“分ける”ことは“分かる”こと？
―クラスター分析―

データ解析演習20100630
M2　中山真孝

もくじ

分けるとは？

われわれ人間は、同時的・経時的に環境から絶え間なく知覚入力を受け取っている

その中から意味ある情報を分節化することで環境を“認知”している

＝その場その場で、場環境情報を圧縮して、いくつかのまとまりとして環境を認知し行動を決定する必要

分けるとは？

長期的に見ても、過去の出来事を記憶するには、個々のエピソードを分類・抽象化・情報圧縮が必要であり、それをよびだして現在の行動に役立てるにも、過去と現在の出来事の類似性を見出さなければ役に立たない

⇒分節化・分類・情報圧縮・抽象化つまり“分ける”ことは人間の認知の基本である

分けるとは？

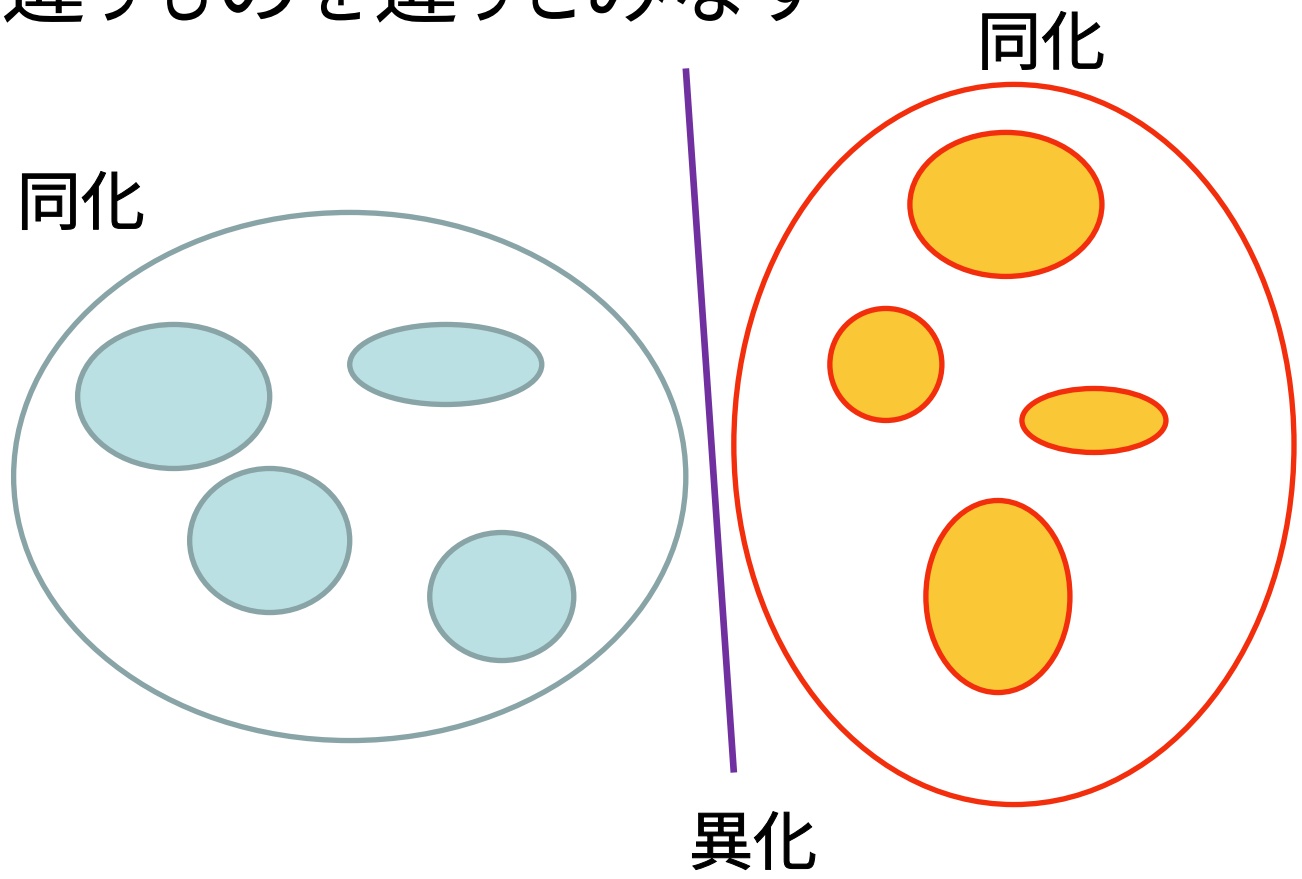
心理学においても、あるいは心理データ解析においても分けることは重要

- なんらかの（複数の）性格特性によっていくつかの類型に分ける
- なんらかの（複数の）心理的特性（属性）によって、刺激を分ける（＝それら刺激がいかに表象されているかの構造を知る）

分けるとは？

同じものを同じとみなす

違うものを違うとみなす



分けるとは？

全体を1つの集合とみなしてそれを分割する

個々の要素からはじめて同じものを併合する

⇒分割も併合も“分ける”ことの一側面

クラスター分析で行うのも大まかにはこの分割と併合（後で紹介する階層的クラスター分析最大に分割された状況から順に併合していく方法）

分けるとは？

よい分類・よく“分かる”分け方とは？

より少ない情報次元で

より多い情報量を得る

＝効率よくかつ正確に

併合：情報次元が減る＝効率よく

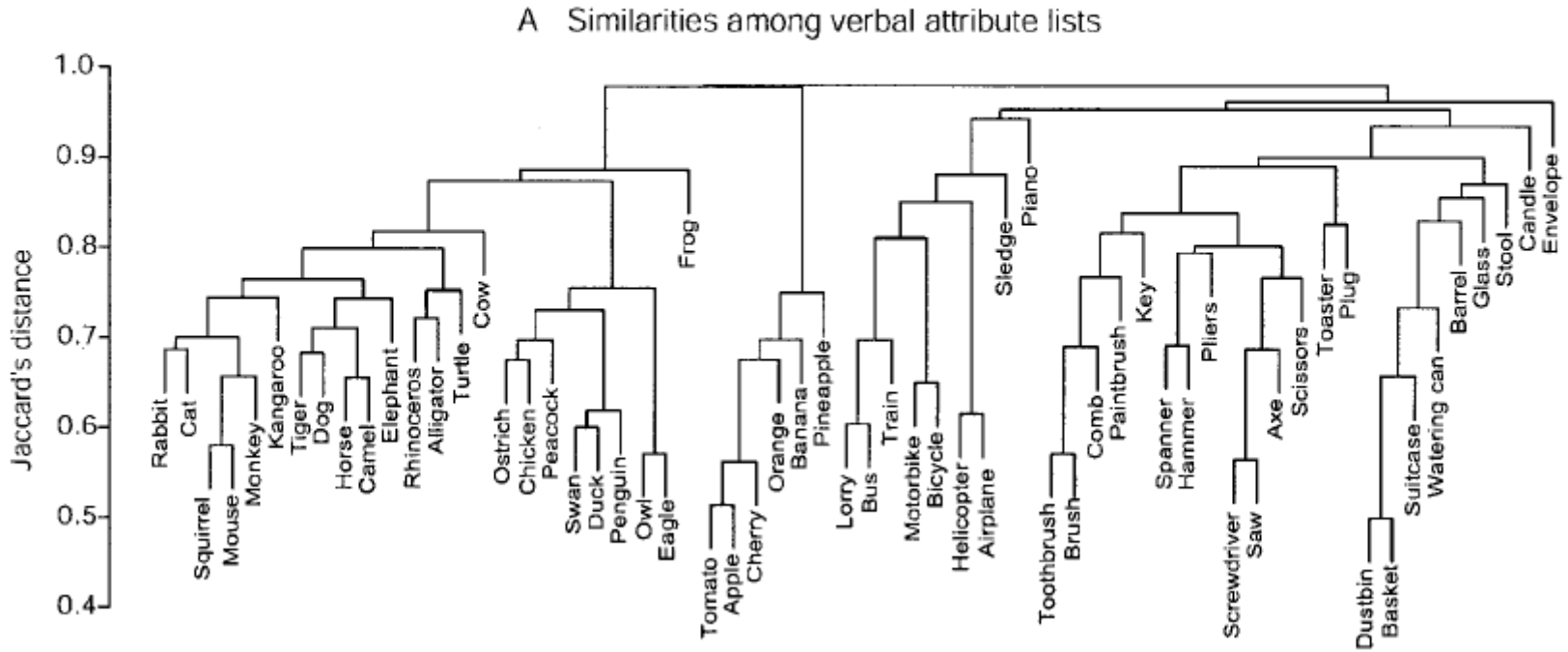
分割：情報量が多くなる＝正確に

クラスター分析では、基本的には情報量（正確さ）の保存され方を定量的に評価し、併合

(階層的) クラスター分析

最終的にこんなデンドログラムを作る

ROGERS ET AL. (2004)



どうやって？

(階層的) クラスター分析の数学的説明

要素Aと要素B (例えば、りんごとバナナ) はどれくらい同じとみなしてよいのか？

・(非) 類似度を直接得る

ーりんごとバナナはどれくらい似ていますか？

・いくつかの観点 (変数) から評価し、そこから (非) 類似度を得る

りんご・バナナはそれぞれ

ーどれくらい甘いですか？

ーどれくらい丸いですか？

(階層的) クラスター分析の数学的説明

これらをもとに、クラスターを併合していく

- まず、すべての要素を異なるクラスターとする
- もっとも似たクラスター2つを併合して新たなクラスターを作る (総クラスター数は1つ減る)
- 新たなクラスターの中でもっとも似たクラスター2つを併合して新たなクラスターを作る (総クラスター数は1つ減る)
- これをクラスター数が1になる (＝すべてが同じクラスターになる) まで繰り返す

(階層的) クラスター分析の数学的説明

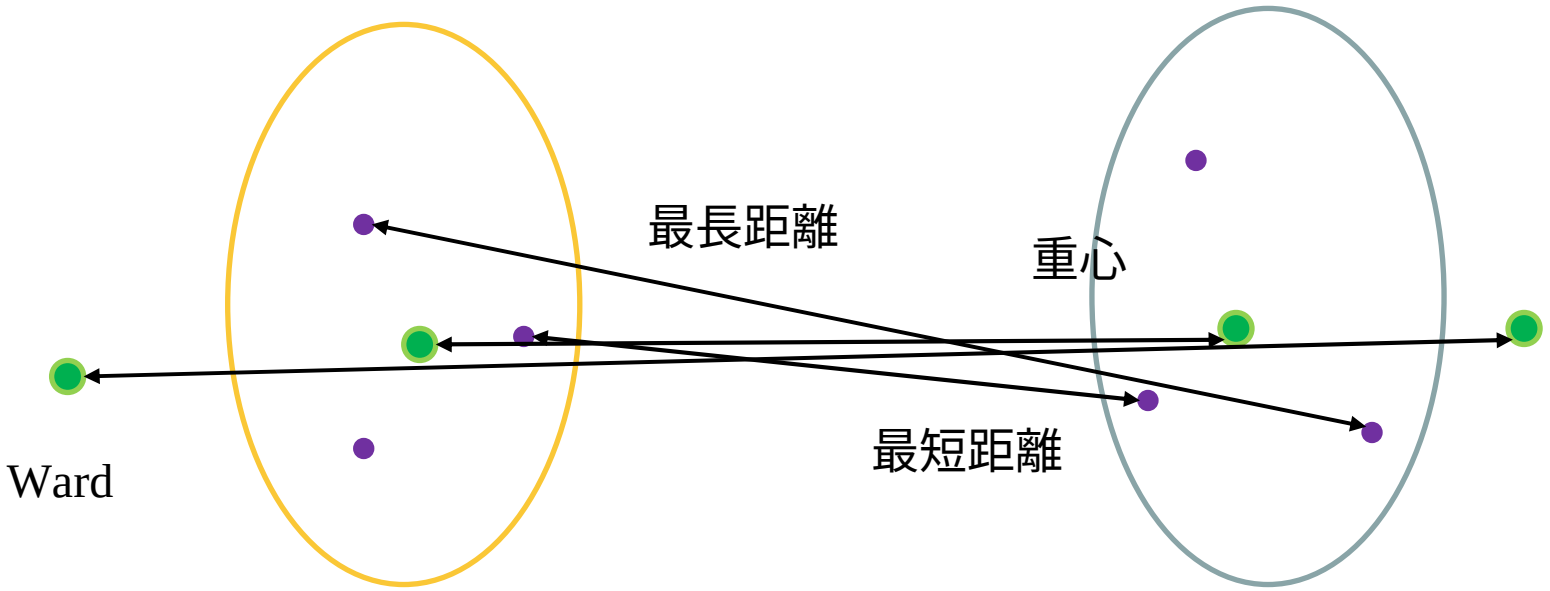
クラスター同士の類似度 (　併合した時の情報量の低下の程度) はいかに評価するのか

→いろいろ方法がある

- 最短距離法 (もっとも近い要素同士の距離)
- 最長距離法 (もっとも遠い要素同士の距離)
- 重心法 (クラスターの中心間距離)
- 群平均法 (クラスター間の要素の距離の総当たりの平均)
- ウォード法 (クラスターを作った時の2乗誤差を最小に)
- 可変法 (上記方法の一般化)

などなど

(階層的) クラスター分析の数学的説明



齋藤・宿久 (2006) を参考に作成

⑷階層的⑷ クラスター分析

＊注意点

クラスター数や分類の妥当性に明確な基準はない

方法の選択によっても結果は異なる

どのような基準で分類するとよいかは理論的に決めるべき！？

クラスター分析は探索的に行うものであり、それによって明確な結論を得るものではない

⑷ただし、次に確証的クラスター分析というべきものの事例を紹介⑷

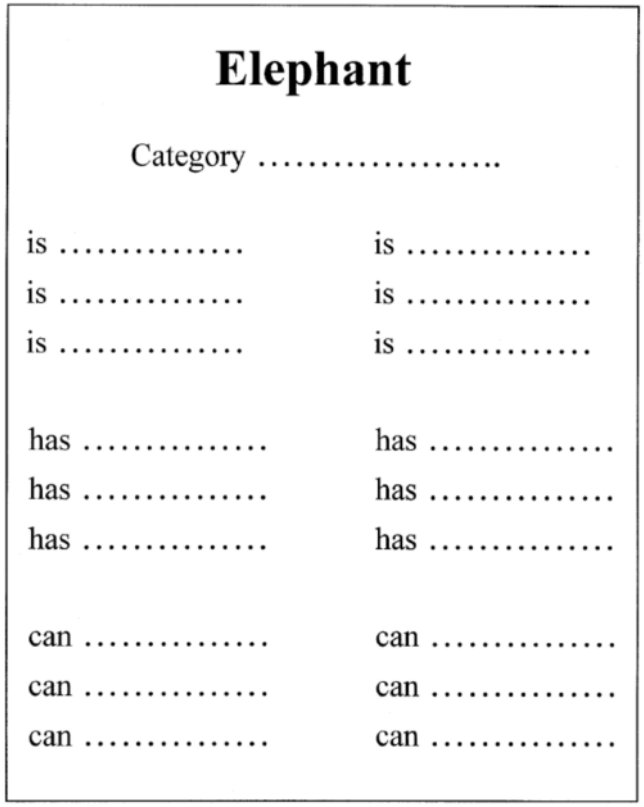
クラスター分析実例

Garrard et al (2001)を紹介 (一部)

意味記憶が 特徴ベースで説明できるか？

64の概念について、どのような属性をもつかを
　被験者に調査

２人以上から挙げられた
属性をその概念がもつ
属性 (特徴) とする



Elephant

Category ……..

is ……	is ……
is ……	is ……
is ……	is ……
has ……	has ……
has ……	has ……
has ……	has ……
can ……	can ……
can ……	can ……
can ……	can ……

Garrard et al (2001)

表現上の違いで同じ属性を表していると思われるものはまとめておく

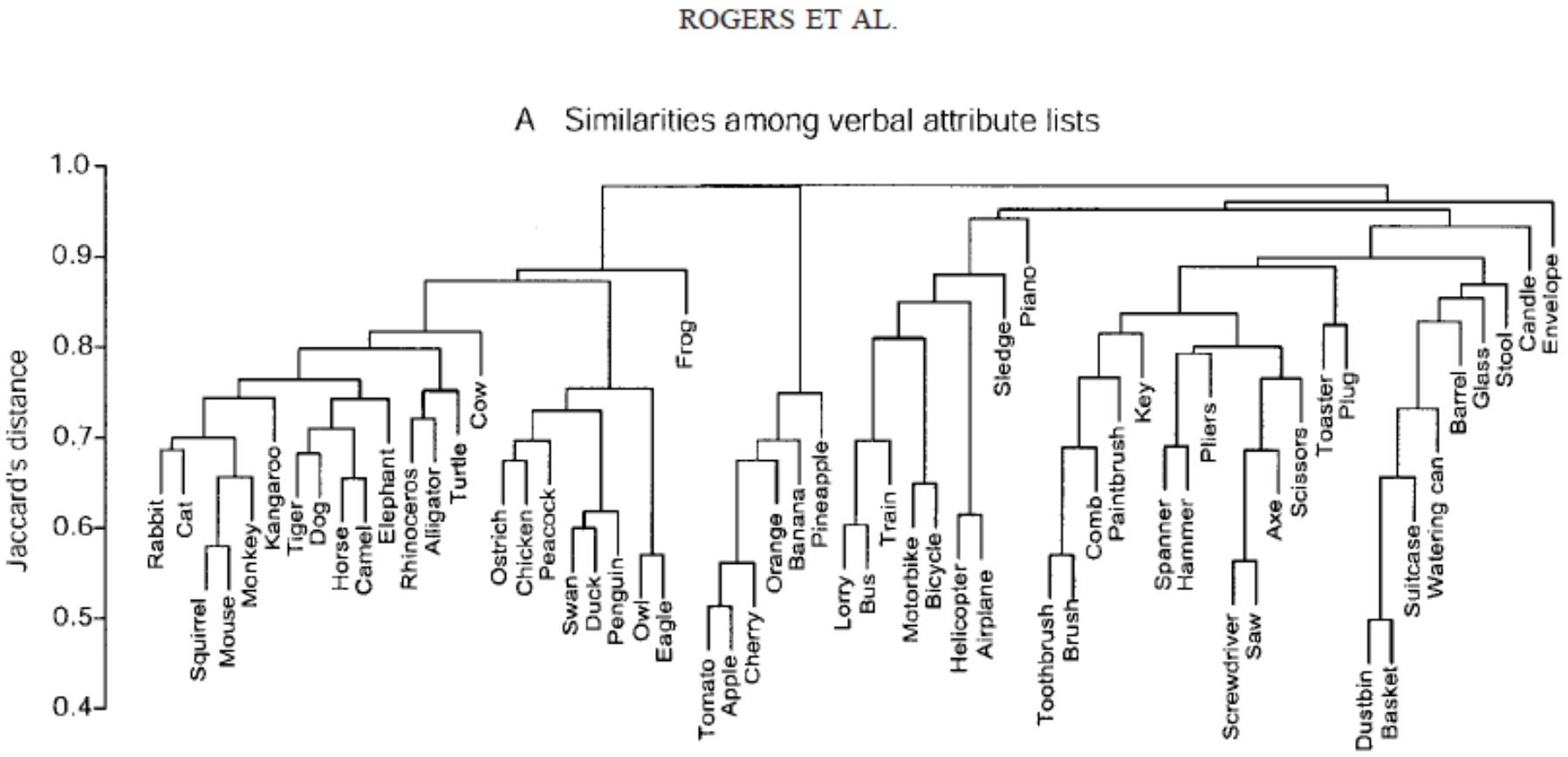
全体で挙げられた属性のありなしを変数として各概念を2値ベクトルで表現

←リンゴ（1 0 0 0 1 1 1 0 .....）といった感じで

これをもとにクラスター分析（先ほど用いたユークリッド距離ではなく2値ベクトル用のJaccardの距離を用いる。たぶん。）

Garrard et al (2001)

こんなデンドログラムに



図自体はRogers et al(2004)より引用

Garrard et al (2001)

同時に（1年後に）同じ被験者から、各概念についてカテゴリ内での典型性と親近性を評定してもらう

この典型性（と親近性）の評定値は、クラスター分析のカテゴリ（クラスター）における中心（重心）からの距離（ベクトル間のコサイン）と相関

→先行研究の行動実験（Rosch & Mervis, 1975）と一致

引用・参考文献（URL）

Garrard, P., Lambon Ralph, M. A., Hodges, J. R. & Patterson, K. Prototypicality, distinctiveness and intercorrelation: analyses of the semantic attributes of living and nonliving items. *Cogn. Neuropsychol.* ***18,*** 125–174 (2001).

Rogers, T. T. *et al.* Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychol. Rev.* ***111,*** 205–235 (2004).

齋藤堯幸・宿久洋 (2006) 関連性データの解析法―多次元尺度構成法とクラスター分析法―　共立出版

http://aoki2.si.gunma-u.ac.jp/lecture/misc/clustan.html（青木先生のページ；WEB上でのクラスター分析も）

http://www.kamishima.net/jp/clustering/（神嶌先生のページ）

http://cse.niaes.affrc.go.jp/minaka/R/R-cluster.html（三中先生のページ；クラスター分析の歴史など概念的な話も）

http://psy.isc.chubu.ac.jp/~oshiolab/teaching_folder/datakaiseki_folder/11_folder/da11_01.html（小塩先生のページ；SPSSの使い方なども）