

多変量解析 <初歩の初歩>に入門

クラスター分析

なかもとけい
kenakamoto@nifty.com

クラスター分析とは?

- ケース(事例, 被験者)あるいは変数をボトムアップ的に分類していく方法.
 - 非類似性(距離)を定義し,
 - 似ているモノ同士 距離が近いもの同士をどんどんくっつけていって,
 - まとまりを作る.
 - 階層的クラスター分析と非階層的クラスター分析に大別できる.
- 「距離」は(多次元空間上での)抽象的な距離を指す.
- 距離の定義方法には色々ある.
- くっつける方法(結合方法)にも色々ある.

階層的 クラスター分析

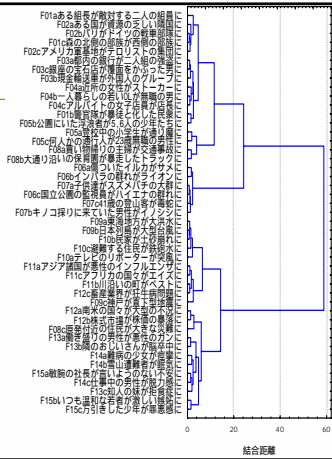
結果の例

おなじみの図

樹形図, あるいはデンドログラムと呼ぶ.

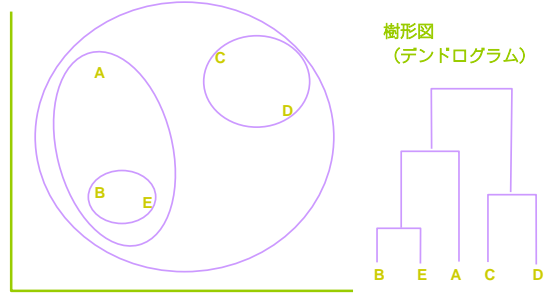
右にいくほどまとまり化(クラスター化が進む)

二分木(Binary Blanching)



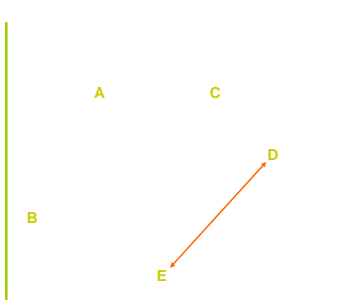
結合過程の模式図

樹形図
(デンドログラム)



(多次元空間上の)距離(1)

変数(素性)
 X_2



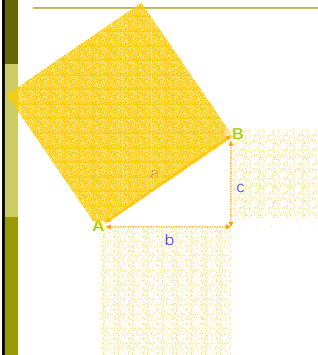
近い位置にある
= 当該の変数群(素性群)に関して, 似たような値を取る
= 類似している.

(2点間の)距離の決め方は色々ある.

ユークリッド距離

変数(素性) X_1

ユークリッド距離とか



$$a^2 = b^2 + c^2$$

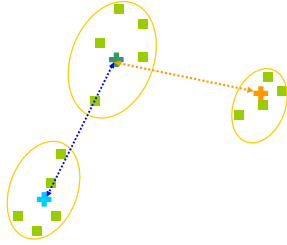
平方ユークリッド距離

$$a = \sqrt{b^2 + c^2}$$

ユークリッド距離

重心法

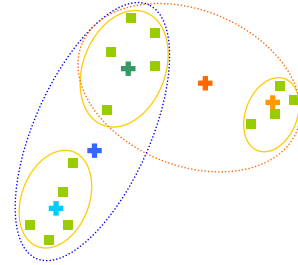
- 各クラスターの重心を求めて、重心間の距離を計算する。距離の近いクラスターを結合する。



計算手続きにクラスターの中心の計算が含まれているので、プロトタイプの議論なんかとなじみがいい。

ワード法

- クラスター内のばらつきがなるべく増えないように、クラスターを結合していく方法。



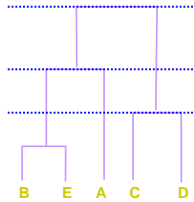
いい感じの結果を出しやすい。

どの水準の解を採用するか

- 数値的な基準は決まっていない。
 - 良い解釈が得られそうな水準でクラスター数を決める。

- その他の基準として、

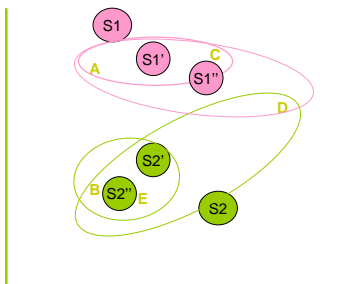
- 結合距離が長くなるところで切る。
- 各クラスターで各変数の平均値をとり、クラスター間で有意な差があるかを(分散分析などで)検討する。
- 判別分析で上手く予測できるクラスター数を採用する。
- 各種統計量基準 (R^2 近似値, 部分的 R^2 , 疑似F値, 疑似 t^2 値, CCC基準) を提案する。
 - <http://www.csc.fi/cschejp/sovellukset/stat/sas/sasdoc/sashtml/stat/index.htm>



非階層的クラスター分析: K-means

- クラスターを階層化せず、単一の水準でクラスターを作る手法。
 - あらかじめクラスター数を決めてから分析する。
 - 分類対象数が多い場合、(1) 階層的クラスター分析よりも結果の解釈が楽かもしれない、(2) (最近のPCではほぼ問題にならないけど) 階層的クラスター分析より計算負荷が低い。

非階層的クラスター分析: 模式図



- クラスター数を決定する。
- シード(クラスターのモト)をランダムに巻く。
- 各観測値を一番近いシードに割り当てる。
- シードを再計算して更新。
- 新しいシードに観測値を割当て。
- [4]->[5]を繰り返す。
- 結果が変わらなくなった (= 収束したら) 終了。

気をつけた方がいいこと(1)

- クラスター分析は探索的な手法であり、「正解」を与えるわけではない。
 - 非常に自由度の高い(ある意味ではいい加減な?)手法
 - 距離の定義や結合方法によって結果が異なる。
 - あまりに大きく異なるようなら、考察を慎重に行う必要がある。
 - 他の統計的手法や(実験やコーパス分析, 作例などの)様々な手法での研究との併用を視野に入れた方がいいでしょう。

気をつけたほうがいいこと(2)

- 現実的に困難なことも多いけど...
- 変数(素性)群の性質に気を配りつつ、距離を考える必要がある。
- 多くの場合(平方)ユークリッド距離モデルを採用すると思われる
 - が、これは(ホントは)変数が直交しているときに使う方法
 - しかし、実際には測定した変数/コードした素性に相関が生じているのが普通
- 解決法
 - 相関係数(など)を算出し、相関の高すぎるペアがあったら片方を削る。
 - 主成分分析や因子分析によって直交する合成変数/(潜在)因子得点を求め、それを用いてクラスター分析を行う。
 - ケースバイケースで判断してください。

クラスター分析のまとめ

- 今回は下記のことを学びました。
 - クラスター分析は階層的クラスター分析と非階層的クラスター分析に大別できる。
 - 距離・非類似性の定義の仕方は色々ある。データ構造などを考慮して良い定義を選ぶ必要がある。
 - クラスターの結合方法には色々ある。
 - 最適なクラスター数を決定する外的な基準はない。幾つか指標はあるけれども、一意には決まらない。
 - ということで、総じて「自由度の高い」というか「いい加減な」というかそういう手法である。
 - 他の多変量解析と同じかそれ以上に「実質科学的知見」による解釈の比重が高い手法といってよいでしょう。
 - できれば、クラスター分析の結果を他の研究結果(他のデータセットから得た結果や、全く異なるデータ収集手続きに基づく結果)と比較・対照することが望ましいです。

おまけ

- 決定木 decision tree
 - クラスター分析とは逆に top-downに領域を分割し、クラスターを作っていく方法
 - JMP, Statisticalには入っている(と思う)。SPSSでは別途オプション購入の必要有り(のような雰囲気)。