Technical Sheet
No. 22-04

データ解析入門 6 <階層的クラスタリングの実践>

キーワード：階層的クラスタリング、多変量解析、ヒートマップ

はじめに

階層的クラスタリングはデータ前処理やデータマイニングなどに用いられる有用な解析手法です。本稿では階層的クラスタリングの計算方法により、異なる結果が得られることを示します。また、ヒートマップを援用したデータマイニングについて説明します。

階層的クラスタリングの実行

以下、178 件のワインの分析結果を解析します [1]。本データセットは、アルコール濃度や色強度などの全 13 項目に関する分析結果です。まずはデータを標準化します。次に、距離尺度としてユークリッド距離を用いることにし、重心法、群平均法、およびウォード法により階層的クラスタリングを実行すると図 1 が得られます。各デンドログラムの最大距離の 70 % をしきい値として分割すると、ウォード法では 3 つの明瞭なクラスタに分割されました。一方、重心法と群平均法では、より多くのクラスタに分割され、クラスタ内のデータ数に偏りが見られました。このように計算方法により得られる結果は変化します。また、用いる距離尺度によっても結果は変わります。

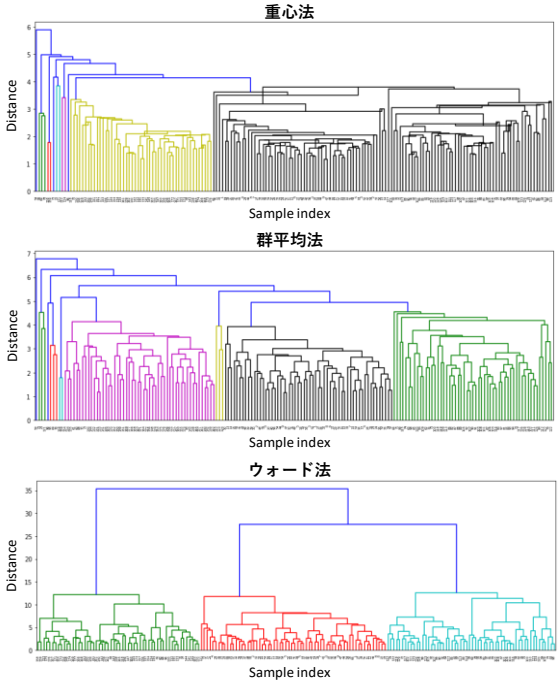

図 1　群平均法、重心法、およびウォード法の結果

変数のクラスタリング

図 1 ではサンプルをクラスタリングした結果を示しましたが、13 個の変数をクラスタリングすることも可能です。図 2 に変数をクラスタリングした結果を示します。これは 178 次元のデータをクラスタリングした結果であり、大きく 3 つのクラスタが存在するように見受けられます。

なお、データを前処理しなかった場合、図 2 とは全く異なる結果となります（図 3）。これは変数のスケールの違いが原因です。階層的クラスタリングに限りませんが、スケーリングの有無により解析結果が変わりうることに注意が必要です。

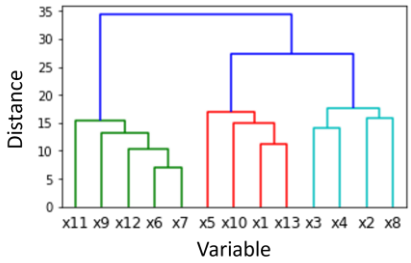

図 2　13 変数の階層的クラスタリング結果
（前処理あり）

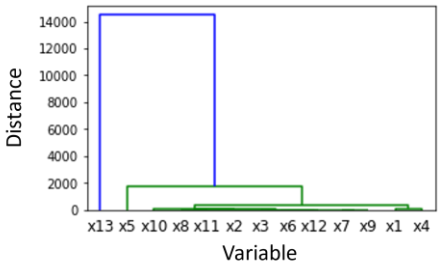

図 3　13 変数の階層的クラスタリング結果
（前処理なし）

階層的クラスタリングとヒートマップ

次に、ヒートマップを紹介します。ヒートマップは行列型のデータの個々の数値を色や濃淡として表現した可視化グラフの一種です。ヒートマップを用いることで、3 次元曲面を 2 次元平面で表現できます（図 4）。ここでは z 軸の数値の大小を色の違いで

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）

表現しています。

今回のデータセットをヒートマップで表現すると図 5 のようになります。ここでは、178 件の分析結果（サンプル）の並びをランダムにシャッフルしており、ヒートマップは無秩序になっています。そのため、図 5 から知見を抽出することは困難です。

しかしながら、階層的クラスタリングを実行することでデータマイニングが容易になることがあります。階層的クラスタリングの結果に基づき、サンプルおよび変数を並び替えると、図 6 のようなヒートマップが得られます。所々、ヒートマップにパターンが浮き上がって見える箇所があります。例えば、ヒートマップ左上には赤いブロックがあるように見えます。そのため、一番上のクラスタに属するサンプルでは、x_{11} や x_9 などの変数の値が小さくなる傾向にあることがわかります。同様に、中央下にも赤いブロックがあるように見えます。一番下のクラスタに属するサンプルでは x_1 や x_{10} が小さくなる傾向にあると推察されます。

最後に主成分分析[2, 3]によりスコアプロットを図示します（図 7）。今回用いたワインのデータセットでは、各データにラベルが割り振られており、左図はその真のラベルごとに色分けしています。一方、右図ではクラスタリング結果をもとにデータを色分けしています。両者の分布は類似しており、階層的クラスタリングにより精度良く層別できていることが確認できます。

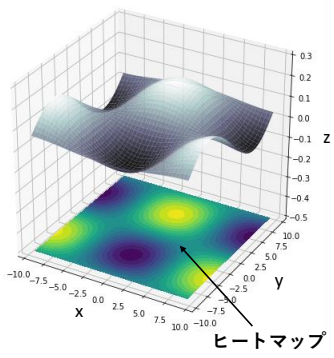

図 4　ヒートマップとは

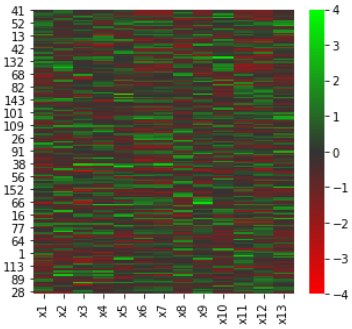

図 5　ランダムなサンプル順序のヒートマップ

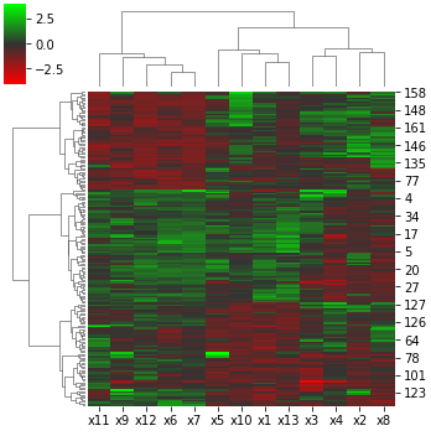

図 6　階層的クラスタリングおよびヒートマップ

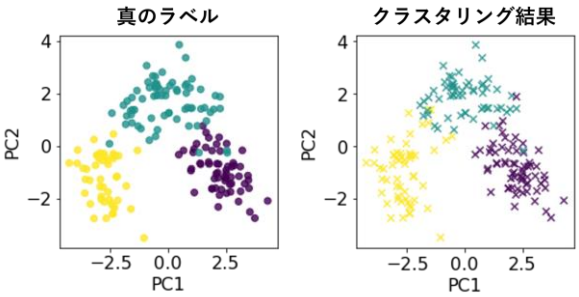

図 7　スコアプロットおよびクラスタリング結果

おわりに

階層的クラスタリングの結果は距離尺度や計算方法に依存します。そのため、データの特徴を考慮した上で、計算条件を選定することが重要になります。ただし、どのような計算方法が適しているか明らかではない場合も多く、本稿で例示したような計算方法の試行錯誤も重要です。今回の解析では、Ward 法において 3 つのクラスタの存在が示唆されましたが、実際にはクラスタ数の推定が難しい場合も少なくありません。次稿では、非階層的手法である k 平均法および適切なクラスタ数を推定するための代表的な手法について説明します。

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)

2) 永廣卓哉：データ解析入門 2 <主成分分析によるデータの可視化>、ORIST テクニカルシート、No. 21-24 (2021)

3) 永廣卓哉：データ解析入門 3 <主成分分析によるデータマイニング>、ORIST テクニカルシート、No. 21-25 (2021)

発行日　2022 年 4 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