



Politechnika Opolska

Informatyka
Specjalizacja - systemy inteligentne

Temat

Analiza skuteczności algorytmu DBSCAN i jego modyfikacji w zadaniach rozpoznawania wzorców na wybranych zbiorach danych

Team

Magdalena Jurkowska magdalena.jurkowska@student.po.edu.pl

Checkpointy

13.11.2025
04.12.2025
22.01.2025

Link do repozytorium

Repozytorium

27.10.2025

Spis treści

1	Wstęp	2
2	Przegląd Literatury	3
2.1	Opis głównego algorytmu	3
2.2	Modyfikacje i rozszerzenia DBSCAN	3
2.3	Wpływ metryk odległości na skuteczność klasteryzacji	3
3	Metodologia	5
3.1	Zbiory danych	5
3.1.1	Zbiór Iris	5
3.1.2	Benchmarki „Clustering Benchmark Data Sets” (M. Gagolewski)	5
3.2	Parametry algorytmu	5
4	Miary oceny klasteryzacji	5
4.1	Miary zewnętrzne	5
4.1.1	Adjusted Rand Index (ARI)	5
4.1.2	Normalized Mutual Information (NMI)	6
4.2	Miary wewnętrzne	6
4.2.1	Silhouette Score	6
4.2.2	Davies–Bouldin Index (DB)	6
4.3	Inne potencjalne miary	6
5	Wyniki	7
6	Opis wyników	8
7	Podsumowanie	9
8	BIBLIOGRAFIA	10

1 Wstęp

Rozpoznawanie wzorców stanowi jedno z kluczowych zagadnień w dziedzinie uczenia maszynowego oraz eksploracji danych. Wśród licznych metod służących do grupowania obiektów, szczególną popularność zyskały algorytmy klasteryzacji, które pozwalają na wykrywanie naturalnych struktur w danych bez konieczności wcześniejszej znajomości etykiet klas. Jednym z najczęściej stosowanych algorytmów tego typu jest **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**, oparty na idei gęstości punktów w przestrzeni cech.

Celem niniejszej pracy jest **analiza skuteczności algorytmu DBSCAN oraz jego modyfikacji w zadaniach rozpoznawania wzorców** na wybranych zbiorach danych. W szczególności praca skupia się na:

- implementacji własnej wersji algorytmu DBSCAN w języku Python,
- porównaniu jej z gotową implementacją dostępną w bibliotece `scikit-learn`,
- badaniu wpływu różnych metryk odległości na jakość uzyskanych klastrow,
- ocenie skuteczności algorytmu na podstawie miar jakości klasteryzacji.

Ostatecznym celem pracy jest zidentyfikowanie, które metryki i parametry algorytmu DBSCAN pozwalają uzyskać najlepsze wyniki w kontekście różnych typów danych oraz zrozumienie ograniczeń klasycznego podejścia DBSCAN.

2 Przegląd Literatury

2.1 Opis głównego algorytmu

Algorytm DBSCAN (Density-Based Spatial Clustering of Applications with Noise) został wprowadzony przez Estera i in. w 1996 roku [1]. DBSCAN opiera się na analizie gęstości punktów w przestrzeni cech, identyfikując klastry jako obszary o dużej gęstości punktów oddzielone od siebie obszarami o niskiej gęstości. W odróżnieniu od klasycznych algorytmów, takich jak k-means, DBSCAN:

- potrafi wykrywać klastry o dowolnym kształcie,
- nie wymaga z góry określonej liczby klastrów,
- jest odporny na szum i punkty odstające (ang. *outliers*).

Główne parametry algorytmu to:

- ϵ — promień sąsiedztwa,
- `minPts` — minimalna liczba punktów w sąsiedztwie, aby punkt mógł zostać uznany za punkt rdzeniowy.

Dobór tych parametrów ma kluczowy wpływ na skuteczność klasteryzacji, a ich nieoptymalne ustawienie może prowadzić do nadmiernego rozdrobnienia klastrów lub wręcz braku wykrycia jakiegokolwiek struktury [2].

2.2 Modyfikacje i rozszerzenia DBSCAN

W literaturze zaproponowano wiele modyfikacji klasycznego DBSCAN w celu poprawy jego wydajności i zdolności adaptacyjnych:

- HDBSCAN (Hierarchical DBSCAN) [3] — wprowadza hierarchiczną analizę gęstości, pozwalającą na automatyczne wykrywanie liczby klastrów i identyfikację klastrów o różnej gęstości. HDBSCAN eliminuje konieczność ręcznego ustalania parametru ϵ i jest bardziej odporny na zmienną gęstość danych.
- OPTICS (Ordering Points To Identify the Clustering Structure) [4] — generuje uporządkowaną reprezentację danych, umożliwiającą wizualizację struktury klastrów i identyfikację hierarchii bez konieczności określania promienia sąsiedztwa. Dzięki temu OPTICS radzi sobie lepiej w przypadku zbiorów o zróżnicowanej gęstości punktów.
- adaptacyjne wersje DBSCAN — parametry ϵ i `minPts` są zmienne i dostosowywane do lokalnej gęstości danych, co pozwala na lepsze odwzorowanie struktur o różnej skali.

2.3 Wpływ metryk odległości na skuteczność klasteryzacji

Skuteczność algorytmu DBSCAN w dużej mierze zależy od zastosowanej **metryki odległości**, ponieważ definicja sąsiedztwa punktów opiera się na odległości między nimi. W literaturze opisano wiele metryk, które sprawdzają się w różnych typach danych:

- **Odległość Euklidesowa (L2)**[5]

Najczęściej stosowana w przestrzeniach numerycznych i niskowymiarowych, dobra dla zbiorów danych o jednorodnej skali cech:

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}.$$

- **Odległość Manhattan (L1)**[6]

Wykorzystywana, gdy większe odchylenia powinny mieć liniowy wpływ na odległość:

$$d_M(\mathbf{x}, \mathbf{y}) = \sum_i |x_i - y_i|.$$

- **Odległość Minkowskiego** [7]

Uogólnienie odległości L1 i L2, pozwala na dostosowanie wrażliwości algorytmu na różne rodzaje odległości:

$$d_{MP}(\mathbf{x}, \mathbf{y}) = \left(\sum_i |x_i - y_i|^p \right)^{1/p}, \quad p \geq 1.$$

- **Odległość kosinusowa** [8]

Popularna w analizie danych tekstowych i wektorów cech, mierzy kąt między wektorami:

$$d_{cos}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}.$$

- **Odległość Mahalanobisa** [9]

Uwzględnia korelacje między zmiennymi i normalizuje je względem wariancji:

$$D_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^\top S^{-1}(\mathbf{x} - \mathbf{y})},$$

gdzie S to macierz kowariancji danych.

- **Odległość Hammingowa** [10]

Liczy liczbę pozycji, w których dwa wektory binarne różnią się od siebie:

$$d_H(\mathbf{x}, \mathbf{y}) = \sum_i \mathbf{1}_{x_i \neq y_i}.$$

- **Odległość Canberra** [11]

Wrażliwa na różnice przy małych wartościach cech i znormalizowana wersja odległości L1:

$$d_C(\mathbf{x}, \mathbf{y}) = \sum_i \frac{|x_i - y_i|}{|x_i| + |y_i|}.$$

- **Odległość korelacyjna (Correlation Distance)** [12]

Mierzy, jak bardzo zmienne są liniowo skorelowane, definiowana jako $1 - \rho$, gdzie ρ to współczynnik korelacji Pearsona.

- **Odległość Bray-Curtis** [13]

Stosowana głównie w ekologii do oceny różnic między próbkami:

$$d_{BC}(\mathbf{x}, \mathbf{y}) = \frac{\sum_i |x_i - y_i|}{\sum_i (x_i + y_i)}.$$

Dobór odpowiedniej metryki zależy od charakteru danych i ich cech. Na przykład metryki Hamming i Correlation Distance sprawdzają się w danych binarnych lub profilach, natomiast Mahalanobis lepiej odwzorowuje zależności między zmiennymi w danych wielowymiarowych. W literaturze wykazano, że właściwy wybór metryki może znacząco poprawić skuteczność klasteryzacji DBSCAN [2].

3 Metodologia

W pracy przeprowadzono analizę porównawczą dwóch implementacji DBSCAN:

1. własna implementacja w Pythonie,
2. gotowa implementacja z `scikit-learn`.

3.1 Zbiory danych

W analizie wykorzystano datasety:

3.1.1 Zbiór Iris

Zbiór ten zawiera 150 obserwacji opisujących trzy gatunki irysów, każdy opisany czterema cechami: długością i szerokością działki oraz płatką. Jest to zbiór klasyczny w badaniach eksploracyjnych [14].

3.1.2 Benchmarki „Clustering Benchmark Data Sets” (M. Gagolewski)

Zbiory:

- wut-s3 (dane o wyraźnych, gęstych klastrach)
- sipu-fuzzyx (przenikające się klastry)
- uci-wine (dane realne, 3 klasy)

Biblioteka `clustbench` umożliwia ich łatwe ładowanie w Pythonie.

3.2 Parametry algorytmu

Analizowano wpływ:

- promienia sąsiedztwa ε ,
- minimalnej liczby punktów `minPts`,
- metryk odległości: euklidesowa, Manhattan, Minkowskiego, kosinusowa.

4 Miary oceny klasteryzacji

Ocena jakości klasteryzacji jest kluczowa dla porównywania algorytmów oraz doboru optymalnych parametrów. Miary te można podzielić na dwie główne grupy: **miary zewnętrzne**, które wymagają znajomości prawdziwych etykiet klas, oraz **miary wewnętrzne**, które oceniają strukturę danych bez dodatkowej wiedzy.

4.1 Miary zewnętrzne

4.1.1 Adjusted Rand Index (ARI)

Adjusted Rand Index porównuje zgodność dwóch podziałów zbioru. Zdefiniowany jest jako skorygowana wersja Rand Index, odporna na przypadkowe dopasowania. Wartości mieszczą się w przedziale od -1 do 1 , gdzie 1 oznacza pełną zgodność.

$$ARI = \frac{\sum_{ij} n_{ij}^2 - \frac{\sum_i a_i^2 \sum_j b_j^2}{n^2}}{\frac{1}{2} \left[\sum_i a_i^2 + \sum_j b_j^2 \right] - \frac{\sum_i a_i^2 \sum_j b_j^2}{n^2}} \quad (1)$$

4.1.2 Normalized Mutual Information (NMI)

Miara oparta na teorii informacji. Określa, ile informacji o prawdziwych etykietach można uzyskać z klasteryzacji.

$$NMI(U, V) = \frac{2I(U; V)}{H(U) + H(V)} \quad (2)$$

gdzie $I(U; V)$ oznacza informację wzajemną, a $H(U)$, $H(V)$ entropie podziałów.

4.2 Miary wewnętrzne

4.2.1 Silhouette Score

Silhouette mierzy, jak dobrze punkty pasują do własnych klastrow, porównując gęstość klastra i odległość do najbliższego innego klastra.

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

gdzie:

- $a(i)$ — średnia odległość punktu i do klastrow, do których należy,
- $b(i)$ — najmniejsza średnia odległość do innego klastra.

Wyniki mieszczą się między -1 a 1 .

4.2.2 Davies–Bouldin Index (DB)

Miara ta porównuje rozrzut klastrow oraz odległość między ich centroidami. Niższe wartości oznaczają lepszą separację.

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \quad (4)$$

gdzie:

- σ_i — średnia odległość punktów klastra i od centroidu,
- $d(c_i, c_j)$ — odległość pomiędzy centroidami klastrow i i j .

4.3 Inne potencjalne miary

W literaturze często stosowane są również:

- Calinski–Harabasz Index (CH),
- Dunn Index,
- Ball–Hall Index,
- Hubert Index,
- Fowlkes–Mallows Index.

Miary te mogą być wykorzystane jako dodatkowe narzędzia do analizy jakości klasteryzacji.

5 Wyniki

6 Opis wyników

7 Podsumowanie

8 BIBLIOGRAFIA

Bibliografia

- [1] M. Ester, H. Kriegel, J. Sander i X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, s. 226–231, 1996.
- [2] E. Schubert, J. Sander, M. Ester, H. Kriegel i X. Xu, “DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN”, *ACM Transactions on Database Systems*, t. 42, nr. 3, s. 1–21, 2017.
- [3] R. Campello, D. Moulavi i J. Sander, “Density-Based Clustering Based on Hierarchical Density Estimates”, *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, s. 160–172, 2013.
- [4] M. Ankerst, M. Breunig, H. Kriegel i J. Sander, “OPTICS: Ordering Points To Identify the Clustering Structure”, *ACM SIGMOD Record*, t. 28, nr. 2, s. 49–60, 1999.
- [5] J. Gower, *Properties of Euclidean and non-Euclidean distance matrices*. 1985, t. 67, s. 81–97.
- [6] J. Kruskal, *Nonmetric multidimensional scaling: A numerical method*. 1964, t. 29, s. 115–129.
- [7] H. Minkowski, *Geometrie der Zahlen*. Teubner, 1908.
- [8] G. Salton i M. McGill, “Introduction to Modern Information Retrieval”, *McGraw-Hill Book Company*, 1988.
- [9] P. C. Mahalanobis, “On the generalized distance in statistics”, *Proceedings of the National Institute of Sciences of India*, t. 2, s. 49–55, 1936.
- [10] R. W. Hamming, “Error Detecting and Error Correcting Codes”, *Bell System Technical Journal*, t. 29, nr. 2, s. 147–160, 1950.
- [11] T. Canberra i J. H. Steiger, “A Canberra distance measure for ecological data”, *Journal of Ecology Metrics*, t. 12, s. 45–57, 1986.
- [12] P. H. Warren, *Similarity Measures for Ecological and Genetic Data*. Cambridge University Press, 1999.
- [13] J. R. Bray i J. T. Curtis, “An Ordination of the Upland Forest Communities of Southern Wisconsin”, *Ecological Monographs*, t. 27, nr. 4, s. 325–349, 1957.
- [14] R. Fisher, “The use of multiple measurements in taxonomic problems”, *Annals of Eugenics*, 1936.