第13章
テキストのクラスター分析

茨城大学工学部
高木真

概要

- 複数のテキストを分析する際に、テキストの何らかの特徴にもとづいて似ているものごとにグループ分けする必要がある場合がある。

- 本章ではテキスト間の類似度(または距離)にもとづいてテキストをグルーピングする方法やその応用例を説明する。

テキストのクラスター分析

- テキストのクラスター分析
  - テキストの分散、相関、類似度や距離の情報を用いてグループ分けすること。
- クラスター分析のアプローチ

i. 個体の特徴の情報にもとづいて、平面や立体空間上で散布図を作成し、分布状況からクラスターの形成状況を分析する。（主成分分析、対応分析、自己組織化マップ法）

ii. データの類似度あるいは距離を用いて、最も似てる個体を近い位置に配置する方法。(多次元尺度法)また、似ている個体順にグルーピングするクラスタリング法(階層的クラスタリング法)

iii. 全体が何グループに分けられるかを指定し、それぞれのグループの中心を機械的に求め、その中心までの距離が近いグループに属すると判断する方法(k-means法)

本章では、多次元尺度法、クラスタリング法、k-means法などについて説明する。

類似度と距離

- 類似度
  - 似ているほど値が大い測度
  - ピアソンの相関係数(最もよく知られている)

$$cor(x,y)=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2(y_i-\bar{y})^2}}$$

x, y : 同じ長さの特徴ベクトル

  - コサイン類似度(テキストの処理)

$$s_{\cos}(x,y)=\frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

- 距離
  - 似ているほど値が小さい測度
  - ユークリッド距離

$$d_E(x,y) = \sqrt{\sum (x_i - y_i)^2}$$

現実の問題を解決するには、より工夫された類似度、距離が必要

- 距離⇒重み付きのユークリッド距離、
  マハラノビス距離、キャンベラ距離など
- 集計したデータが相対頻度である場合⇒IR距離

$$d_{IR}(x,y) = \sum \left(x_i \log \frac{2x_i}{x_i + y_i} + y_i \log \frac{2y_i}{x_i + y_i} \right)$$

- 類似度は距離に変換して用いることも可能
- コサイン類似度は以下のような変換で距離として用いられる。(定数kは1，2が多く用いられる)

$$d_{\cos}(x,y) = k(1 - s_{\cos}(x,y))$$

- 二値データの類似度
  - 一致係数
  - Jaccard係数

表1　両ベクトル0，1の対応表

		ベクトルy 1	0
ベクトルx	1	n_{11}	n_{10}
	0	n_{01}	n_{00}

2つのベクトルにおける1,0を表1のように集計して計算したとき、jaccard係数とJaccardの距離は

$$s_J = \frac{n_{11}}{n_{11}+n_{10}+n_{01}}, d_J = 1 - \frac{n_{11}}{n_{11}+n_{10}+n_{01}} = \frac{n_{10}+n_{01}}{n_{11}+n_{10}+n_{01}}$$

となる。

多次元尺度法

- 多次元尺度法(MDS)
  - 距離(あるいは類似度)データから何らかの方法で2～3次元に配置する方法
  - 地図上の地点間の距離データから地点のマップを作成する方法
  - いくつかのアルゴリズムがある(古典的多次元尺度法)
  - 距離(あるいは類似度)の測度に依存する

- 古典的多次元尺度法
  - 2点間の距離行列の要素d_{ij}について

$$z_{ij} = d_{ij} - \frac{1}{n}\sum_{i=1}^{n} d_{ij} - \frac{1}{n}\sum_{j=1}^{n} d_{ij} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}$$

のように、あるいはこれに比例するように変換を施したデータの固有値と固有ベクトルを求める方法

本章では、11人が2つのテーマ(住まい、車)について、1000文字前後で書いた作文をテーマ別にグループ分けすることを例とする。作文データを形態素解析し、出現頻度が高い名詞32項目を抽出して用いる。

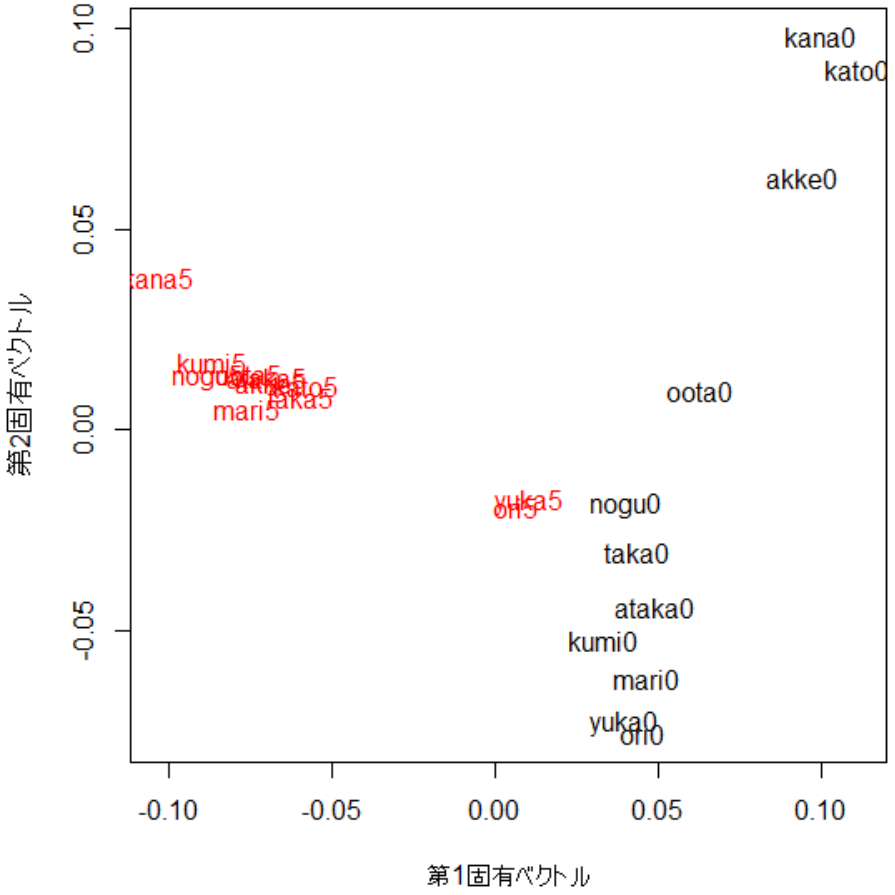


図1 ユークリッド距離

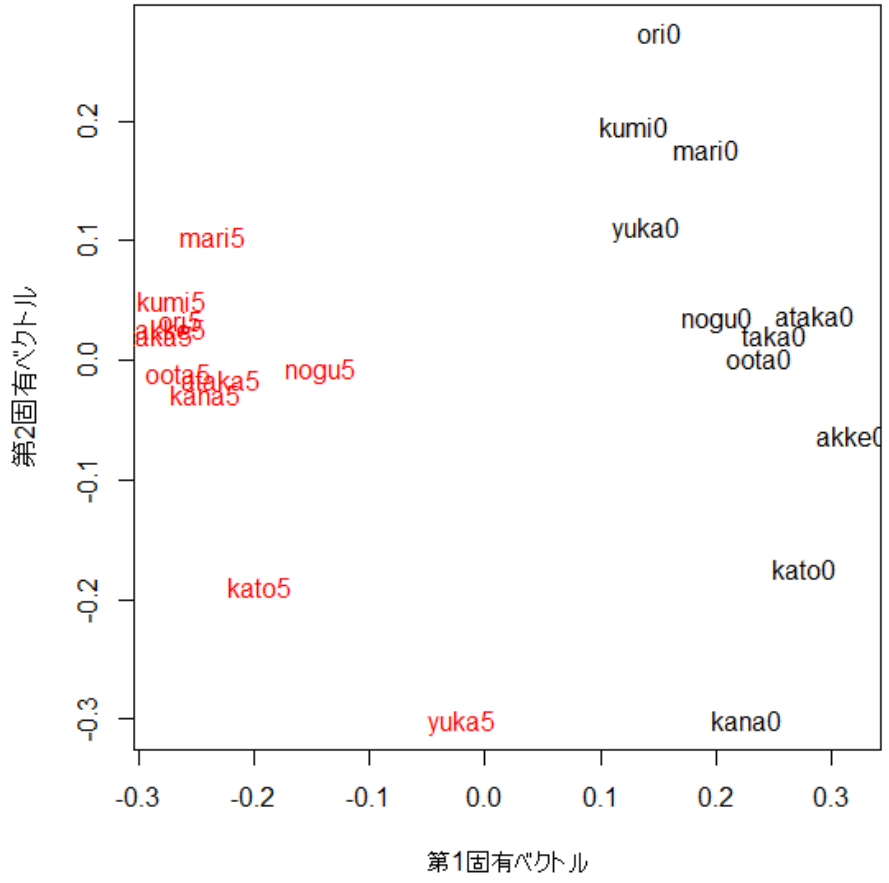


図2 IR距離

- ユークリッド距離を用いた多次元尺度法の結果は、主成分分析の主成分得点の結果に等しい。
- 古典的多次元尺度法をさらに発展させた多次元尺度法としてsammon,isoMDS,metaMDSなどがある。

階層的クラスタリング

- 階層的クラスタリング
  - 距離行列を用いて似ているものを段階的にグルーピングする
  - 単連結法、完全連結法、群平均法、重心法、メディア法、ウォード法などのアルゴリズムがある
  - 大量のデータには不向き
  - アルゴリズムごとに距離データから似ている者同士をグルーピングする方法が異なる
    - 第1段階はすべて同じで、最も距離の値が小さい2つの個体を1つのグループにする
    - 形成されたグループc_iとグループc_jの間の距離を求める方法が異なる。

説明の便利のため、グループc_iに属する任意の個体k($k \in c_i$)とグループc_jに属する任意の個体s($s \in c_j$)の間の距離をd_{ks}とする。

- 単連結法(最近隣法)
  - グループc_iとc_jの個体間の距離の中で最小の距離をグループ間の距離とする。

$$d(c_i, c_j) = \min(d_{ks}), k \in c_i, s \in c_j$$

- 完全連結法(最遠隣法)
  - グループc_iとc_jの個体間の距離の中で最小の距離をグループ間の距離とする。

$$d(c_i, c_j) = \max(d_{ks}), k \in c_i, s \in c_j$$

- 群平均法
  - グループc_iとc_jの個体間の距離の平均値を両グループ間の距離とする

$$d(c_i, c_j) = \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{s=1}^{n_j} d_{ks}, k \in c_i, s \in c_j$$

- ウォード法
  - 比較的多く用いられている
  - 分散の情報を用いる
  - グループ内の分散が小さく、かつグループ間の分散大きい組み合わせでグループ分けする。
  - 全体の偏差の2乗和をT、グループ内の偏差の2乗和をW,グループ間の偏差の2乗和をBで表すと

    T=W+B

    が成り立つ

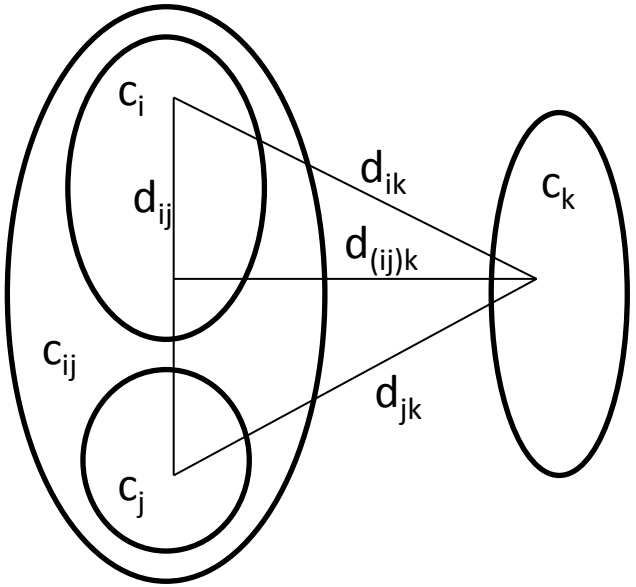


図3 クラスター間の距離

階層的クラスタリング法のグループ間の距離は
つぎの式で求めることができる

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma \left| d_{ik} - d_{jk} \right|$$

$d_{ij}, d_{(ij)k}, d_{ik}, d_{jk}$: 左図に示すクラス管の距離

$\alpha_i, \alpha_j, \beta, \gamma$: パラメータ(係数)

表2　方法とパラメータの対応表(n_i:クラスターc_iの個体数)

方法の名称	α_i	α_j	β	γ
最近隣法	1/2	1/2	0	-1/2
最遠隣法	1/2	1/2	0	0
群平均法	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	1/2
重心法	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$-\alpha_i\alpha_j$	0
メディアン法	1/2	1/2	-1/4	0
ウォード法	$\frac{n_i+n_k}{n_i+n_j+n_k}$	$\frac{n_j+n_k}{n_i+n_j+n_k}$	$-\frac{n_k}{n_i+n_j+n_k}$	0

- 階層的クラスターのグラフを樹形図と呼ぶ

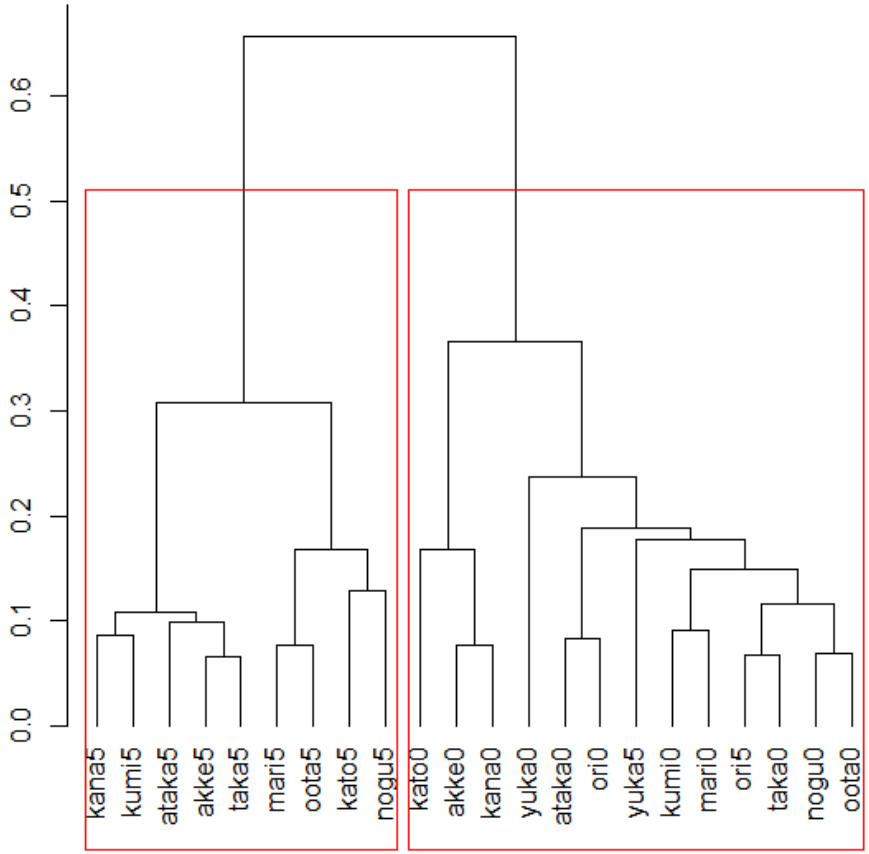


図4　作文のクラスター樹形図(ユークリッド距離、ウォード法)

- 樹形図は大まかに2つのグループに分かれている
- 左側のクラスターが「車」に関する作文であり、右側のクラスターのほとんどが「住まい」に関する作文である
- 実際のテーマごとの分類と完全には一致しない

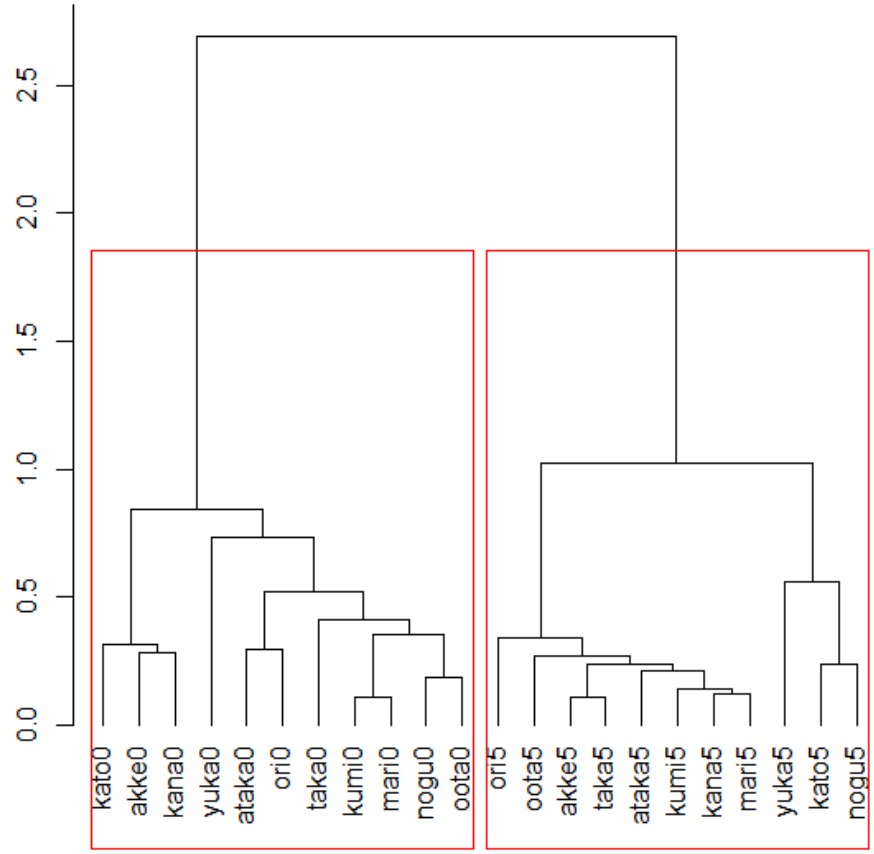


図5　作文のクラスター樹形図(IR距離、ウォード法)

- すべての作文がテーマ別に正しく2つのクラスターに分かれている

- クラスター樹形図を作成する際に用いた変数についてクラスター分析が必要な場合

  ⇒変数のクラスター分析

    - データを転置する
    - 転置したデータセットのキャンベラ距離を求める
    - ウォード法によるクラスター樹形図を作成する
- キャンベラ距離

$$d_C(x,y) = \sum \frac{|x_i - y_i|}{|x_i + y_i|}$$

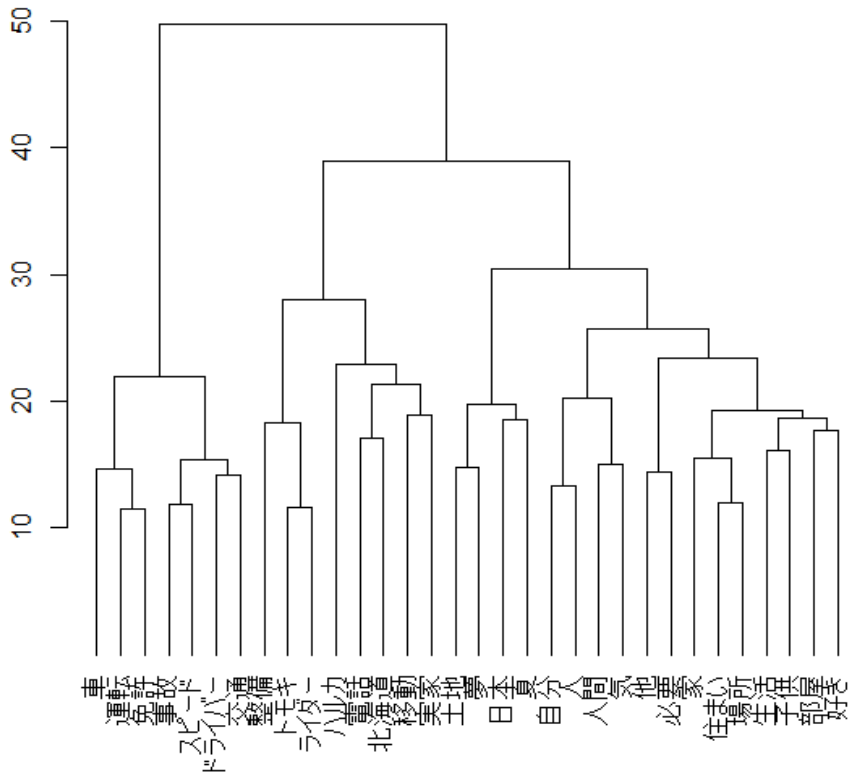


図6 作文の名詞のクラスター樹形図(キャンベりゃ距離、ウォード法)

- 左のクラスターは「車」と関連する語で、右側は「住まい」に関する語

階層的クラスタリングは、用いる距離および方法によって、しばしば得られる結果が異なる

⇒如何に客観的に分析するかが大きな課題

k-means法

- k-means
  - 比階層的クラスター分析
  - 大量のデータのクラスター分析に用いられる
  - データをいくつのグループに分けるかについてユーザが指定することが必要
  - いくつかのアルゴリズムが提案されている。以下に大まかな流れを示す
    1. k個の仮の初期グループの中心を何らかの方法で与える
    2. すべてのデータについてk個のグループの中心との距離を求め、それぞれ最も近いグループに配属させる
    3. 新たに形成されたグループの中心を求める
    4. ステップ2,3を繰り返し、グループおよびグループの中心が前の結果と同じであれば終了する

```
akke0  akke5 ataka0 ataka5  kana0  kana5  kato0  kato5  kumi0  kumi5  mari0  mari5  nogu0
    2      1      1      1      2      1      2      1      1      1      1      1      1
 nogu5  oota0  oota5   ori0   ori5  taka0  taka5  yuka0  yuka5
    1      2      1      1      1      1      1      1      1
> table(rep(2:1,11),sb.km$clust)
     1  2
  1 11  0
  2  7  4
```

誤分類：7個

```
akke0  akke5 ataka0 ataka5  kana0  kana5  kato0  kato5  kumi0  kumi5  mari0  mari5  nogu0
    2      1      2      1      2      1      2      1      2      1      2      1      2
 nogu5  oota0  oota5   ori0   ori5  taka0  taka5  yuka0  yuka5
    1      2      1      2      2      2      1      2      2
> table(rep(2:1,11),sb.km$clust)
     1  2
  1  9  2
  2  0 11
```

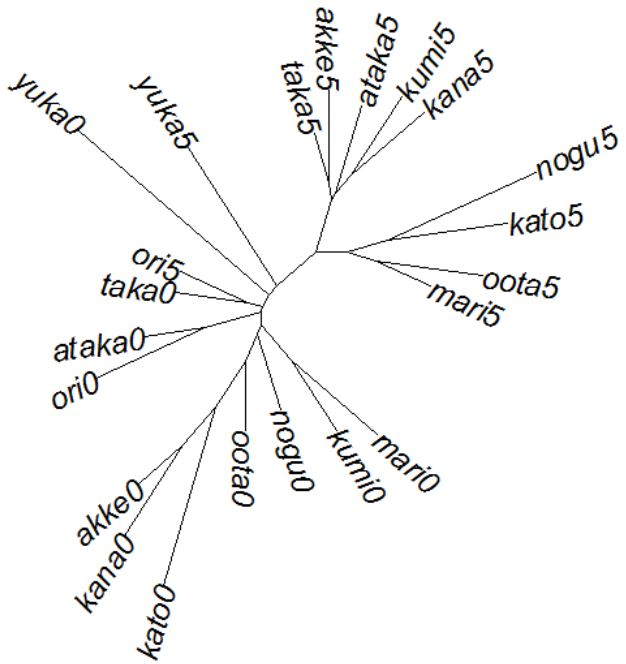
誤分類：2個

- k-means法のグループ分けの精度は決して高くない
- k-means法の長所は、大量のデータにおいて計算が速いことである

系統樹

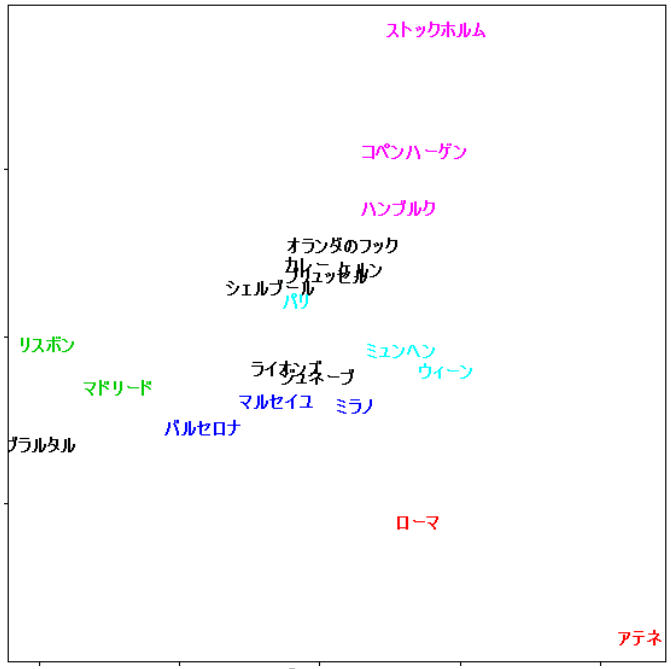
- 系統樹
  - 有根系統樹と無根系統樹に分類される
  - 階層的クラスター樹形図に似ている
  - クラスター分類に用いることもできる
  - 分岐させる方法でより妥当な系統関係を求めるという発想
  - 距離データをもとに作成するNJ法というアルゴリズムがある

- Neighbor-Net法
  - 系統機を発展させたネットワーク・ツリーという方法
  - NJ法を発展させたもの



参考1

- ヨーロッパの21都市間の道路距離のデータをもとに古典的多次元尺度法をRで実行してみる
- eurodistというヨーロッパの21都市間の道路距離が記されているデータセットを用いた



データとして都市間の距離だけをもとにして2次元座標を復元している

- Rのコマンド

```
data(eurodist)
data <- as.matrix(eurodist)
result <- cmdscale(data, k = 2)
citynames <- c("アテネ", "バルセロナ", "ブリュッセル", "カレー", "シェルブール", "ケルン", "コペンハーゲン", "ジュネーブ", "ジブラルタル", "ハンブルク", "オランダのフック", "リスボン","ライオンズ", "マドリード", "マルセイユ", "ミラノ", "ミュンヘン", "パリ","ローマ", "ストックホルム", "ウィーン")
colors <- c(2, 4, 1, 1, 1, 1, 6, 1, 1, 6, 1, 3, 1, 3, 4, 4, 5, 5, 2, 6, 5)
plot(result[,1], -result[,2], type = "n")
text(result[,1], -result[,2], labels = citynames, col = colors, cex = 0.8, font = 2)
```

※データセットeurodist を利用するには mva パッケージを読み込む必要がある。

参考2

- 7人の5教科の成績データを用いて階層的クラスター分析をRで実行してみる

表 3　成績データ

	算数	理科	国語	英語	社会
田中	89	90	67	46	50
佐藤	57	70	80	85	90
鈴木	80	90	35	40	50
本田	40	60	50	45	55
川端	78	85	45	55	60
吉野	55	65	80	75	85
斉藤	90	85	88	92	95

- 成績データのユークリッド距離を示す

	田中	佐藤	鈴木	本田	川端	吉野
佐藤	69					
鈴木	34	81				
本田	60	64	53			
川端	28	61	21	47		
吉野	63	12	76	54	56	
斉藤	68	38	88	92	68	46

- Rのコマンド

```
seiseki<-matrix(c(89,90,67,46,50, 57,70,80,85,90,80,90,35,40,50,
    40,60,50,45,55,78,85,45,55,60,
    55,65,80,75,85,90,85,88,92,95),7,5,byrow = TRUE)
colnames(seiseki)<-c("算数","理科","国語","英語","社会")
rownames(seiseki)<-c("田中","佐藤","鈴木","本田","川端","吉野","
    斉藤")
seiseki.d<-dist(seiseki)
(sei.hc<-hclust(sei.d))
par(mfrow=c(2,2))
plot(sei.hc,main="最遠隣法")
plot(sei.hc,hang=-1,main="最遠隣法")
s.hc2<-hclust(sei.d,method="centroid")
plot(s.hc2,hang=-1,main="重心法")
s.hc3<-hclust(sei.d,method="ward")
plot(s.hc3,hang=-1,main="ウォード法")
```

最遠隣法
斉藤
佐藤
吉野
本田
田中
鈴木
川端

最遠隣法
0
20
40
60
80
斉藤
佐藤
吉野
本田
田中
鈴木
川端

重心法
0
20
40
60
80
100
120
140
斉藤
佐藤
吉野
本田
田中
鈴木
川端

ワオード法
0
20
40
60
80
100
120
140
斉藤
佐藤
吉野
本田
田中
鈴木
川端