Technical Sheet
No. 22-03

データ解析入門 5 <階層的クラスタリングの基礎>

キーワード：階層的クラスタリング、距離、多変量解析、次元の呪い

はじめに

複数のデータを分割するための解析手法のことをクラスタリングやクラスター分析といいます。クラスタリングは類似したデータからなる複数のグループを得ることを目的としており、データマイニングやデータ前処理などに用いられます。本稿では、階層的クラスタリングに関する基礎について紹介します。

データ間の距離（非類似度）について

クラスタリングは階層的手法と非階層的手法に大別できます。階層的手法では近接するデータ点同士を逐次的につなぎ、デンドログラム（樹形図）を作成します（図 1）。一方、非階層的手法では階層構造を形成せずデータを分割します。

これから紹介する階層的クラスタリングでは、データ間の距離を基準にグループ分けを行います。すなわち、2 点のデータ間距離を非類似度として用い、距離が近い 2 つのデータは類似しているとみなします。なお、距離には様々な定義が存在しますが、代表的な距離の定義を図 2 に示します。

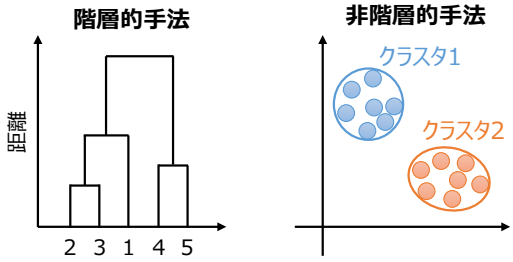

図 1　クラスタリングの種類

ユークリッド距離はなじみのある直線的な距離のことであり、データ解析ではよく用いられます。

マンハッタン距離は、各変数の差の絶対値から計算されます。一般に、ユークリッド距離よりも外れ値に対して頑健になります。

マハラノビス距離は異常検知などでよく用いられます。多変量データでは変数が相関することがありますが、このときユークリッド距離を用いると変数間の相関により異常検出力が低下することがあります。そこで、等方的な距離尺度ではなく、異方的な距離尺度でデータ分布の異方性を補正します。これは変数間の相関を考慮するということであり、このような観点のもと、変数間の相関をキャンセルした距離がマハラノビス距離です。マハラノビス距離の計算過程では、主成分分析を実行しているとみなすことができます。全ての主成分から算出した T^2 統計量 [1]の正の平方根がマハラノビス距離に対応します。

そのほか、スペクトルデータなどの非類似度として用いられるコサイン距離や化学構造のクラスタリングなどに用いられる Tanimoto 距離などがあります。

階層的クラスタリングの手順

以下に階層的クラスタリングの手順を示します。

1) クラスタリングで用いる距離の定義を指定
2) クラスタ間距離の計算方法を指定
3) 距離が最も近い 2 つのデータ点を結合し、クラスタを作成
4) 距離が近いデータ点（クラスタ）を順番に結合していき、クラスタが 1 つになったら終了

ユークリッド距離

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

最もなじみのある直線的な距離

マンハッタン距離

$$d_{ij} = \sum_{k=1}^{m} |x_{ik} - x_{jk}|$$

カクカクと座標点を動いたときの距離

マハラノビス距離

$$d_{ij} = \sqrt{(\boldsymbol{x_i} - \boldsymbol{x_j})\Sigma^{-1}(\boldsymbol{x_i} - \boldsymbol{x_j})}$$

データが異方的に分布している

変数間の相関をキャンセルした距離

d_{ij}：データ点$\boldsymbol{x_i}$, $\boldsymbol{x_j}$との距離
x_{ik}：i番目のデータ点のk番目の変数の数値
m：変数の数
$\boldsymbol{x_i}, \boldsymbol{x_j}$: i, j番目のデータ点
$\boldsymbol{\mu}$：変数の平均値を格納したベクトル
Σ：分散共分散行列

図 2　代表的な距離の定義

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）

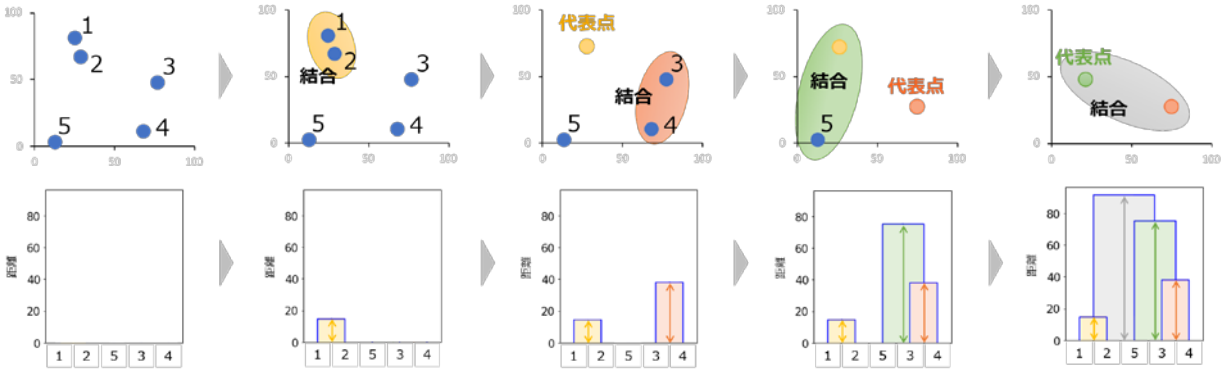

図 3　階層的クラスタリングの流れ

図 3 に階層的クラスタリングの流れを示します。まず、計算に用いる距離の定義を決めておきます。次に、距離計算のためにクラスタ間距離の計算方法を指定します。ここでは、各クラスタの重心を代表点として計算する重心法を用いることにします。その後、距離が最も近いデータ点 1、2 を結合し、クラスタを作成します。同様に、クラスタが 1 つになるまでクラスタを結合していきます。以上の処理により、デンドログラムが得られます（図 3 右下）。

図 3 ではクラスタ作成のために重心法を用いましたが、そのほかの計算手法として群平均法やウォード法（Ward 法）などがあります（図 4）。ウォード法ではクラスタ内のデータのばらつきが小さくなるようにクラスタが逐次的に形成されます。

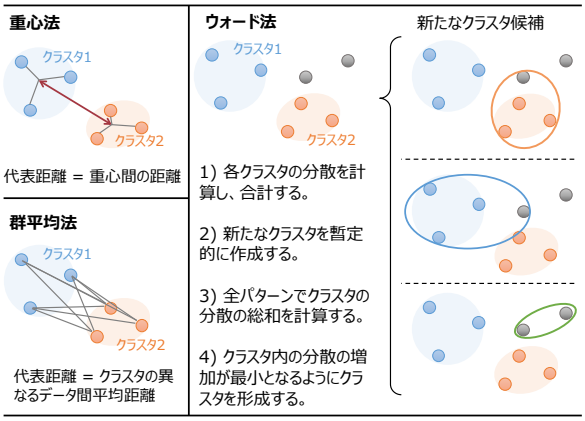

図 4　重心法、群平均法およびウォード法

次元の呪い

上述のとおり、階層的クラスタリングは距離を用いる手法ですが、変数が多くなりすぎると次元の呪いという厄介な問題に直面することがあります。

一般に、データの次元が大きくなるにつれ、必要なデータ量や計算量は増加します。また、相関の強い変数の組やノイズが増加する可能性もあります。そのため、高次元であることが原因でデータ解析が難しくなることがあります。高次元データが抱える問題は、総称して「次元の呪い」と呼ばれます。

以下にデータが高次元になるにつれて、2 点のデータ間距離の分布がどのように変化するかを示します。ここでは、各要素が 0 以上 1 未満の一様乱数である d 次元ベクトル（d = 2, 8, 32, 128, 512, 1024）を 1000 個生成し、2 点間距離の分布を得ました。図 5 に示すように、データが高次元になるに従い、データ間距離が増加していることがわかります。一方、ヒストグラムは次第にシャープになっており、データ間距離のばらつきは縮小しています。つまり、高次元空間では 2 点間距離の差が小さくなるということです。このような高次元データの特性には注意が必要です。そのため、次元削減はデータ解析における重要な課題のひとつとなっています。

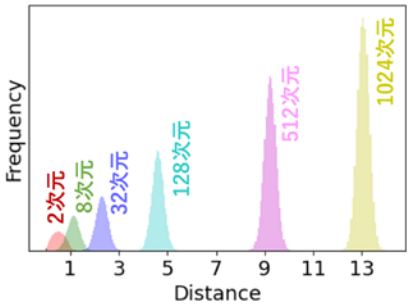

図 5　データ間距離のヒストグラム

おわりに

本稿では階層的クラスタリングの動作原理などについて紹介しました。次稿では、具体的な多変量データを用いて、階層的クラスタリングを実行します。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。

参考文献

1) 永廣卓哉：データ解析入門 4　<T^2 統計量と Q 統計量を用いた異常検知>、ORIST テクニカルシート、No. 21-26 (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