Technical Sheet
No. 22-12

データ解析入門 8 <DBSCAN>

キーワード：DBSCAN、クラスタリング、エルボー法、外れ値検出、機械学習

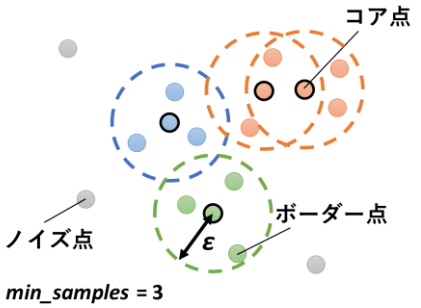
はじめに

代表的なクラスタリング手法である k 平均法（*k*-means 法）[1]は球状のデータ分布を前提にしており、複雑な分布への適用が難しいことがあります。本稿では、データ密度をもとにクラスタリングを実行する DBSCAN（Density-based spatial clustering of applications with noise）を紹介します。*k*-means 法と比べて、DBSCAN は複雑なデータ分布に対応できますが、パラメータ調整が必要になります。以下、DBSCAN の特徴について説明します。

DBSCAN の特徴

DBSCAN は密度に準拠したクラスタリング手法であり、多くのデータが近接している高密度領域内のデータを同一クラスタに割り当てます。具体的には、各データ点を、コア点、ノイズ点、あるいはボーダー点のいずれかに割り当て、クラスタリングを行います（図 1）。このとき、2 つのパラメータ ε および *min_samples* を設定しておきます。

まず、データ点を適当に 1 つ選び、半径 ε 以内に *min_samples* 個以上のデータ点が存在するかどうかを判定します。近傍点数が *min_samples* 個未満の場合、そのデータ点をノイズ点とし、クラスタを割り当てません。一方、*min_samples* 個以上の近傍点が存在する場合、そのデータ点をコア点とし、新たなクラスタラベルを割り当てます。このとき、コア点から半径 ε 以内の近傍点にも同じクラスタラベルを割り当てます。次に、それらの近傍点についても上記と同様の判定を行い、コア点と判定された場合は同じ処理を実行します。コア点の判定が続くと、クラスタは成長していきますが、半径 ε 以内の近傍点にコア点が存在しなくなればいったん成長は終了します。その後、まだ判定されていないデータ点を適当に選び、同様の処理を継続し、クラスタリングを実行します。



	半径ε以内の近傍点の数
コア点	*min_samples*個以上
ノイズ点	*min_samples*個未満 （近傍点にコア点を含まない）
ボーダー点	1個以上*min_samples*個未満 （近傍点にコア点を含む）

図 1　コア点、ボーダー点、およびノイズ点

なお、コア点から半径 ε 以内に存在し、コア点でもノイズ点でもない点をボーダー点といいます。DBSCAN では、コア点とノイズ点のクラスタは一意に決まりますが、ボーダー点のクラスタは判定するデータ点の順序によって変化することがあります。

DBSCAN の主な特徴を以下に示します。*k*-means 法と異なり、クラスタ数を事前に指定する必要はありません。

【長所】
- クラスタ数を事前に指定する必要がない
- データ分布の形状を仮定しない
- 外れ値に頑健である

【短所】
- 各クラスタの密度は同程度であることが前提
- パラメータ調整が必要

k-means 法 vs. DBSCAN

k-means 法では、各クラスタの重心からの距離に基づきクラスタリングを実行します。そのため、分布が複雑であるデータでは、クラスタリングがうまくいかないことがあります。たとえば、図 2 に示す月形の

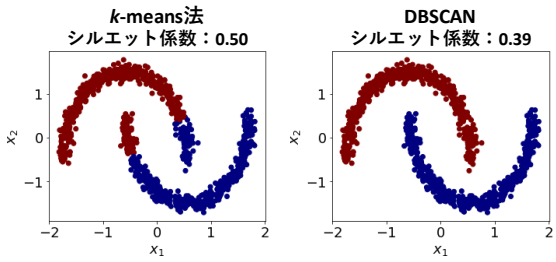


図 2　*k*-means 法および DBSCAN

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）

人工データではクラスタリングがうまくいっていないことがわかります。一方で、密度ベースの手法である DBSCAN では、クラスタ数を指定することなく月形の 2 つの高密度領域を分割できています。

ここで、両手法のシルエット係数を確認してみます。*k*-means 法のシルエット係数は 0.50 であり、DBSCAN の 0.39 よりも大きくなっています。シルエット係数は各クラスタの凝集度と乖離度に基づいて算出され、1 に近くなるほどよいクラスタとみなします。直観的には DBSCAN によるデータ分割の方が良好ですので、今回の解析例ではシルエット係数はうまく機能していないといえます。このように、単純な距離ベースの評価指標であるシルエット係数は、複雑な分布に不向きであるため注意が必要です。

DBSCAN のパラメータについて

繰り返しになりますが、DBSCAN にはパラメータ ε と *min_samples* が存在します。パラメータ ε が小さいと、ノイズ点と判定されるデータ点が増加し、クラスタの成長が抑制されます。一方、*min_samples* はクラスタに含まれる最小データ数を表します。*min_samples* = 5 とし、ε を変更した際の結果を図 3 に示します。プロットの色の違いは異なるクラスタを表しており、上述の ε の影響を確認できます。

ここまで 2 変数の人工データを用いて DBSCAN の特徴を確認してきましたが、変数がより多くなると図示しながらパラメータ調整を行うことは難しくなります。そこで、あるデータ点の近傍点を用いて、適切な ε の値にあたりをつけることを考えます。たとえば、図 4 左の概念図に示した *k* 番目の近傍点までの距離が参考になります。DBSCAN ではコア点かどうかの判定に近傍点を用いていました（図 4 右）。そのため、すべてのデータ点について、*k* 番目の近傍点までの距離の推移を確認することで、適切な ε の範囲を推察することができます。

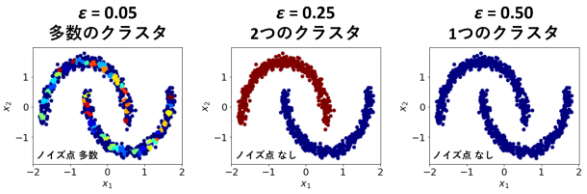


図 3　DBSCAN におけるパラメータ ε の影響

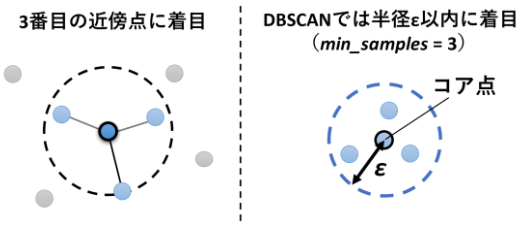


図 4　近傍点について

ここでは、5 番目の近傍点までの距離を降順に並べ替えました（図 5）。距離が 0.25 あたりになると、近傍距離の減少が緩やかになっていることがわかります。つまり、このあたりでデータ密度の差が小さくなっていることを示します。したがって、ε = 0.25 の周辺を検討すると良いように思われます。実際、ε = 0.25 とすると期待するようなクラスタリング結果が得られました（図 3）。

本手法はエルボー法と呼ばれることがありますが、前稿 [1]で紹介した解析例と同様に、判断が曖昧になることがありますので、最適化ではなく、あくまで ε の探索範囲を限定するための手法であると理解しておくと無難です。

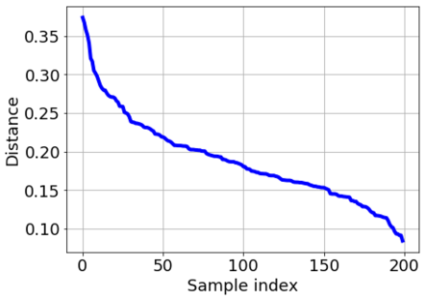


図 5　5 番目の近傍点までの距離

DBSCAN による外れ値検出

上述のように、DBSCAN のアルゴリズムでは、各データ点をコア点、ノイズ点、あるいはボーダー点に割り当てます。そのため、外れ値に頑健な手法であるといわれますが、これは外れ値（ノイズ点）を検出可能であるということでもあります。なお、解析ソフトウェアによっては実行後に得られるクラスタラベルに外れ値ラベルも含まれることがあるため、結果の取り扱いに注意すべき場合があります。

おわりに

DBSCAN は密度ベースのクラスタリング手法であり、*k*-means 法にはない長所を有することを紹介しました。ただし、DBSCAN では、各クラスタの密度が同程度であることを前提とします。次稿では、密度の異なるクラスタに適用できる階層的 DBSCAN（HDBSCAN）を紹介します。

参考文献

1) 永廣卓哉：データ解析入門 7 <*k*-means 法>、ORIST テクニカルシート、No. 22-11 (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