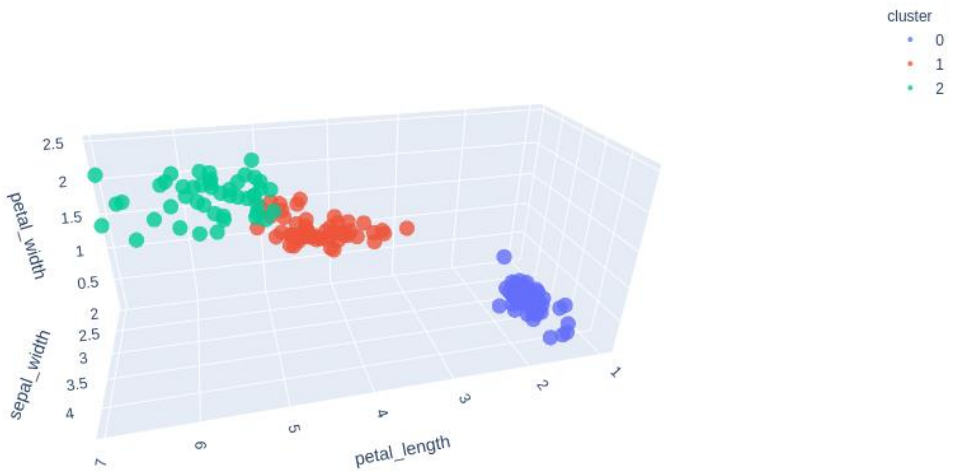

Рис. 6. Візуалізація просторової кластеризації даних за допомогою кластеризації основаної на BFS з підібраними параметрами на основі статистичного аналізу набору даних

Проведемо порівняння кластеризації з KMeans задаючи параметр кількості кластерів 3. Згідно отриманих результатів, та візуалізації на Рис. 7 KNM має помилкове кластеризування 17% даних. Згідно цих результатів, правильно підібрані параметри для методу основаного на BFS втричі точнішу кластеризацію використанням KNM алгоритму на заданому датасеті.

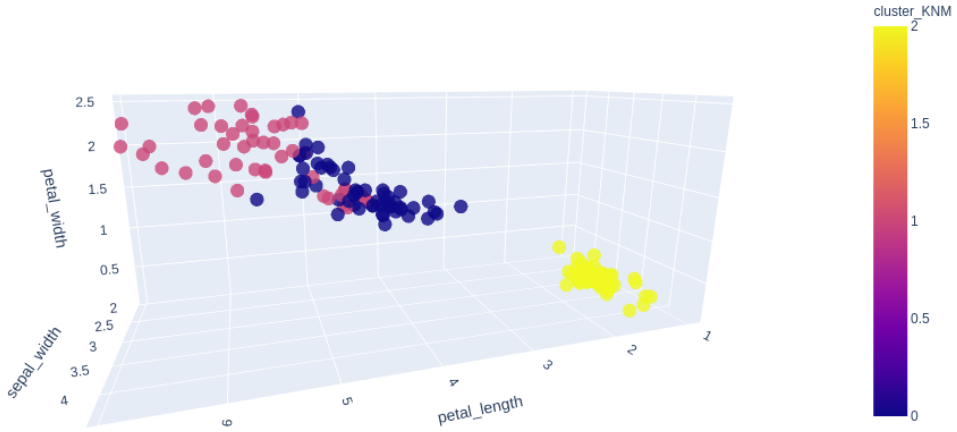

Рис. 7. Візуалізація просторової кластеризації даних за допомогою кластеризації K-means

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

Згідно з проведеними експериментами побудований авторами метод на основі методу обходу графа в ширину являється гнучким інструментом для проведення кластерного аналізу. Підбір параметрів дозволяє провести різноманітний аналіз даних з урахуванням особливостей даних чи навіть метаданих, що дозволить збільшити можливість виявлення прихованих залежностей даних. Даний підхід проведення кластерного аналізу цілком конкурентоспроможний не зважаючи на свою простоту побудови та обчислення. Виявлення інших практичних переваг та недоліків методу потребує подальшого дослідження.

Література

1. Tan P.-N., Steinbach M., Kumar V. Data mining cluster analysis: basic concepts and algorithms. Introduction to data mining. Pearson Education India, 2013. Vol. 487. P. 533.

2. Kurant M., Markopoulou A., Thiran P. On the bias of BFS (breadth first search). 2010 22nd International Teletraffic Congress (lTC 22). IEEE, 2010. P. 1–8.

3. Likas A., Vlassis N., Verbeek J.J. The global k-means clustering algorithm. Pattern recognition. Elsevier, 2003. Vol. 36, № 2. P. 451–461.

4. Barlow H.B. Unsupervised learning. Neural computation. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …, 1989. Vol. 1, № 3. P. 295–311.

5. West D.B. Introduction to graph theory. Prentice hall Upper Saddle River, 2001. Vol. 2.

6. Gower J.C. Properties of Euclidean and non-Euclidean distance matrices. Linear algebra and its applications. Elsevier, 1985. Vol. 67. P. 81–97.

7. Hoey P.S. Statistical Analysis of the Iris Flower Dataset. University of Massachusetts. 2004.

8. Pantone color list · toolstud.io [Electronic resource]. URL: https://toolstud.io/color/pantone.php (accessed: 13.04.2022).

References

1. Tan P.-N., Steinbach M., Kumar V. Data mining cluster analysis: basic concepts and algorithms. Introduction to data mining. Pearson Education India, 2013. Vol. 487. P. 533.

2. Kurant M., Markopoulou A., Thiran P. On the bias of BFS (breadth first search). 2010 22nd International Teletraffic Congress (lTC 22). IEEE, 2010. P. 1–8.

3. Likas A., Vlassis N., Verbeek J.J. The global k-means clustering algorithm. Pattern recognition. Elsevier, 2003. Vol. 36, № 2. P. 451–461.

4. Barlow H.B. Unsupervised learning. Neural computation. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …, 1989. Vol. 1, № 3. P. 295–311.

5. West D.B. Introduction to graph theory. Prentice hall Upper Saddle River, 2001. Vol. 2.

6. Gower J.C. Properties of Euclidean and non-Euclidean distance matrices. Linear algebra and its applications. Elsevier, 1985. Vol. 67. P. 81–97.

7. Hoey P.S. Statistical Analysis of the Iris Flower Dataset // University of Massachusetts. 2004.

8. Pantone color list · toolstud.io. URL: https://toolstud.io/color/pantone.php (accessed: 13.04.2022).