ORCID: 0000-0002-1443-9052

Д. О. СТУПАК
к.т.н., доцент,
старший аналітик консолідованої інформації, EPAM Digital, Черкаси, Україна,
stupak0810@gmail.com

DOI: 10.31651/2076-5851-2024-37-45 PACS: 02.50.Sk, 07.05.Mh, 89.75.Fb, 02.70.Rr

РОЗРОБКА МЕТРИКИ ЯКОСТІ КЛАСТЕРИЗАЦІЇ ОБ'ЄКТІВ ДЛЯ БАГАТОВИМІРНИХ ПРОСТОРІВ

В сучасному світі нові дані генеруються в геометричній прогресії. Кластеризація є одним із методів машинного навчання, що не вимагає розмічених даних, а тому дає можливість швидко визначити структуру даних і зробити певні висновки. В статті розглянуто проблему кластеризації об'єктів в багатовимірному просторі. Ця проблема не нова. Поняття «прокляття розмірності» саме в кластеризації є критичним, оскільки алгоритм має спочатку поділити об'єкти в багатовимірному просторі на кластери, а потім застосовувати метрики якості кластеризації для знаходження оптимальної структури. Існуючі метрики оцінки якості кластеризації часто залежать від розмірності простору, а тому їх використання за таких умов може бути утруднено або призводити до невірних результатів. Метою статті є розробка метрики якості кластеризації, значення якої не залежало б від розмірності простору, в якому описані об'єкти. Для дослідження кластеризації було згенеровано два набори датасетів. В першому наборі об'єкти згруповані у 5 добре розділених кластерів, в другому – кластери майже «торкаються» один одного. Кожен набір містить 6 датасетів із розмірністю простору 10, 100, 300, 1024, 2048 і 4096. Розроблена метрика якості кластеризації базується на порівнянні міжкластерної характеристики поділу об'єктів на кластери і внутрішньокластерної характеристики. В метриці враховані розмірність простору і умова пріоритету поділу об'єктів на меншу кількість кластерів. Використані методи чисельного експерименту для доведення ефективності застосування розробленої метрики якості кластеризації. Зроблена перевірка на синтетичних датасетах, що є близькими по розподілу об'єктів в існуючих датасетах в практичних задачах. Показно, що розроблена метрика якості кластеризації об'єктів дозволяє за «методом ліктя» визначити вірну оптимальну кількість кластерів, не залежить від розмірності простору і може бути застосована навіть в складних випадках, коли кластери розташовані близько обин від одного.

Ключові слова: кластеризація, багатовимірний простір, метрика якості, метод ліктя.

1. Вступ

Постановка проблеми; аналіз останніх досліджень та публікацій; мета статті.

В сучасному діджіталізованому світі кількість інформації, яка накопичується людством збільшується кратно щороку. Все більше даних збирають автоматизовані системи, розвиваються і покращуються великі мовні моделі типу ChatGPT, LLaMa, збільшується можливості по зберіганню і обробці великих об'ємів даних [1]. Часто

виникає ситуація, коли дослідникам невідома структура даних, тобто чи поєднується вони в групи і що ці групи можуть означати. В цьому випадку застосовують один із методів навчання без учителя – кластеризацію. Важливим етапом для застосування методів кластеризації є визначення оптимальної кількості кластерів або груп, на які оптимально слід поділити об’єкти. Для цього використовують метрики якості кластеризації. Однак всі ці метрики якості перевірені для випадків, коли об’єкт описується вектором довжиною до 10. Автори зазначають, що збільшення розмірності векторного простору знижує ефективність використання цих метрик якості. Звідси актуальною є задача розробки метрики якості кластеризації, значення якої б не залежали від розмірності векторного простору, що дозволить використовувати її в багатовимірних просторах.

2. Кластеризація об’єктів

2.1. Методи кластеризації

Існує багато методів кластеризації об’єктів, але всі вони можеть бути поділені на три групи [2]. ***Відцентрові методи*** (k-Means) намагаються визначити положення центрів кластерів і віднести об’єкти до того кластера, центр якого є ближчим до об’єкта. У цих методів є певні недоліки і обмеження, але в цілому вони досить непогано виконують свою задачу. **Ієрархічні методи** (агломеративна кластеризація) будують дерево або ієрархію на базі відстаней між об’єктами. В більшості випадків ці методи працюють повільніше за попередні але якість кластеризації зазвичай трохи краща. Методи кластеризації, засновані на визначенні **густини розташування об’єктів** в просторі (HDBSCAN). Це найповільніша група методів, що дає можливість корректно визначити досить складні структури, але дуже часто ці методи відносіть всі об’єкти до одного кластера, що робить їх досить нестабільними в роботі.

2.2. Існуючі метрики якості кластеризації

Обрані методи кластеризації можуть поділити об’єкти на певну кількість кластерів. Але задача кластеризації полягає не стільки в поділі об’єктів на кластери, скільки у визначенні структури об’єктів в датасеті, тобто визначенні такої кількості кластерів, при яких розбиття на кластри буде найбільше відповідати структурі даних.

Визначити це можна за допомогою метрик якості кластеризації [3]. Існує і активно використовується не менше десяти метрик якості кластеризації.За визначенням оптимального розбиття на кластери метрики поділяються на такі, для яких потрібно шукати максимальне значення, мінімальне значення або використати метод ліктя [4].

Розглянемо ті метрики якості, що найчастіше використовуються при кластеризації.

Silhouette або Silhouette score [5] розраховує те, наскільки кожен об'єкт схожий на об'єкти свого кластеру і не схожий на об'єкти в інших кластерах.

Для кожної точки розраховується середня схожіть із об’єктами свого кластеру

$$a(i) = \frac{1}{|C_I|-1}\sum_{j\in C_I, i\neq j} d(i,j), \quad (1)$$

де C_I – кластер, до якого відноситься об’єкт.

Потім розраховується середня несхожість на об’єкти іншого кластері і обирається найменше значення з порахованих для інших кластерів

$$b(i) = \min_{J\neq I} \frac{1}{|C_J|}\sum_{j\in C_J} d(i,j). \quad (2)$$

Значення метрики розраховується за формулою

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}\ . \quad (3)$$

Якщо значення метрики вище 0,7 вважається що структура даних виявлена ідеально, при значеннях від 0,5 – виявлена, при значеннях менше ніж 0,25 – слабка відповідність. А значення самої метрики можуть знаходитись в діапазоні від -1 до +1. Але ці значення можуть бути досягнені лише для датасетів із невеликою кількістю ознак. Як буде показано в розділі 2, навіть у просторі розмірністю 10 при ідеальному поділі на кластери отримуємо значення 0,45.

Davies-Bouldin Index (DBI) [6] був винайдений в 1979 році. Він також показує наскільки добре поділені кластери, але в інший спосіб. Для кожного кластера розраховується середня відстань від об'єктів в кластері до їх центру. При чому формула (1.4) при різних значеннях *q* відповідає різним метрикам відстані. При *q=1* розраховується манхеттенська відстань, при *q=2* - евклідова, більші значення відповідають метриці Мінковського.

$$S_i = \left(\frac{1}{T_i} \sum_{j=1}^{T_i} ||X_j - A_i||_p^q \right)^{1/q} \tag{4}$$

Після цього розраховується відстань між центрами кластерів. При чому так саме можуть бути використані різні метрики дистанції

$$M_{i,j} = ||A_i - A_j||_p = \left(\sum_{k=1}^{n} |a_{k,i} - a_{k,j}|^p \right)^{\frac{1}{p}} \tag{5}$$

І фінальне значення метрики розраховується за формулою:

$$DB \equiv \frac{1}{N} \sum_{i=1}^{N} D_i \quad , \tag{6}$$

де $D_i \equiv \max_{j \neq i} R_{i,j}$ - найбільша несхожість пари кластерів,

$R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$ - несхожість кластера *i* та *j*.

Хоча в метрику і закладено можливість використання різник метрик дистанції для застосування в просторах більших порядків, реалізація в бібліотеці scikit-learn [7] використовує лише евклідову дистанцію.

Кращому розбиттю на кластери відповідає менше значення, в ідеальному випадку близьке до 0. Більші значення відповідають гіршим варіантам розбиття об'єктів на кластери.

Таблиця 1
Table 1

Залежність евклідової відстані для двох пар точок від зміни значень координат та розмірності простору
Dependence of the Euclidean distance for two pairs of points on the change in coordinate values and the dimension of space

№ зп	Розмірність простору	ЕВ I пара	ЕВ II пара, зміна значень		
			0.05	0.1	0.15
1	10	2	0.158114	0.316228	0.474342
2	100	2	0.5	1	1.5
3	300	2	0.866025	1.732051	2.598076
4	1024	2	1.6	3.2	4.8
5	2048	2	2.262742	4.525483	6.788225
6	4096	2	3.2	6.4	9.6

Інші метрики [3] в різний спосіб визначають внутрішньокластерну відстань та міжкластерну відстань та фінальну формулу для поєднання цих відстаней.

В роботі будемо орієнтуватись саме на Silhouette score і Davies-Bouldin Index як на метрики якості, що найбільш часто застосовуються на практиці.

2.3. Розробка метрики якості кластеризації

Будь-яка метрика кластеризації має показати наскільки добре кластери розділені в просторі.

Метрика буде базуватись на міжкластерній дистанції та внутрішньокластерній дистанції.

Міжкластерна дистанція має показати наскільки далеко кластери розійшлись в просторі. Крім того, кластери можуть містити різну кількість об'єктів. Візьмемо за цю характеристику середньовзважену відстань центрів кластерів від загального центру датасету:

$$bcd = \frac{\sum_{i=1}^{k} \|z_i - z_{all}\| \cdot n_i}{n}, \quad (7)$$

де n – загальна кількість об’єктів в датасеті,

z_i – центр окремого кластера,

z_{all} - загальний центр датасету,

n_i – кількість об’єктів в i-му кластері.

Внутрішньокластерна дистанція має показати наскільки далеко об’єкти всередині кластера знаходяться від центру свого кластеру. Візьмемо середньоарифметичне значення відстані від кожного об’єкту до центру свого кластеру. Також при приблизно однакових значеннях метрика повинна показати, що розбиття на меншу кількість кластерів є доцільнішим. Тому отримані значення поділимо на кількість кластерів. Також, для компенсації збільшення розмірності векторного простору, поділимо на корінь квадратний із довжини вектору.

$$wcd = \frac{1}{k \cdot \sqrt{n_{dim}}} \cdot \sum_{i=1}^{k} \left(\frac{1}{n_{i_i}} \cdot \sum_{x \in c_i} \|x - z_i\| \right), \quad (8)$$

де k – кількість кластерів,

c_i – кластер,

z_i – центр кластера,

x – об’єкт в кластері,

n_i – кількість об’єктів в кластері c_i,

n_{dim} – розмірність векторного простору.

Таким чином, обидві складові містять середні значення відстаней, не залежать від кількості точок в кластерах і датасеті, враховують розмірність простору.

Метрика якості кластеризації має задовільняти наступним вимогам:

- збільшуватись, коли відстань між центрами кластерів збільшується;
- зменшуватись, коли середня відстань всередині кластерів збільшується;
- бути обмеженою в діапазоні.

Цим вимогам відповідає функція

$$cl_{score} = 1 - \frac{1}{e^{e^{(bcd - wcd)}}} \quad (9)$$

В кращому випадку bcd має бути значно більше, ніж wcd. Це означатиме, що кластери відокремлені один від одного. Значення cl_{score} буде близьким до 1. В гіршому випадку bcd менше, ніж wcd. Це означає, що кластери завеликі для центрів (наприклад, об’єднано два відокремлених кластери в один). Тоді значення cl_{score} буде близьким до 0.

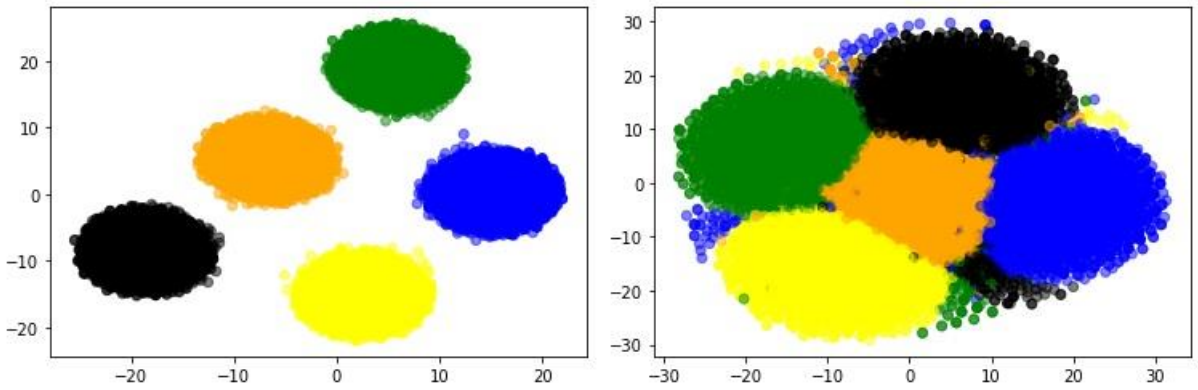
2.4. Перевірка ефективності використання розробленої метрики кластеризації на синтетичних даних

Для дослідження методів кластеризації об’єктів в багатовимірному просторі створимо два набори датасетів (рис.1). В першому наборі групи об’єктів будуть розташовані на значній відстані один від одного. Таке розташування називається розділеним (well-separated). В другому наборі відстань між кластерами майже відсутня. Таке розташування називається центроїдним (center-based). Всередині кожного набору згенеруємо датасети, що складаються із векторів довжиною 10, 100, 300, 1024, 2048, 4096. Кожен датасет має 5 груп по 4000 об'єктів. Фінальний датасет підчищається від викидів і містить від від 19500 до 19800 об'єктів.

Для кожного датасету було проведено алгомеративну кластеризацію із поділом на 2, 3, 4, 5, 6, 7 кластерів і розраховані значення cl_{score}.

3. Результати

Для першого набору датасетів, в якому об’єкті явно виділенні в кластери, отримані значення cl_{score} зведені в таблицю 2.



а) б)

Рис. 1. Візуалізація розташування об’єктів в датасетах при розділених даних (а) і центроїдному розташуванні (б).

Fig. 1. Visualization of the location of objects in datasets with separated data (a) and centroid location (b).

Таблиця 2

Table 2

Значення метрики cl_{score} для агломеративної кластеризації об’єктів, що добре поділені на 5 кластерів, в просторах розмірністю від 10 до 4096

Values of the cl_{score} metric for agglomerative clustering of objects that are well separated into 5 clusters, in spaces with dimensions from 10 to 4096

№ зп	Кластерів	Значення cl_{score} для векторних просторів розмірністю					
		10	100	300	1024	2048	4096
1	2	0.677	0.694	0.693	0.694	0.692	0.694
2	3	0.721	0.732	0.721	0.734	0.734	0.727
3	4	0.751	0.758	0.761	0.763	0.765	0.768
4	**5**	**0.766**	**0.781**	**0.788**	**0.792**	**0.795**	**0.798**
5	6	0.768	0.783	0.791	0.794	0.798	0.801
6	7	0.771	0.785	0.793	0.797	0.801	0.805

Як бачимо із задних в таблиці 2, при розбитті на меншу кількість кластерів, ніж є реально в датасеті, різниця між значення розробленної метрики істотно змінюється, при розбитті на більше кількість кластерів – зміна значень є значно меншою. Саме ця властивість дозволяє використати метод ліктя для визначення оптимальної кількості кластерів в датасеті (рис. 2).

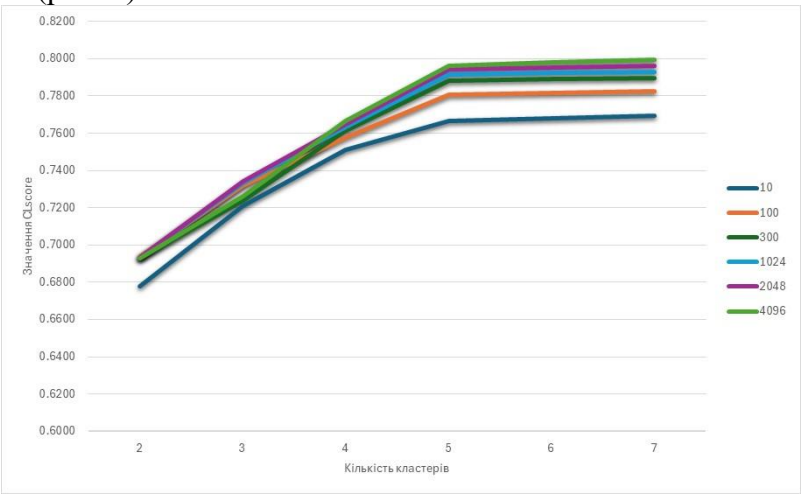


Рис. 2. Залежність значення cl_{score} від кількості кластерів та розмірності даних для добре розділених в просторі кластерів.

Fig. 2. Dependence of the cl_{score} value on the number of clusters and data dimension for well-separated clusters in space.

Для кластерів, що майже не розділені в просторі проведено аналогічні розбиття на кластери, а результати розрахунку розробленої метрики якості наведено в таблиці 3.

Таблиця 3
Table 3

Значення розробленої метрики якості cl_{score} для нерозділених кластерів
The value of the developed clscore quality metric for non-partitioned clusters

№ зп	Кластерів	Значення cl_{score} для розмірностей простору					
		10	100	300	1024	2048	4096
1	2	0.634	0.626	0.624	0.623	0.622	0.623
2	3	0.643	0.641	0.638	0.634	0.637	0.638
3	4	0.651	0.650	0.648	0.648	0.647	0.650
4	**5**	**0.658**	**0.659**	**0.658**	**0.658**	**0.657**	**0.660**
5	6	0.662	0.661	0.660	0.660	0.661	0.665
6	7	0.665	0.664	0.662	0.662	0.663	0.669

Хоча в даному випадку об'єкти в датасеті майже не розділені на окремі кластери, однак і тут можна побачити різницю в значенні приросту метрики для випадків, коли датасет розбито на меншу кількість кластерів, і для випадків, коли датасет розбито на більшу кількість кластерів. Графіки залежності cl_{score} для цього набору датасетів наведено на рис.3.

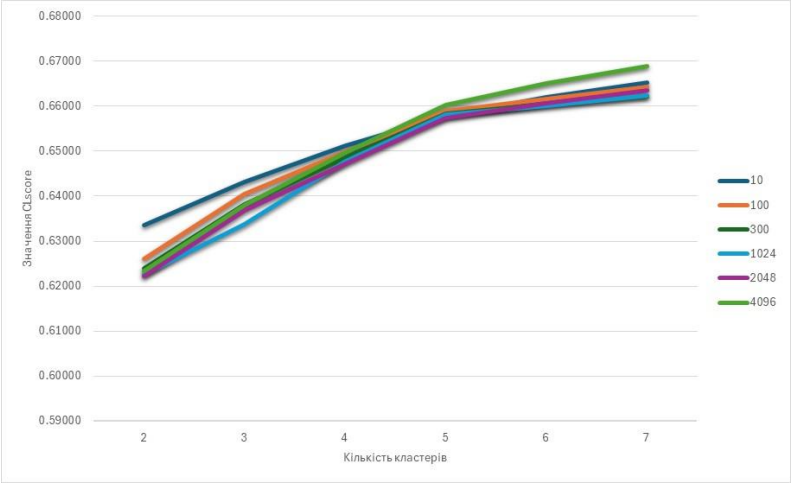


Рис.3. Залежність значення cl_{score} від кількості кластерів та розмірності даних для не розділених в просторі кластерів

Fig. 3. Dependence of the clscore value on the number of clusters and data dimension for clusters not separated in space

З графіків на рис.3 також можна зробити висновок, що метод ліктя для розробленої метрики якості кластеризації може бути застосований навіть у випадках, коли кластери в датасеті є майже не розділеними, тобто структура датасету не явна.

4. Висновки

В статті розроблена нова метрика оцінки якості кластеризації. Показано, що визначити оптимальну кількість кластерів можна за допомогою методу ліктя. Отримані значення метрики доводять, що метрика може бути застосована як для датасетів із невеликою розмірністю (5-10), так і для датасетів із розмірністю 2000-4000, при цьому значення метрики залежать від розташування об’єктів в багатовимірному просторі і майже не залежать від розмірності простору.

Список використаної літератури:

1. LLM Basics: Embedding Spaces - Transformer Token Vectors Are Not Points in Space [Електронний ресурс] — Режим доступу: https://www.lesswrong.com/posts/pHPmMGEMYefk9jLeh/llm-basics-embedding-spaces-transformer-token-vectors-are.
2. Madhulatha TS. An overview on clustering methods. IOSR J Eng. 2012; 2(4): pp. 719–725. Режим доступу: https://doi.org/10.48550/arXiv.1205.1117
3. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J. and Wu, S. Understanding and enhancement of internal clustering validation measures. IEEE Transactions on Cybernetics, 2013; 43(3), pp. 982–994. Режим доступу: https://doi.org/10.1109/TSMCB.2012.2220543
4. Elbow method (clustering) [Електронний ресурс] — Режим доступу: https://en.wikipedia.org/wiki/Elbow_method_(clustering).
5. Silhouette (clustering) [Електронний ресурс] — Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Silhouette_(clustering).
6. Davies–Bouldin index [Електронний ресурс] — Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index.
7. davies_bouldin_score [Електронний ресурс] — Режим доступу до ресурсу: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html.

References:

1. LLM Basics: Embedding Spaces - Transformer Token Vectors Are Not Points in Space [Електронний ресурс] — Режим доступу: https://www.lesswrong.com/posts/pHPmMGEMYefk9jLeh/llm-basics-embedding-spaces-transformer-token-vectors-are.
2. Madhulatha TS. An overview on clustering methods. IOSR J Eng. 2012; 2(4): pp. 719–725. Retrieved from https://doi.org/10.48550/arXiv.1205.1117
3. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J. and Wu, S. Understanding and enhancement of internal clustering validation measures. IEEE Transactions on Cybernetics, 2013; 43(3), pp. 982–994. Retrieved from https://doi.org/10.1109/TSMCB.2012.2220543
4. Elbow method (clustering) [Електронний ресурс] — Retrieved from https://en.wikipedia.org/wiki/Elbow_method_(clustering).
5. Silhouette (clustering) [Електронний ресурс] — Retrieved from https://en.wikipedia.org/wiki/Silhouette_(clustering).
6. Davies–Bouldin index [Електронний ресурс] — Retrieved from https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index.
7. davies_bouldin_score [Електронний ресурс] — Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html.

D.O. STUPAK
P.h.d, Academic Degree,
senior data analyst, EPAM Digital, Cherkasy, Ukraine,
stupak0810@gmail.com

DEVELOPMENT OF OBJECT CLUSTERING QUALITY METRIC FOR MULTIDIMENSIONAL SPACES

DOI: 10.31651/2076-5851-2024-37-45 PACS: 02.50.Sk, 07.05.Mh, 89.75.Fb, 02.70.Rr

In the modern world, new data is generated in geometric progression. Clustering is one of the machine learning methods that does not require labeled data, and therefore makes it possible to quickly determine the structure of the data and draw certain conclusions. The article considers the problem of clustering objects in a multidimensional space. This problem is not new. The concept of the “curse of dimensionality” is critical in clustering, since the algorithm must first divide objects in a multidimensional space into clusters, and then apply clustering quality metrics to find the optimal structure. Existing metrics for assessing the quality of clustering often depend on the dimensionality of the space, and therefore their use under such conditions can be difficult or lead to incorrect results. The aim of the article is to develop a clustering quality metric, the value of which would not depend on the dimensionality of the space in which the objects are described. Two sets of datasets were generated to study clustering. In the first set, the objects are grouped into 5 well-separated clusters, in the second, the clusters almost “touch” each other. Each set contains 6 datasets with a space dimension of 10, 100, 300, 1024, 2048 and 4096. The developed clustering quality metric is based on a comparison of the intercluster characteristic of dividing objects into clusters and the intracluster characteristic. The metric takes into account the dimension of space and the condition of priority of dividing objects into a smaller number of clusters. Numerical experiment methods were used to prove the effectiveness of the application of the

developed clustering quality metric. The test was performed on synthetic datasets that are close in terms of the distribution of objects in existing datasets in practical problems. It is significant that the developed metric of the quality of clustering of objects allows us to determine the correct optimal number of clusters using the “elbow method”, does not depend on the dimensionality of the space and can be applied even in complex cases when the clusters are located close to each other.

Keywords: clustering, multidimensional space, quality metric, elbow method

Одержано редакцією 16.06.2024
Прийнято до друку 22.07.2024