Technical Sheet
No. 22-11

データ解析入門 7 <*k*-means 法>

キーワード：*k*-means 法、クラスタリング、エルボー法、シルエット係数、機械学習

はじめに

本稿では、非階層的クラスタリング手法である *k* 平均法（*k*-means 法）について紹介します。階層的クラスタリングと異なり、*k*-means 法では事前にクラスタ数を指定してから解析する必要があります。そのため、適切なクラスタ数を推察するための手法として、エルボー法とシルエット係数を用いる手法についても併せて紹介します。

k-means 法の手順

k-means 法は代表的な非階層的クラスタリング手法であり、事前にクラスタ数を指定してからデータ分割を行います。以下に *k*-means 法の流れを示します（図 1）。

1. クラスタ数 *k* を指定する。
2. 乱数をもとに、各データにいずれかのクラスタを割り当てる。
3. 各クラスタ重心を計算する。
4. クラスタ重心と各データとの距離を計算する。
5. 各データに重心が最も近いクラスタを割り当て直す。
6. クラスタの割り当てが収束するまで上記ステップ 3〜5 を繰り返す。

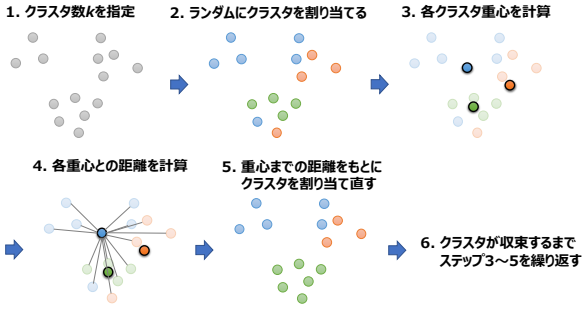


図 1　*k*-means 法の流れ

k-means 法の特徴

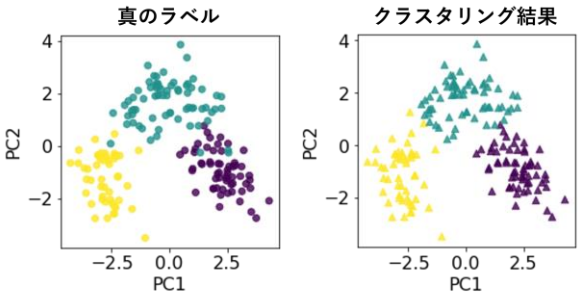
k-means 法は動作原理がシンプルであり、解釈しやすいクラスタリング手法ですが、事前にクラスタ数を設定する必要があります。また、得られる結果はステップ 1 における初期値に依存します。そのため、実用的には *k*-means 法によるクラスタリングを 100 回以上実行し、各クラスタにおける重心と各データ間距離のばらつきが小さくなる実行結果を採用などの対応が実施されます。

クラスタ数について

以下、ワインのデータセット [1]を *k*-means 法によりクラスタリングします。本データセットには、アルコール濃度や色強度などの 13 項目の分析結果が含まれています。

k-means 法の実行にあたり、はじめにクラスタ数を指定します。今回のデータには、ワインの種類に応じて 3 つのクラスラベルが割り当てられています。そこで、まずはクラスタ数を 3 として *k*-means クラスタリングを実行します。続いて、データ可視化のため、主成分分析（PCA）を実行します。実際のワインのクラスラベルおよび *k*-means 法で得られたクラスタリング結果をもとに、PCA のスコアプロットを色分けすると図 2 が得られます。クラスの境界付近の色の分布に違いが見られますが、おおむね良好にクラスタリングされているといえます。

今回の解析例では、3 つのクラスタ数が適切であるとわかっていましたが、一般的には適切なクラスタ数は不明である状況が想定されます。そのため、これまでに適切なクラスタ数を推察するための数多くの手法が提案されてきました。次頁では、エルボー法およびシルエット係数を用いた適切なクラスタ数の推定法について紹介します。



PC1: 第1主成分、PC2: 第2主成分

図 2　PCA によるスコアプロット

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）

エルボー法

ここで、改めてクラスタリングの良し悪しについて考えます。同じクラスタ内のデータは近くに位置し、異なるクラスタのデータは遠くに位置しているという状態であれば、クラスタリングは良好であるといえます。つまり、同一クラスタにおけるデータの凝集度と各クラスタの乖離度が高くなるようなクラスタリングが望まれます。

エルボー法では、クラスタ内におけるデータの凝集性に焦点を当て、クラスタ内誤差平方和(SSE)を指標としてクラスタリングの良し悪しを評価します。図 3 に示すように、クラスタ数を増やしていくと、SSE は減少していきます。ただし、たとえ SSE が 0 であっても、データの数だけクラスタを生成してしまっては意味がありません。そこで、クラスタ数に対して SSE をプロットし、SSE とクラスタ数の両者が小さくなるクラスタ数を探ります。

今回のデータでは図 3 に示すプロットが得られました。クラスタ数が 2 から 3 に変化すると SSE は大きく減少しますが、クラスタ数が 3 から 4 に増加してもあまり SSE は減少しません。エルボー法では、このようなひじ(エルボー)のように折れ曲がる点を探します。SSE のプロットから、クラスタ数は 3 つが適切であるように思われます。

なお、エルボー法は代表的なクラスタ数推定手法として知られていますが、明瞭な折れ曲がりが見られず、適切なクラスタ数の推定が難しい場合もあります。

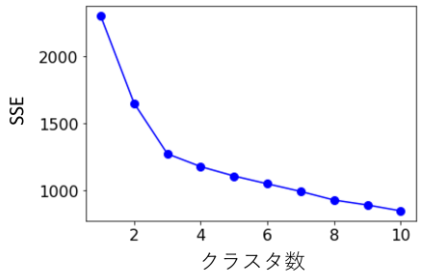


図 3　エルボー法で用いる SSE のプロット

シルエット係数

次に、シルエット係数について紹介します。シルエット係数は、データの凝集度に加え、クラスタの乖離度も考慮した指標です。以下にシルエット係数の算出式を示します。

$$s_i = \frac{b_i - a_i}{max(a_i,\ b_i)} \qquad (1)$$

ここで、s_i、a_i、b_i はそれぞれ i 番目のサンプルのシルエット係数、同一クラスタ内のその他すべてのサンプルとの平均距離、最近傍のクラスタのすべてのサンプルとの平均距離を表します。また、$max(a_i, b_i)$は括弧内の変数の最大値を表します。シルエット係数は、$-1 \le s_i \le 1$ の値をとり、s_i が 1 に近くなると良いクラスタリングであるとみなします。

すべてのサンプルのシルエット係数 s_i を計算し、横向きの棒グラフとして並べると図 4 が得られます。なお、クラスタごとに棒グラフを色分けしています。クラスタ数を 3 に設定すると、全サンプルのシルエット係数の平均値は 0.285 となりました。クラスタ数を 2～10 とし、k = 2～10 におけるシルエット係数の平均値をプロットすると図 5 が得られます。シルエット係数のグラフからもクラスタ数を 3 と設定するのがよさそうです。

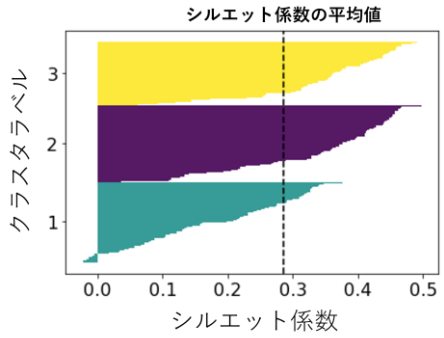


図 4　k = 3 におけるシルエット係数

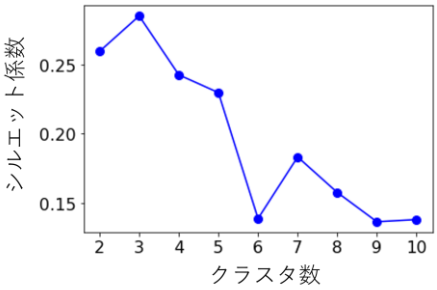


図 5　各クラスタ数におけるシルエット係数

おわりに

k-means 法自体はシンプルなクラスタリング手法ですが、より発展的な関連手法も開発されています。例えば、k-means 法よりも外れ値に頑健な k-medoids 法などがあります。なお、k-means 法は球状のデータ分布を前提としており、複雑なデータ分布に対応できないことがあるため注意が必要です。次稿では、密度ベースのクラスタリング手法について紹介します。

参考文献

1) UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data (accessed on April 28th, 2021)

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