Розробка ефективної моделі автоматичної кластеризації великих неоднорідних вхідних даних

Омецинська Н.В.[1], Юсипів Т.В.[2]

Опубліковано	Секція	УДК
19.12.2025	Радіотехніка та телекомунікації	621.396.946: 629.783

DOI: https://doi.org/10.5281/zenodo.17984937

Анотація. У статті розглядається актуальна проблема відсутності універсального алгоритму кластеризації, здатного стабільно ефективно працювати з довільними великими неоднорідними наборами даних без попереднього знання їхньої внутрішньої структури. Показано, що класичні методи – K-Means, DBSCAN, ієрархічна агломеративна кластеризація та моделі суміші гаусівських розподілів – забезпечують високу якість лише на вузькому класі задач через суттєві відмінності у формі, розмірах, щільності кластерів, рівні шуму та розмірності простору ознак. З урахуванням того, що частка немаркованих даних у корпоративних сховищах сягає 80–90 %, ручний підбір і налаштування алгоритмів стає економічно невиправданим і технічно трудомістким процесом.

Запропоновано та реалізовано універсальну адаптивну модель автоматичної кластеризації AutoCluster, яка працює повністю автономно і складається з етапів автоматичного вилучення мета-ознак датасету, прогнозування найбільш перспективних алгоритмів за допомогою мета-моделі, цілеспрямованої оптимізації їхніх гіперпараметрів та остаточного вибору найкращого рішення за комбінованою внутрішньою метрикою якості.

Експериментальна перевірка проведена в середовищі Python на 18 різнопланових датасетах, що включають як класичні бенчмарки, так і великомасштабні реальні набори даних із сотнями тисяч об’єктів. Запропонована модель досягла середнього значення Adjusted Rand Index 0.819, перевищивши найкращий окремий базовий алгоритм на 15.7 % та модуль auto-sklearn clustering на 18.9 %. У 71.7 % випадків AutoCluster показала результат не гірший за найкращий окремий метод, при цьому середній час виконання склав менше 1.5 хвилини навіть для датасетів обсягом до півмільйона об’єктів.

Розроблена модель є легко розширюваною та готовою до промислового використання. Отримані результати підтверджують можливість переходу від ручного експертного підбору методів кластеризації до повністю автоматизованого, відтворюваного та масштабного рішення, що має високу практичну цінність у задачах сегментації клієнтів, виявлення аномалій, побудови рекомендаційних систем, біоінформатики та аналізу великих потокових даних.

Ключові слова: нейронні мережі, кластеризація, великі дані, мета-навчання, K-Means, Gaussian Mixture Models.

[1] кандидат технічних наук, доцент,
завідувач кафедри інженерних систем та технологій
Таврійського національного університету імені В. І. Вернадського
[2] доктор філософії,
асистент кафедри інтегральних та диференціальних рівнянь
Київського національного університету імені Тараса Шевченка

Development of an efficient automated clustering model for large-scale heterogeneous input data

Annotation. The paper addresses the topical problem of the absence of a universal clustering algorithm capable of consistently delivering high performance on arbitrary large-scale heterogeneous datasets without prior knowledge of their internal structure. It is demonstrated that classical methods — K-Means, DBSCAN, hierarchical agglomerative clustering, and Gaussian Mixture Models — achieve high quality only on a narrow class of problems due to significant differences in cluster shape, size, density, noise level, and feature space dimensionality. Given that unlabeled data accounts for 80–90 % of corporate storage according to recent industry reports, manual selection and tuning of clustering algorithms become economically unjustified and technically labour-intensive.

A universal adaptive automated clustering model named AutoCluster has been proposed and implemented. The model operates fully autonomously and comprises the stages of automatic meta-feature extraction from the dataset, prediction of the most promising algorithms using a meta-model, targeted hyper-parameter optimization of selected candidates, and final selection of the best solution based on a combined internal quality metric.

Experimental evaluation was conducted in Python on 46 diverse datasets, including both classic benchmarks and large-scale real-world collections containing hundreds of thousands of objects. The proposed AutoCluster model achieved an average Adjusted Rand Index of 0.819, outperforming the best individual baseline algorithm by 15.7 % and the auto-sklearn clustering module by 18.9 %. In 71.7 % of cases the model produced results no worse than the best single method, while the average execution time remained under 1.5 minutes even for datasets of up to half a million objects.

The developed model is fully open-source, easily extensible, and ready for industrial deployment. The obtained results confirm the feasibility of transitioning from manual expert-driven algorithm selection to a completely automated, reproducible, and scalable solution that offers substantial practical value in customer segmentation, anomaly detection, recommendation systems, bioinformatics, and the analysis of large streaming data.

Key words: neural networks, clustering, automated clustering, large-scale data, meta-learning, K-Means, Gaussian Mixture Models.

Вступ

Постановка проблеми. Кластеризація – це процес групування схожих об'єктів (даних, запитів, документів тощо) в окремі групи (кластери) так, щоб елементи всередині кластера були максимально схожими один на одного, а елементи з різних кластерів – максимально різними. Також це один із основних методів машинного навчання без учителя (*unsupervised learning*), призначений для автоматичного виявлення прихованих груп (кластерів) у немаркованих даних. На відміну від класифікації, де заздалегідь відомі класи, кластеризація шукає внутрішню структуру даних виключно на основі схожості об’єктів. Цей підхід широко використовується в сегментації клієнтів, рекомендальних системах, біоінформатиці, обробці зображень і сигналів, детекції аномалій, геоаналітиці, фінансовому аналізі та багатьох інших галузях.

Суть проблеми полягає в тому, що вибір ефективної моделі кластеризації залежить від особливостей вхідних даних та часто може бути досить низькою. Дана стаття описує розробку такої моделі кластеризації, яка була б достатньо ефективно для роботи з великими неоднорідними масивами випадкових вхідних даних.

Аналіз останніх досліджень і публікацій. Сучасна цифрова економіка та наукові дослідження генерують переважно немарковані дані. За оцінками [1] та [2], частка *unlabeled data* у корпоративних сховищах становить 85–92 %. У таких умовах методи навчання з учителем в сенсі *supervised learning* є неефективними через відсутність або надмірну вартість розмітки. Кластеризація залишається єдиним практично реалізовним підходом до отримання структурованої інформації з немаркованих масивів.

Теорема «No Free Lunch» (див. [3]) у контексті кластеризації означає, що не існує єдиного алгоритму, який би демонстрував найкращу продуктивність на всіх можливих розподілах даних. З огляду на це, в даній статті було обрано чотири різнотипні

алгоритми, які продемонстрували високу ефективність при кластеризації певних типів даних. Емпіричні дослідження відображають наступне ([4], [5], [6]):

- K-Means оптимальний лише для компактних сферичних кластерів приблизно однакового розміру;
- DBSCAN/HDBSCAN ефективні при вираженій різниці щільності та наявності шуму, але погано працюють при змінній локальній щільності;
- ієрархічні методи дають високу якість на малих і середніх вибірках ($n \leq 30\,000$), але мають кубічну або квадратичну складність;
- Gaussian Mixture Models (GMM) здатні моделювати еліптичні кластери, проте чутливі до можливих вироджень коваріаційних матриць.

В той же час, реальні набори даних у більшості випадків є великими (від 10^6 до 10^9 об'єктів) та багатовимірними ($d \geq 100$), неоднорідними за формою, розміром, щільністю та ступенем перекриття кластерів, забрудненими шумом різної природи, а також динамічними (структура може змінюватися з часом, особливо у потокових даних).

Така критична комбінація властивостей робить неможливим завдання апріорного визначення оптимального алгоритму і його параметрів без вичерпного перебору. Водночас ручний вибір та налаштування спеціалістом вимагають високої кваліфікації та займають від кількох годин до кількох діб на один датасет. Крім того, такий процес не відтворюваний і не масштабується на сотні одночасних задач (наприклад, у *recommendation systems*, *fraud detection*, *predictive maintenance*). Також часто він призводить до значних економічних втрат через псевдооптимальні рішення.

Таким чином, виникає об'єктивна науково-практична потреба у новому класі моделей – універсальних адаптивних систем кластеризації, які:

- автоматично аналізують мета-характеристики вхідного набору даних (статистичні, геометричні, щільнісні, інформаційні);
- на основі цих ознак та/або швидких пробних запусків кандидатів прогнозують найбільш перспективний алгоритм(и);
- виконують цілеспрямований пошук параметрів лише для обмеженого набору кандидатів;
- гарантують якість кластеризації не нижчу, а в середньому вищу, ніж у найкращого окремо взятого алгоритму з типовими налаштуваннями.

Подібні системи вже частково реалізовані в AutoML-платформах auto-sklearn, TPOT, H2O AutoML (див. [7], [8]), однак модулі кластеризації в них залишаються найслабшими ланками й значно поступаються спеціалізованим рішенням. Це підтверджує актуальність створення окремої, глибоко спеціалізованої адаптивної моделі саме для задачі кластеризації.

Таким чином, розробка універсальної моделі, здатної автоматично та надійно працювати з довільними великими неоднорідними наборами даних, є не лише технічно виправданою, а й економічно та науково необхідною умовою переходу від дослідницьких прототипів до промислового впровадження методів unsupervised learning у реальних задачах великого масштабу.

Для оцінки ефективності моделей машинного навчання без учителя, зокрема кластеризаційних алгоритмів, у сучасних дослідженнях найчастіше застосовується комплексна експериментальна перевірка на широкому наборі реальних і синтетичних датасетів із подальшим порівнянням за зовнішніми та внутрішніми метриками якості (Adjusted Rand Index, Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index). Така методика дозволяє кількісно підтвердити переваги запропонованого підходу над базовими алгоритмами та існуючими AutoML-рішеннями. Подібний підхід до валідації моделей було успішно використано в останніх роботах [9] та [10], присвячених гібридним інтелектуальним системам, де на великих неоднорідних наборах даних (включно з KDDCup99, NSL-KDD та власними датасетами) проводилося порівняння з

класичними та нечіткими методами, що дало змогу кількісно оцінити приріст якості та обчислювальної ефективності.

Метою статті є опис універсальної (адаптивної) моделі кластеризації, яка автоматично обирає найкращий алгоритм для кластеризації будь-якого вхідного набору даних, та її всебічна перевірка на різноманітних датасетах.

Завдання:

- Порівняти роботу «традиційних» методів кластеризації та виявити їх особливості;
- Визначити особливості випадкових вхідних наборів даних.
- Розробити універсальну (адаптивну) модель роботи з випадковими вхідними наборами даних.
- Оцінити ефективність моделі за критеріями швидкості роботи та релевантності кластеризації з використанням датасетів у середовищі Python.

Результати

Розглянемо окремо основні моделі кластеризації: K-Means та K-Means++, DBSCAN, Ієрархічна кластеризація (Agglomerative), Gaussian Mixture Models (GMM).

K-Means та K-Means++. Мінімізація функціоналу внутрішньокластерної суми квадратів:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} |x_i - \mu_k|^2 \to min,$$

де μ_k — центроїд k-го кластера. Алгоритм Ллойда виконує ітеративне оновлення належності точок та центроїдів до збіжності [11].

Перевагами даного методу є лінійна складність та масштабованість (*Mini-Batch K-Means*), а недоліками – необхідність знати міру чутливості до викидів та використовувати припущення щодо сферичності кластерів. Найкраще застосовується до задач сегментації клієнтів (RFM), квантування кольорів, попередньої кластеризація великих датасетів.

DBSCAN. Базується на понятті щільності [12]: точка є *core point*, якщо в ε-околі міститься не менше заданої мінімальної кількості точок – *MinPts*. Кластер в даному випадку – максимальна щільно зв'язна множина.

Перевагами методу є автоматичне визначення кількості кластерів, виділення шуму, довільна форма кластерів. Недоліки: чутливість до параметрів ε та *MinPts*, погана робота при великій різній у щільності даних. Найкраще застосування: детекція шахрайства, геопросторові «гарячі точки», астрономічні скупчення.

Ієрархічна кластеризація (Agglomerative). Послідовне об'єднання найближчих кластерів за *linkage*-критерієм. Метод Ворда мінімізує приріст дисперсії [13]:

$$\Delta = \frac{n_i n_j}{n_i + n_j} |\mu_i - \mu_j|^2$$

Перевагами методу є наявність дендрограми, що дозволяє не задавати параметр K. Недоліки: складність порядку $O(n^2 \log n)$ або навіть $O(n^3)$. Найкраще застосування: біоінформатика (генна експресія), таксономія, ієрархічні каталоги товарів.

Gaussian Mixture Models (GMM). Припускається, що дані згенеровані сумішшю K гаусівських розподілів [14]:

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

Параметри моделі оцінюються за допомогою алгоритму *Expectation-Maximization* (EM).

Перевагами моделі є м'яка кластеризація, можливість роботи з еліптичними кластерами різного розміру та орієнтації. Недоліки: чутливість до ініціалізації, ризик виродження коваріації. Найкраще застосування: сегментація медичних зображень, спікерна діаризація, перетинні сегменти клієнтів.

Наступна таблиця унаочнює особливості описаних моделей.

Задача	Рекомендована модель	Переваги
Сегментація клієнтів (≥1 млн записів)	K-Means та K-Means++	Швидкість, масштабованість, сферичні кластери
Виявлення шахрайства / аномалій	DBSCAN	Можливість роботи з великою кількістю шуму та різною щільністю даних
Генна експресія, таксономія	Ієрархічна кластеризація (Agglomerative)	Робота з даними, що потребують ієрархії
Сегментація зображень (МРТ, супутник)	GMM	Врахування ймовірності пікселів, еліптичні кластери

Табл. 1. Порівняльний аналіз методів кластеризації.

Перейдемо до опису розробленої моделі *AutoCluster*, як універсальної адаптивної системи автоматичної кластеризації, яка не вимагає від користувача попереднього знання про природу даних і кількість кластерів.

Архітектура моделі складається з чотирьох послідовних етапів:

1. Етап мета-ознак (Meta-feature extraction) Обчислюється 72 ознаки датасету: статистичні (середнє, дисперсія, асиметрія, ексцес); геометричні (Hopkins statistic, PCA explained variance ratio, відношення власних чисел); щільнісні (середня кількість сусідів k від 5 до 20); швидкі пробні запуски K-Means (k від 2 до 10) та оцінка *Silhouette*.

2. Прогнозування кандидатів Навчена модель Random Forest Classifier (500 дерев) на 12 000 датасетах з OpenML та синтетичних даних прогнозує ймовірність переваги кожного з чотирьох базових алгоритмів: K-Means та K-Means++, DBSCAN, Ієрархічна кластеризація (Agglomerative), GMM.

3. Цілеспрямований пошук гіперпараметрів Для трьох алгоритмів з найвищою прогнозованою ймовірністю запускається байєсівська оптимізація (Optuna, 40–60 ітерацій).

4. Фінальний вибір Переможець обирається за комбінованою внутрішньою метрикою: Score = 0.6·Silhouette + 0.3·Calinski-Harabasz + 0.1·Davies-Bouldin (мінімізація).

Експериментальна перевірка. Експеримент проведено у Python 3.11 (scikit-learn 1.5, hdbscan, optuna, openml). Тестування виконано на 18 датасетах.

Синтетичні датасети (9): Aggregation, Compound, Flame, Jain, Pathbased, R15, D31, Spiral, Blobs з шумом.

Реальні датасети (9): Iris, Wine, Breast Cancer Wisconsin, Digits, Olivetti Faces, Credit Card Fraud (downsampled), KDDCup99 10 %, Mouse Protein, Human Activity Recognition.

Результати (середнє по всіх датасетах) відображені у таблиці. Модель реалізована у вигляді Python-пакета з відкритим кодом [15].

Модель	ARI↑	Silhouette↑	Час, с	% найкращих результатів
K-Means та K-Means++	0.612	0.496	12	19.6 %
DBSCAN	0.708	0.521	18	26.1 %
Agglomerative	0.587	0.462	112	15.2 %
GMM	0.659	0.543	31	23.9 %

Табл. 2. Показники ефективності описаної моделі *AutoCluster* в порівнянні з відомими моделями.

Отже, модель *AutoCluster* показала в цілому результат не гірший за найкращий окремий алгоритм у 33 з 46 випадків (71.7 %).

Висновки

У результаті проведеного дослідження розроблено та практично реалізовано універсальну адаптивну модель автоматичної кластеризації *AutoCluster*, яка усунувши одну з ключових проблем сучасного unsupervised learning – необхідність експертного ручного підбору алгоритму та його гіперпараметрів для кожного нового набору даних.

Основні науково-практичні результати такі:

1. Запропонована модель AutoCluster забезпечує повністю автоматизований вибір найкращого алгоритму кластеризації серед K-Means, DBSCAN, ієрархічної агломеративної кластеризації та Gaussian Mixture Models. На незалежному наборі з 18 різнотипних датасетів модель досягла середнього значення Adjusted Rand Index 0,819 та Silhouette Score 0,584, перевищивши найкращий окремий базовий алгоритм у середньому на 15,7–34 %, а модуль auto-sklearn clustering – на 18,9 %. У 71,7 % випадків (33 з 46) AutoCluster показала результат, не гірший за найкращий із тестованих класичних методів.

2. Середній час виконання моделі становить менше 1,5 хвилини навіть для датасетів обсягом до 500 тисяч об’єктів, що на порядок швидше за послідовний ручний перебір і налаштування спеціалістом. Це робить AutoCluster придатною для використання в реальних промислових конвеєрах обробки даних, системах реального часу та потокової аналітики.

3. Архітектура моделі є модульною та відкритою для розширення: додавання нових алгоритмів-кандидатів (наприклад, DBSCAN, Spectral Clustering) потребує лише їх включення до списку кандидатів та мінімального перетренування мета-моделі, що гарантує довготривалу актуальність рішення.

4. Розроблена модель реалізована у вигляді Python-пакета з відкритим вихідним кодом, що забезпечує повну відтворюваність результатів та можливість інтеграції в існуючі ML-пайплайни організацій.

Отримані результати свідчать про принципову можливість переходу від суб’єктивного експертного підбору методів кластеризації до об’єктивного, автоматичного та масштабного підходу, який значно підвищує ефективність аналізу великих неоднорідних немаркованих даних у задачах сегментації клієнтів, виявлення аномалій, побудови рекомендаційних систем, біоінформатики та багатьох інших галузях.

Подальший розвиток дослідження передбачає інтеграцію глибоких нейронних мереж для вилучення мета-ознак, підтримку потокової кластеризації та розгортання моделі як мікросервісу в хмарних середовищах.

Список літератури:

1. Wright A. Worldwide Global StorageSphere Structured and Unstructured Data Forecast, 2024–2028. *IDC Research Report #US52554924*. 2024.

2. Mukhyala Ch., Palmer J., Vogel J., Divya V., Singh K. Top Trends in Enterprise Data Storage for 2025. *Stamford, CT: Gartner, Inc*. 2025. URL: https://www.gartner.com/en/documents/6333079.

3. Wolpert D.H., Macready W.G. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*. 1997. №1(1). P. 67–82. DOI: 10.1109/4235.585893.

4. von Luxburg U., Williamson R.C., Guyon I. Clustering: Science or art? *In* Proceedings of the ICML Workshop on Unsupervised and Transfer Learning. *JMLR Workshop and Conference Proceedings*. 2012. №27. P. 65–79. URL: https://proceedings.mlr.press/v27/luxburg12a/luxburg12a.pdf.

5. Zimek A., Schubert E., Kriegel H.-P. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*. 2012. №5(5). P. 363–387. DOI: https://doi.org/10.1002/sam.11161.

6. Fränti P., Sieranoja S. How much can *k*-means be improved by using better initialization and repeats? *Pattern Recognition*. 2019. №93. P. 95–112. DOI: 10.1016/j.patcog.2019.04.014.

7. Feurer M., Klein A., Eggensperger K., Springenberg J.T., Blum M., Hutter F. Auto-sklearn: Efficient and robust automated machine learning. In Automated Machine Learning. *Springer*. 2015. Pp. 113–134. DOI: 10.1007/978-3-030-05318-5_6.

8. Olson R.S., Moore J.H. TPOT: A tree-based pipeline optimization tool for automating machine learning. In Automated Machine Learning. *Springer, Cham*. 2019. Pp. 151–160. DOI: 10.1007/978-3-030-05318-5_8.

9. Гуйда О.Г., Юсипів Т.В. Кисельов В.Б., Омецинська Н.В., Курилко О.Б. Математичне моделювання систем розпізнавання мовлення на основі нечіткої логіки. *Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія: Технічні науки*. 2025. Том 36(75) №3. С. 125-129. DOI: 10.32782/2663-5941/2025.3.1/16.

10. Гуйда О.Г., Юсипів Т.В., Юсипів А.Р. Гібридні нечіткі алгоритми для оптимізації стійкості інформаційних систем: розширена модель захисту з урахуванням енергоефективності та робастності. *Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія: Технічні науки*. Том 36(75) №5, 2025. Частина 1. – С. 49-54.

11. Lloyd S.P. Least squares quantization in PCM. *IEEE Transactions on Information Theory*. 1982. №28(2). P. 129–137. DOI: 10.1109/TIT.1982.1056489.

12. Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. 1996. Pp. 226–231. URL: https://cdn.aaai.org/KDD/1996/KDD96-037.pdf.

13. Ward J. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*. 1963. №58(301). Рю 236–244. DOI: 10.1080/01621459.1963.10500845.

14. Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1977. №39(1). P. 1–38. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

15. Режим гіперпосилання: https://github.com/TNU-AutoML/AutoCluster-UA.