Analyzing Linguistic Data 2009/1/22 野村真理子

5.1.5 Tables with distances: hierarchical cluster analysis

＊階層的クラスター分析：データをクラスタリングし、樹形図で表示する手法
インプットとして距離オブジェクトが必要

＊クラスターの作り方

①分岐型クラスタリング：関数 diana( )

すべてのデータ点を含むクラスターから出発して、より小さいクラスターに分割する。
より小さいクラスターへの最適分割が難しいと言われているが、数個の大きなクラスターを見つけるのが目的の場合は、魅力的な方法

②凝集型クラスタリング：関数 hclust( )

１つずつのデータ点から出発して、グループを作り、そのグループをより大きなグループにまとめていく。
クラスタリングは、データ点やデータ点のかたまりをより大きなクラスターに融合するのに使われる基準にかなり依存している。
h clust がどの基準を使用するかはオプション method によって指定される。
R のデフォルト：complete (最長距離法)

＊分析例：data set--english (2233 の英語の単一形態素、単音節の語)を特徴づける 23 の尺度
単語とそれに付いている尺度に関する情報→deta set--lexicalMeasures
尺度に関する情報→?lexicalMeasures / help (lexicalMeasures)

```
> lexicalMeasures[1:5, 1:6]
```
尺度はある程度相関がある

相関行列は cor( )で得られる

```
> lexicalMeasures.cor = cor(lexicalMeasures[, -1])
```
数値でない第 1 列を除く

```
> lexicalMeasures.cor[1:5, 1:5]
```

低いように見える相関でも data set の単語の多さのために有意である。
（例） CelS (frequency)と Ient (inflectional entropy)

```
> cor.test(lexicalMeasures$CelS, lexicalMeasures$Ient)
```

Baayen et al. (2006)：興味がある問題は、語の頻度(CelS)が語の形式の尺度とより強い相関があるか、意味の尺度とより強い相関があるか。

＊階層的クラスター分析：23 の尺度の相関構造を探るのに理想的

・相関関係は、正と負両方があるが、距離行列は正の値のみを持つのが望ましい。

→相関行列を 2 乗する

```
> (lexicalMeasures.cor^2)[1:5, 1:5]
```

・cor( )はかなり対称的なベクトルに最もよく作用するが、尺度の多くはゆがんだ分布（1 つ以上の山がある：mutimodality）

→Spearman correlations を用いる

```
> lexicalMeasures.cor = cor(lexicalMeasures[ ,-1], method="spearman")^2
> lexicalMeasures.cor[1:5, 1:5]
```

この行列を距離オブジェクトに変換する

```
> lexicalMeasures.dist = dist(lexicalMeasures.cor)
```

↓

クラスター分析を行う

➢ 凝集型クラスタリング

クラスター分析 hclust( )→樹形図（デンドログラム）作成 plclust( )

```
> lexicalMeasures.clust = hclust(lexicalMeasures.dist)
> plclust(lexicalMeasures.clust)
```

分析結果：ある構造を示しているが、形式の尺度と意味の尺度の明確な区別は得られなかった

➢ 分岐型クラスタリング

diana( ) (cluster package より)のアウトプットを pltree( )に入れる

```
> library(cluster)
> pltree(diana(lexicalMeasures.dist))
```

分析結果：頻度(CelS)は語の形式の尺度と並んでいない

＊ 変数がどのクラスターに割り当てられているか知りたいとき

クラスターがいくつ必要だと思うかを決め、この数を cutree( )の 2 番目の引数として使う

```
> cutree(diana(lexicalMeasures.dist), 5)
```

尺度の名前と data set—lexicalMeasuresClasses における尺度の分類と結びつけると、変数の分類とクラスターの番号の間に非常に密接な対応関係がある(Fdif は唯一の例外)。

```
> x = data.frame(measure = rownames(lexicalMeasures.cor),
+cluster = cutree(diana(lexicalMeasures.dist), 5),
+class = lexicalMeasuresClasses$Class)
> x = x[order(x$cluster), ]
> x
```