区間値関数データの階層的クラスタリング

清水　信夫　　データ科学研究系　助教

区間値関数データとは何か?

- 解析対象とする個体のデータがそれぞれ系統的に観測された場合において、各々の系統的データが関数化されたものを関数データという。
- 関数データのうち、定義域内の個々の値に対応する値域が区間データとして表されるものを**区間値関数データ**という。
- 区間データはシンボリックデータの一種であることから、区間値関数データもシンボリックデータの一種と考えることができる。

区間値関数データの何をどのようにクラスタリングする?

- n個の区間値関数データがあるとき、各個体における区間値関数データは定義域によって分割せず、定義域全体で1つのデータとして考える。
- シンボリックデータ解析においては、区間データ間の距離規準がいくつか定義されているが、これを関数データ解析における距離規準の定義に適用することで、区間値関数データ間の距離規準を定義する。
- クラスタリング手法は階層的クラスタリングと非階層的クラスタリングに大別されるが、前者においては区間値関数データに関する距離規準をデータ集合に適用し、クラスターを形成するための手法（アルゴリズム）を選んでデンドログラムを作成する。

区間値関数データ間の距離規準の定義

異なるn個の区間値関数データ $f_1^{INT}(t), f_2^{INT}(t), \cdots, f_n^{INT}(t)$ が、実数区間 $T=[t^L, t^U]$ において

$$f_i^{INT}(t)=[f_i^L(t), f_i^U(t)], \quad t\in T, \quad i=1,\cdots,n$$

と表されるものとする。ここで $f_i^L(t)\in L^2(T)$ および $f_i^U(t)\in L^2(T)$ はそれぞれ $f_i^{INT}(t)$ の下界関数および上界関数であり、$t\in T$ において $f_i^L(t)\le f_i^U(t)$ である。以下では $f_k^{INT}(t)$ と $f_l^{INT}(t)$ の間の距離規準について示す。ただし $1\le k\le n, 1\le l\le n, k\ne l$ とする。

- 関数Hausdorff 距離

$$D_H^{INT}(f_k,f_l)=\int_{t\in T}\max\{|f_k^L(t)-f_l^L(t)|,|f_k^U(t)-f_l^U(t)|\}dt$$

- 関数ユークリッドHausdorff 距離

$$D_{EH}^{INT}(f_k,f_l)=\sqrt{\int_{t\in T}\max\{|f_k^L(t)-f_l^L(t)|^2,|f_k^U(t)-f_l^U(t)|^2\}dt}$$

- 関数Gowda-Diday 非類似度

$$D_{GD}^{INT}(f_k,f_l)=\int_{t\in T}\left[|f_k^L(t)-f_l^L(t)|/|y(t)|\right]dt$$
$$+2\int_{t\in T}\left[\max\{|f_k^U(t)-f_k^L(t)|,|f_l^U(t)-f_l^L(t)|\}-I_{kl}(t)\right]/U_{kl}(t)\,dt$$

ただし

$$|y(t)|=\max_{1\le i\le n} f_i^U(t)-\min_{1\le j\le n} f_j^L(t)$$
$$U_{kl}(t)=\left|\max\{f_k^U(t),f_l^U(t)\}-\min\{f_k^L(t),f_l^L(t)\}\right|$$
$$I_{kl}(t)=\max[\min\{f_k^U(t),f_l^U(t)\}-\max\{f_k^L(t),f_l^L(t)\},0]$$

- 関数Ichino-Yaguchi 非類似度

$$D_{IY}^{INT}(f_k,f_l)=\int_{t\in T}\{U_{kl}(t)-I_{kl}(t)\}dt \qquad (0\le\gamma\le 0.5)$$
$$-\gamma\int_{t\in T}\{|f_k^L(t)-f_k^L(t)|+|f_l^U(t)-f_l^U(t)|-2I_{kl}(t)\}dt$$

区間値関数データの階層的クラスタリングアルゴリズム

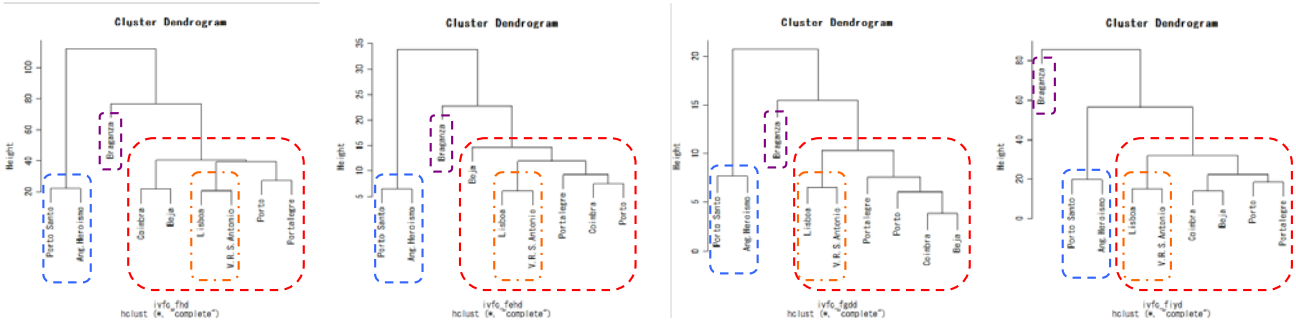
区間値関数データの階層的クラスタリングは、従来型のデータの階層的クラスタリングを行う場合の拡張として、以下に示すアルゴリズムにより行われる。

1. 初期状態としてn個の区間値関数データがそれぞれ1つのクラスターを形成しているものと考える。すなわちこの時点でのクラスターの個数KはK=nとする。
2. K個のクラスターの中で最も非類似度(距離)の小さいクラスターの対 $C_i, C_j (i\ne j)$ を求め、それを1つのクラスターに融合する。ここでK→K-1としてK>1ならば3.に進み、そうでなければ4.に進む。
3. 新しく作られたクラスター $C_i\cup C_j$ と他のクラスターとの非類似度(距離)を計算し、その情報を得て2.に戻る。
4. 必要な情報を出力して終了する。

階層的クラスタリングにおいては、融合すべきクラスターの選択規準の違いにより様々な方法が提案されているが、いずれの方法も上記アルゴリズムにおける2.および3.におけるクラスター間の非類似度(距離)の計算において区間値関数データ間の距離規準を適用することで拡張可能である。

ポルトガル9都市の月別平均気温データ(1971～2000年)の解析例

最長距離法(Complete Linkage)を用いた場合においてそれぞれの距離規準を適用した場合のデンドログラムを下記に示す。



関数Hausdorff距離　　関数ユークリッドHausdorff距離　　関数Gowda-Diday非類似度　　関数Ichino-Yaguchi非類似度 ($\gamma=0.5$)