DBSCAN の概要と実装結果

大阪大学 工学部 電子情報学科 3 年
情報システム工学コース
08D23091 辻 孝弥

2025 年 4 月 29 日

1　DBSCAN とは

Density-Based Spatial Clustering of Applications with Noise（DBSCAN）は，データ分布の**密度**に着目してクラスタを抽出する手法である．1996 年に Martin Ester らが提案して以来，密度が高い領域をクラスタとしてまとめ，低密度領域の点をノイズ（外れ値）として切り捨てるシンプルさと汎用性から，機械学習分野で広く採用されている．

DBSCAN の特徴は，**クラスタ数を事前に与える必要がない**点にある．データ空間内で互いに近い点が多数存在する高密度領域を自動で検出し，それぞれをクラスタとみなす．周囲に点がほとんど存在しない孤立点はノイズとして扱われるため，異常検知にも応用できる．

2　検索トレンドの動向

図 1 は Google Trends による「DBSCAN」と「k means」の検索数推移を示す．近年，AI・データサイエンスへの関心が高まったことで DBSCAN の検索数も増加傾向にある．k means は古くから用いられてきたが，クラスタ数の事前指定が不要という利点から DBSCAN も同程度に検索されるようになってきたと考えられる．

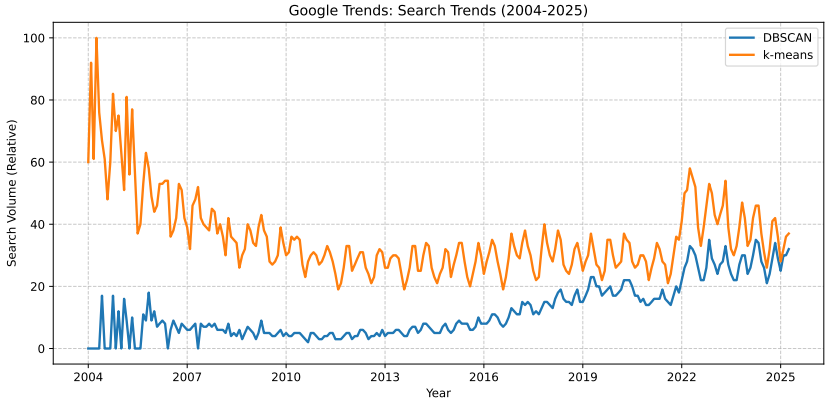

図 1　k means と DBSCAN の検索数比較

3　アルゴリズムの要点

3.1　2 つの主要パラメータ

DBSCAN は，半径 ε の近傍と最小点数 MinPts という 2 つのパラメータで動作する．

ε:　半径 ε 以内に存在する点集合をその点の ε ‐近傍と定義する．

$$N_\varepsilon(p) = \{\, q \mid \mathrm{dist}(p, q) \le \varepsilon \,\}$$

MinPts:　近傍に含まれる点数が MinPts 以上なら，点 p を**コア点**と呼ぶ．

- **コア点**: ε ‐近傍に MinPts 以上の点を持つ点．
- **境界点**: 自身はコア点でないが，あるコア点の ε ‐近傍に含まれる点．
- **ノイズ点**: いずれのコア点からも到達できない点．

3.2　処理手順

1. すべての点を未訪問とする．
2. 未訪問点 p を 1 つ取り出し，$N_\varepsilon(p)$ を求める．
3. $|N_\varepsilon(p)| \ge$ MinPts であれば新しいクラスタを生成し，p および $N_\varepsilon(p)$ をクラスタに割り当てる．
4. クラスタに追加された各点 q について，再帰的に $N_\varepsilon(q)$ を調べ，条件を満たす点をクラスタへ拡張する．
5. $|N_\varepsilon(p)| <$ MinPts の場合，p を一旦ノイズとラベル付けする．
6. 未訪問点がなくなるまで手順 2–5 を繰り返す．

3.3　パラメータ選択の指針

ε が小さすぎるとノイズが増え，大きすぎると異なるクラスタを誤って結合する．適切な値はデータに依存するため，k**距離プロット**でエルボー点を探す方法がよく用いられる．MinPts は次元数 D に対して $D+1$ 以上を目安とするのが経験則である．

4　k means との比較

k means はクラスタを凸形状（球状）と仮定し，クラスタ数を事前に指定する必要がある．非凸形状や異なる密度のクラスタが混在する場合，k means は適切に分割できないことが多い．これに対し DBSCAN は形状の仮定を置かず，密度の差を利用して図 2 のような複雑なクラスタも検出できる．

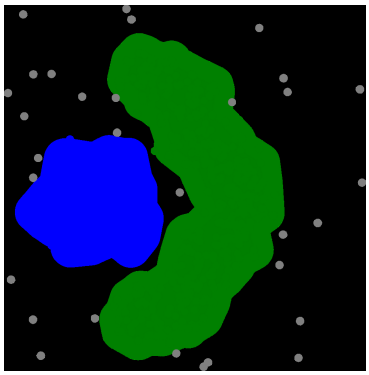

図 2　DBSCAN による非凸クラスタの抽出例

4.1　実装結果

図 3 は DBSCAN による非線形形状のクラスタリングの結果である。k-means 法ではクラスタリング困難だった非線形データセットに対してのクラスタリングが可能となっているのが視覚的にわかる。また, DBSCAN の実装では, 外れ値検出/除外を明示的に別で実装しているわけではなかったが, 図 4 のように孤立点がノイズとして扱われているのがわかる。

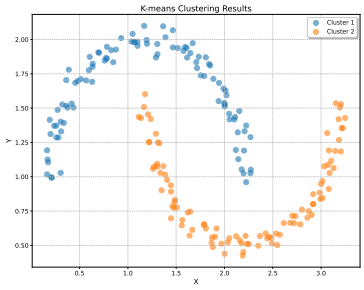

図 3　DBSCAN による非線形形状のクラスタリング

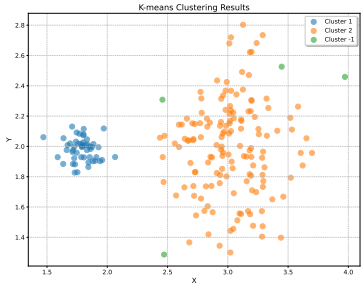

図 4　DBSCAN による孤立点の検出

5　おわりに

DBSCAN は，クラスタ数の事前設定が不要であり，任意形状のクラスタを検出できるという強みを持つ．適切なパラメータ設定が難点ではあるが，k 距離プロットなどを活用することで実用上の課題は緩和できる．今後も異常検知や空間データ解析といった分野で重要な役割を果たすだろう．

参考文献

[1] DBSCAN Clustering in ML — Density based clustering, `https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/`,
(最終アクセス: 2025-01-29).

[2] Wikipedia, “DBSCAN,”
https://ja.wikipedia.org/wiki/DBSCAN,
(最終アクセス: 2017-04-26).