Technical Sheet
No. 22-13

データ解析入門 9 <HDBSCAN>

キーワード：HDBSCAN、クラスタリング、機械学習

はじめに

テクニカルシート「データ解析入門 8」で紹介した DBSCAN（Density-based spatial clustering of applications with noise）はデータ密度に基づき、データ分割を行うクラスタリング手法です[1]。DBSCAN では特定のデータ分布を仮定することなくクラスタリングを行いますが、各クラスタのデータ密度は同程度であることが前提になります。そのため、密度の異なるクラスタでは期待するようなデータ分割が困難となる場合があります。本稿では、局所的データ密度を考慮し、クラスタリングを実行できる階層的 DBSCAN（HDBSCAN）を紹介します。

クラスタリング手法の比較

HDBSCAN の動作原理を説明する前に、まず各種手法による 2 次元データ[2]のクラスタリング結果を確認します。本データには 6 つのクラスタが存在します（図 1）。カーネル密度推定により得られたデータ密度の等高線から、各クラスタ周辺のデータ密度（等高線の高さ）が異なることがわかります（図 1(b)）。

この 2 次元データに *k*-means 法や階層的クラスタリングを実行しても、期待するようなクラスタは抽出されません（図 2）。

一方、DBSCAN は密度準拠の手法であるため、複雑なクラスタの形状に適用できます。しかしながら、DBSCAN ではすべてのデータに対して一律にクラスタを成長させるため、局所的なデータの疎密には対応できません。たとえば、橙色や赤色のクラスタでは、本来分割されるべき 2 つのクラスタを含みます。これは局所的なデータ密度の違いを考慮せずにクラスタを抽出したことが原因です。

他方、HDBSCAN では直観に適ったデータ分割

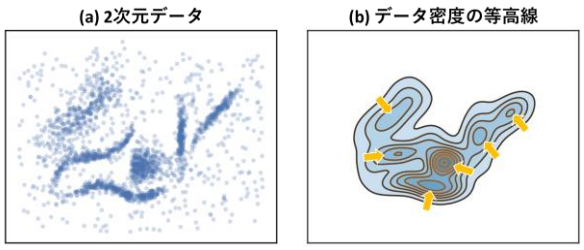

図 1　(a) 2 次元データ、(b) データ密度の等高線

が達成されています。それでは、ここからは HDBSCAN の動作原理について概説します。

HDBSCAN の手順

HDBSCAN の処理の流れは以下の通りです。

1. データ間の相互到達可能距離（Mutual reachability distance）を計算し、最小全域木（MST, Minimum spanning tree）を作成する。
2. MST を基に圧縮したデンドログラムを作成する。
3. 圧縮したデンドログラムからクラスタを抽出する。

簡単のため、少数の 2 次元データに HDBSCAN を適用し、上記手順を説明します（図 3(a)）。

ステップ 1 では、まず一般的な距離の代わりに相互到達可能距離$d_{mreach-k}$を計算します（式(1)）。相互到達可能距離を用いることで、低密度領域におけるデータ間距離が大きくなります。

$$d_{mreach-k} = max\{core_k(\boldsymbol{a}), core_k(\boldsymbol{b}), d(\boldsymbol{a},\boldsymbol{b})\} \quad (1)$$

ここで、$core_k(\boldsymbol{a})$および$core_k(\boldsymbol{b})$はコア点$\boldsymbol{a}$および$\boldsymbol{b}$の k 番目の近傍点までの距離、$d(\boldsymbol{a},\boldsymbol{b})$は$\boldsymbol{a}$-$\boldsymbol{b}$間の距離を表します。次に、MST を作成します。MST とは、閉路を持たず、各データ点（頂点）をつなぐ辺（エッジ）の重みの総和が最小になるグラフ理論に

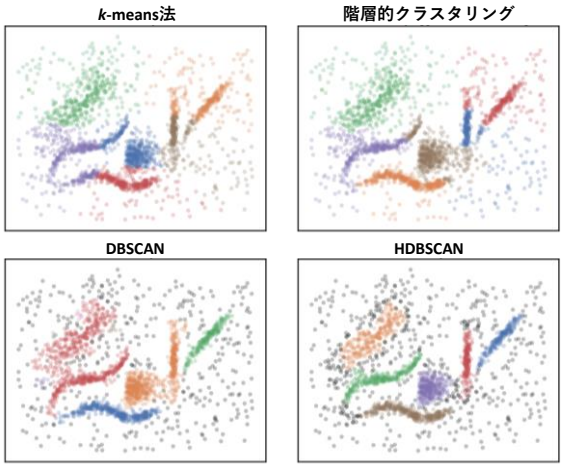

図 2　*k*-means 法、階層的クラスタリング、DBSCAN、および HDBSCAN の実行結果

地方独立行政法人
大阪産業技術研究所　本部・和泉センター
https://orist.jp/
〒594-1157　和泉市あゆみ野 2 丁目 7 番 1 号
Phone: 0725-51-2525（総合受付）

おける重み付きグラフのことです（図 3(b)）。エッジの重みを距離とした MST を用いることで効率的な計算が可能になります。

ステップ 2 では MST を基に、シンプルに圧縮したデンドログラムを得ます。まず、MST における全データに同一のクラスタを割り当てておきます。その後、重みの大きい順に MST のエッジを切断していきます。切断後に生成するデータのつながり（クラスタ）が、事前に設定するクラスタの最小データ数を上回れば、それらデータのクラスタラベルを新たなラベルに更新します。一方、新たなクラスタ内のデータが少なく、データが孤立している状態であれば、外れ値ラベルを割り当てます。以上の操作を、クラスタ数がデータ数に等しくなるまで繰り返します。

ステップ 3 では局所的なデータ密度を考慮し、データ密度の等高線におけるどのピークをクラスタとして抽出し、また、どのピークをひとまとまりのクラスタとみなすかを決定します。そのために、クラスタの安定性（stability）という尺度を算出します（式(2)）。

$$stability = \sum_{p \in Cluster}(\lambda_p - \lambda_{birth}) \qquad (2)$$

この安定性の計算には距離（エッジの重み）の逆数λを用います。MST のエッジ切断時に新たなクラスタが生成しますが、そのときのλをλ_{birth}とします。この新たなクラスタに属するあるデータが別のクラスタラベルに更新される際のλをλ_pとします。このクラスタに割り当てられたすべてのデータについて、$(\lambda_p - \lambda_{birth})$を計算し、それらの総和をクラスタの安定性とします。

次に、この安定性に基づいてクラスタを抽出します。図 3(d)のカラーバーおよびデンドログラムのノードの太さはデータ数を示し、縦軸はλを表します。したがって、クラスタの安定性はノードの面積に対応します。図 4 に示す階層では、丸で囲んだ 2 つのカラー部分の面積の合計が大きいため、それら 2 つのクラスタが抽出されることになります（図 3(e)）。

なお、HDBSCAN にはいくつかパラメータがありますが、種々のパラメータの影響などに関する実践的な内容については The hdbscan Clustering Library のドキュメント[3]が参考になります。

おわりに

本稿では 2 次元データに HDBSCAN を実行しましたが、より一般には高次元のデータが解析対象となります。高次元データのクラスタリング結果を視覚化する方法として、次元削減手法が有用です。次稿では、非線形なデータ分布でもよく機能する次元削減手法を紹介します。

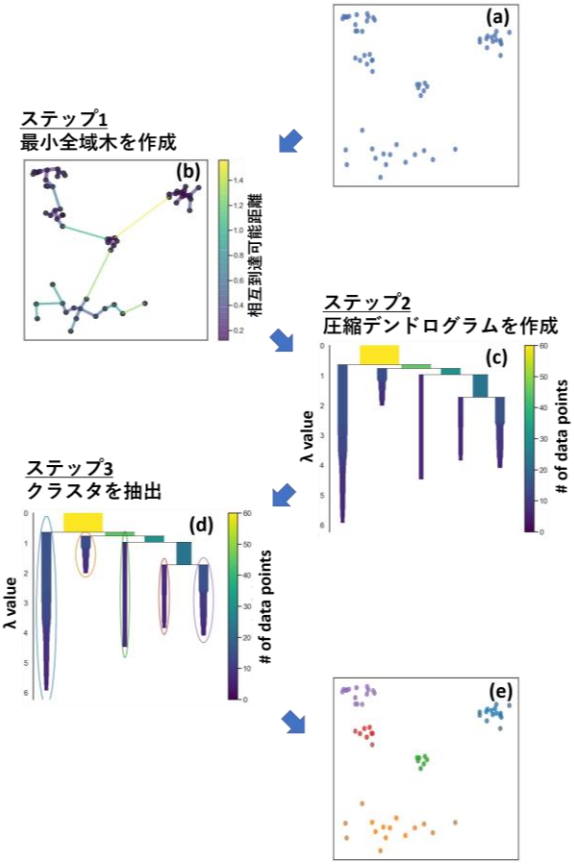

図 3　HDBSCAN の手順について

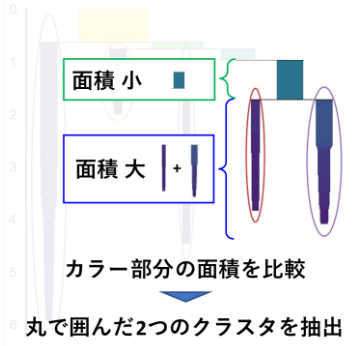

図 4　安定性に基づくクラスタ抽出

参考文献

1) 永廣卓哉：データ解析入門 8 <DBSCAN>、ORIST テクニカルシート、No. 22-12(2022)
2) https://github.com/scikit-learn-contrib/hdbscan/blob/master/notebooks/clusterable_data.npy (accessed on March 11th, 2022)
3) HDBScan - Read the Docs, https://hdbscan.readthedocs.io/en/latest/index.html (accessed on March 11th, 2022)

発行日　2022 年 6 月 1 日
作成者　高分子機能材料研究部　生活環境材料研究室　永廣卓哉
Phone: 0725-51-2611　　E-mail: ehirot@orist.jp