mwp-110

クラスタ分析概要　【 評価版 】

Stata におけるクラスタ分析に関する機能は `cluster` 系の各種サブコマンドによって提供されます。個々のサブコマンドの機能、用法については別の whitepaper に譲るとして、本 whitepaper ではクラスタ分析全般に関する概念や用語について説明を行います。

1. クラスタ分析

クラスタ分析とはデータの自然なグルーピング（クラスタリング）を決定する操作のことを言います。そのための手法は分割型と階層型とに区分されます。

(1) 分割型手法

分割型手法 (partition clustering methods) の場合、観測データは指定された数のグループに分割されます。Stata では kmeans 法と kmedians 法という 2 種類の手法が実装されています。

kmeans **法**の場合、それぞれの観測データはその平均値が最も近いグループに一旦アサインされます。その後で平均値の値が再度計算され、それに基づきグルーピングが再評価されます。これらのステップは観測データの移動がなくなるまで繰り返されることになります。このアルゴリズムは k 個のシード値から開始されることになりますが、その指定方法には様々なものがあります。kmedians **法**でも同様の反復計算が行われますが、kmeans 法との違いはグループ中心が平均値ではなく中央値によって規定されるという点にあります。これらの機能はそれぞれ `cluster kmeans`, `cluster kmedians` コマンドによって提供されます。

(2) 階層型手法

階層型手法 (hierarchical clustering methods) の場合には階層的に関係したクラスタ群が生成されます。この階層型手法はさらに凝集型 (agglomerative) と分枝型 (divisive) に二分することができます。

凝集型の階層クラスタリングは個々の観測データ (observation) を別個のグループとみなす形で処理を開始します。すなわちこの段階ではサイズが 1 の N 個のグループが構成されたことになります。次に最も近接した 2 つのグループが結合され、グループの数は $N-1$ 個に減じられます。このプロセスは繰返し実行され、最終的にはすべての観測データが 1 つのグループに帰属する形となります。結果として階層構造のクラスタ群が形成されます。

2 つの観測データを比較する再には**類似度**、あるいは**非類似度**をどのような尺度で測るか (similarity/dissimilarity measures) が問題となるわけですが、複数の観測データを含むグループ同士をどのように比較するかについても基準を設けておく必要があります。このグループ間の比較に用いられる手法は**連結法** (linkage method) と呼ばれ、種々のものが選択できるようになっています。

一方、**分枝型**の階層クラスタリングの場合には、すべての観測データが単一のグループに集約された状態を出発点とし、それらを次々に分割して行くというアプローチを取ります。ただしこの分枝型は余り用いられることがないため、Stata では実装されていません。

2. 連結法

Stata の階層型クラスタリングにおいては次の連結法がサポートされています。

- 単連結法
- 完全連結法
- 平均連結法
- Ward 連結法
- 重心連結法
- メディアン連結法
- 加重平均連結法

(1) 単連結法

単連結法 (single linkage method) の場合、2 グループ間の類似性/非類似性はグループ間で最も近傍に位置する 2 つの観測データ間の類似性/非類似性として計算されます。この手法の場合、連鎖 (chaining) と呼ばれる問題を生じることがあります。2 グループ間の最も近傍に位置する 2 点が次の併合を決するため、細長いクラスタが形成されてしまうわけです。これを避けるには完全連結法や平均連結法等の他の結法等を使用する必要があります。Stata のコマンドとしては `cluster singlelinkage`（または `clustermat singlelinkage`）が該当します。

(2) **完全連結法**

評価版では割愛しています。

(3) **平均連結法**

評価版では割愛しています。

(4) Ward **連結法**

評価版では割愛しています。

(5) **重心連結法**

評価版では割愛しています。

(6) **加重平均連結法、メディアン連結法**

評価版では割愛しています。

3. 類似度/非類似度

3.1 連続変数用の尺度

評価版では割愛しています。

3.2 二値変数用の尺度

評価版では割愛しています。

4. Postclustering コマンド

評価版では割愛しています。

5. クラスタ管理コマンド

評価版では割愛しています。

■