D-019

デンドログラムに基づくグラフィカルなクラスタ分析支援ツールの開発
Development of Graphical Support Tool for Cluster Analysis Based on Dendrogram

筧　和政† 青砥　直弘† 蓬莱　尚幸†
Kazumasa Kakehi Naohiro Aoto Hisayuki Horai

1. はじめに

クラスタ分析[1]とは、分析対象データの集合をクラスタと呼ばれる部分集合に分けて分析する手法であり、心理学[2]や生物学[3]など様々な分野で用いられている分析手法である。クラスタ分析には階層的クラスタリングという手法がある。その結果はデンドログラムで表され、分析者はデンドログラムを解釈することで知見を得る。しかし、分析対象データの集合が膨大になるとデンドログラムが膨大となるので、人手によって知見を得ることは非効率的であり困難となる。したがって、デンドログラムの解釈を支援するツールを開発し、クラスタ分析を支援することが本研究の目的である。

2. 階層的クラスタリング

階層的クラスタリング(以降、クラスタリングと呼ぶ)とは、分析対象データの集合を階層的な包含関係をもつ複数のクラスタに分けて分析する手法である。クラスタリング対象データは共通の属性をもち、各データがそれぞれの属性に対して一つの属性値をもつ。したがって、データ集合 Data と属性集合 Attribute と属性値 Value の間には以下の(1)式の関係があり、各データは属性値によって特徴づけられる。

$$\mathrm{Data} \times \mathrm{Attribute} \rightarrow \mathrm{Value} \quad \cdots (1)$$

相関の度合いとして、各データがもつ属性値からデータ間の距離を定義する。クラスタリングの手順は、(1)最も距離の小さい 2 個のデータ(またはクラスタ)を結合してクラスタを生成する。(2)生成したクラスタと、結合されずに残っているデータ(またはクラスタ)の間の距離を求める。(3)1 個のクラスタに全データが含まれるまで(1),(2)を繰り返す。

以下の(2)式はクラスタが包含関係をもつことを表した式である。C、C_1、C_2はクラスタである。

$$C \supset C_1, C_2 \quad \cdots (2)$$

(2)式において Cを親、C_1、C_2を子と呼ぶ。各データはそのデータ自身を含むサイズ 1 のクラスタとみなす。親クラスタのサイズ$|C|$には式(3)で示す性質がある。$|C_1|$、$|C_2|$はそれぞれ子クラスタ C_1、C_2のサイズである。

$$|C| = |C_1| + |C_2| \quad \cdots (3)$$

デンドログラムはクラスタの包含関係を 2 分木として表す図である。葉は分析対象の各データであり、根は分析対象データの全体集合である。枝はクラスタを表す。節はクラスタの結合を表し、葉から節までの長さが結合した子クラスタ間の距離を表す。図 1 で示すデンドログラムのクラスタ a, b は、それぞれ、データ 1 のみを含むクラスタとデータ 1,2 を含むクラスタである。

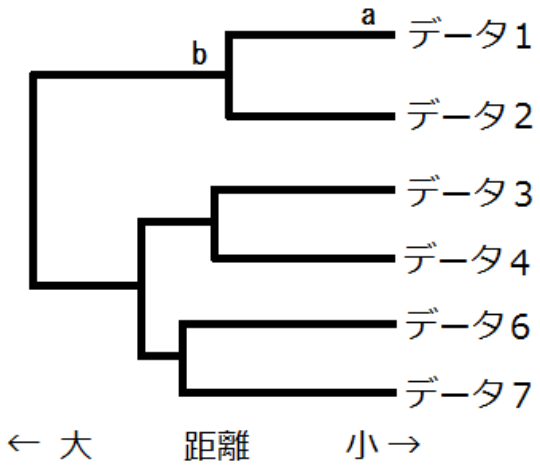

図 1 デンドログラムの例

3. クラスタ分析支援ツール

3.1 本ツール概要

本ツールは、クラスタリング対象データベースとデンドログラムを用いてクラスタ分析支援を行う。クラスタリング対象データベースには、2 節で述べたクラスタリング対象データと属性と属性値の関係が格納されている。

クラスタ分析では、特徴的なクラスタから知見を得られることが多い。そのため、特徴的なクラスタを発見することは非常に重要である。したがって、本ツールは、特徴的なクラスタを発見する機能をユーザに提供する。本研究では、特徴的なクラスタを「ある特徴的なデータを多く含むクラスタ」あるいは「ある特徴的なデータを偏って含むクラスタ」とする。前者のクラスタを発見する機能を「割合に基づくクラスタ発見機能」と呼ぶ。後者のクラスタを発見する機能を「偏りに基づくクラスタ発見機能」と呼ぶ。ここで特徴的なデータとは、特定の属性値の組をもつデータである。また、クラスタを発見する機能を利用するために、ユーザは属性値の組を指定する必要がある。

本ツールは、クラスタを発見する機能のほかに、発見したクラスタを図示する機能として、「印を付加する機能(マーキング機能)」と「発見したクラスタを単純な長方形に置き換えて表現する機能(クラスタ - 長方形置換機能)」をもつ。

3.2 割合に基づくクラスタ発見機能

割合に基づくクラスタ発見機能では、特徴的なデータを含む割合 P を指定する。特徴的なデータの総数を N、

† （独）国立高等専門学校機構　茨城工業高等専門学校, NIT Ibaraki College

あるクラスタに含まれる特徴的なデータの個数を n とすると、以下の式(4)を満たすクラスタを候補とする。

$$\frac{n}{N} > P \quad \cdots (4)$$

その中で、極小のクラスタを採用する。また、ツールでは割合をグラフィカルに指定できる。

3.3 偏りに基づくクラスタ発見機能

偏りに基づくクラスタ発見機能では、各クラスタ内での特徴的なデータが占める割合によって、特徴的なデータが偏ったクラスタを発見する。本ツールでは、クラスタ内の偏りの度合い D_C をサイズ $|C|$ に対する特徴的なデータの個数 n の割合と定義する。しかし、サイズの小さいクラスタが選ばれ易いという問題が生じるので、クラスタのサイズによって補正するために重み付けを行う。今回は、重み付けに $|C|$ の平方根を乗ずることにした。したがって D_C は式(5)のように表す。

$$D_C = \sqrt{|C|}\frac{n}{|C|} = \frac{n}{\sqrt{|C|}} \quad \cdots (5)$$

式(5)から、すべてのクラスタの D_C を求め、最大の D_C を与えるクラスタを発見する。

また、ユーザが閾値として距離を指定し、その閾値までクラスタリングして生成されるクラスタ集合について、属性値ごとに最大の D_C を与えるクラスタを発見する。ここで利用する属性値としては、あらかじめユーザが閾値とともに指定した属性がとりうるすべての属性値が利用される。また、ツールでは閾値をグラフィカルに指定できる。

3.4 マーキング機能

マーキング機能は、発見したクラスタと特徴的なデータが一目でわかるように、それらをデンドログラム上で明示する機能である。発見したクラスタはそのクラスタが含むデータに直線を付けて明示する。特徴的なデータはそのデータに色を付けて明示する。図 2 はデータ 1, 2, 3 からなるクラスタとデータ 5, 6 からなるクラスタが発見された場合の表示例である。データ 1,3,5,6 は特徴的なデータである。

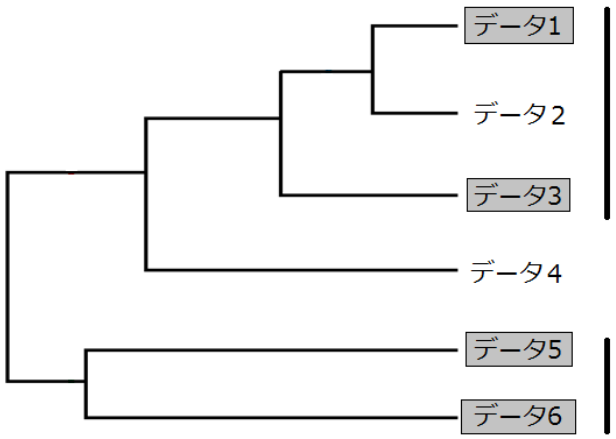

図 2 印をつける機能の表示例

3.5 クラスタ - 長方形置換機能

クラスタ - 長方形置換機能は、発見したクラスタを単純な長方形に置き換えることで、デンドログラム全体を小さくまとめ、その俯瞰性を向上させる機能である。長方形の横の長さはクラスタの距離を表し、縦の長さはクラスタのサイズを表す。さらに、縦の長さはクラスタ間のサイズの比を保ったまま縮小できる。図 3 は図 2 で示したクラスタにこの機能を使用した表示例である。

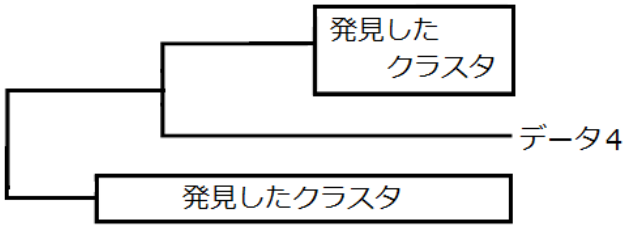

図 3 長方形に置き換える機能の表示例

4. 外部知識データによる拡張

外部知識データとは、クラスタリング対象データに関する既知の分類と性質を扱うために属性を拡張したものである。したがって、分類と性質にはそれぞれデータ集合との間に式(1)と同様な関係がある。データ集合 Data と分類 Classification とクラス Class の間の関係を式(6)に、データ集合 Data と性質 Property と性質値 PValue の間の関係を式(7)に示す。

$$\text{Data} \times \text{Classification} \rightarrow \text{Class} \quad \cdots (6)$$

$$\text{Data} \times \text{Property} \rightarrow \text{PValue} \quad \cdots (7)$$

分類と性質は属性に対応し、クラスと性質値は属性値に対応している。したがって、クラスや性質値の組を指定してクラスタ発見機能を使用することができる。この拡張によって、クラスタリング対象データがもつ属性値以外の値を用いた分析の支援ができるようになるのでデンドログラムの解釈の多様性が高まる。

5. おわりに

本研究では、デンドログラムを用いたクラスタ分析支援ツールを考案した。今後、実装を行い KNApSAcK データベース[4]の漢方薬データなどの分析に試用して評価する。

参考文献

[1]B.S.Everitt *et al.* ñCluster Analysisò,Wiley,(2011).
[2]東　正訓, ñ社会的ルールの認知構造ò, 追手門学院大学人間学部紀要, 4 号, pp.49-56 (1997).
[3]岡田　吉史, 三林　光, 長島　知正, ñ生物学的知識を導入した遺伝子発現データの自動分類ò, SVBL 年報, Vol.6, pp.33-36 (2004).
[4]F.M.Afendi *et al.* "KNApSAcK Family Databases: Integrated Metabolite-Plant Species Databases for Multifaceted Plant Research", Plant Cell Physil 53(2) (2012).