言語処理のための機械学習入門

3章　クラスタリング
3.1～3.3

茨城大学工学部情報工学科
08T4038G　篠塚 晃一

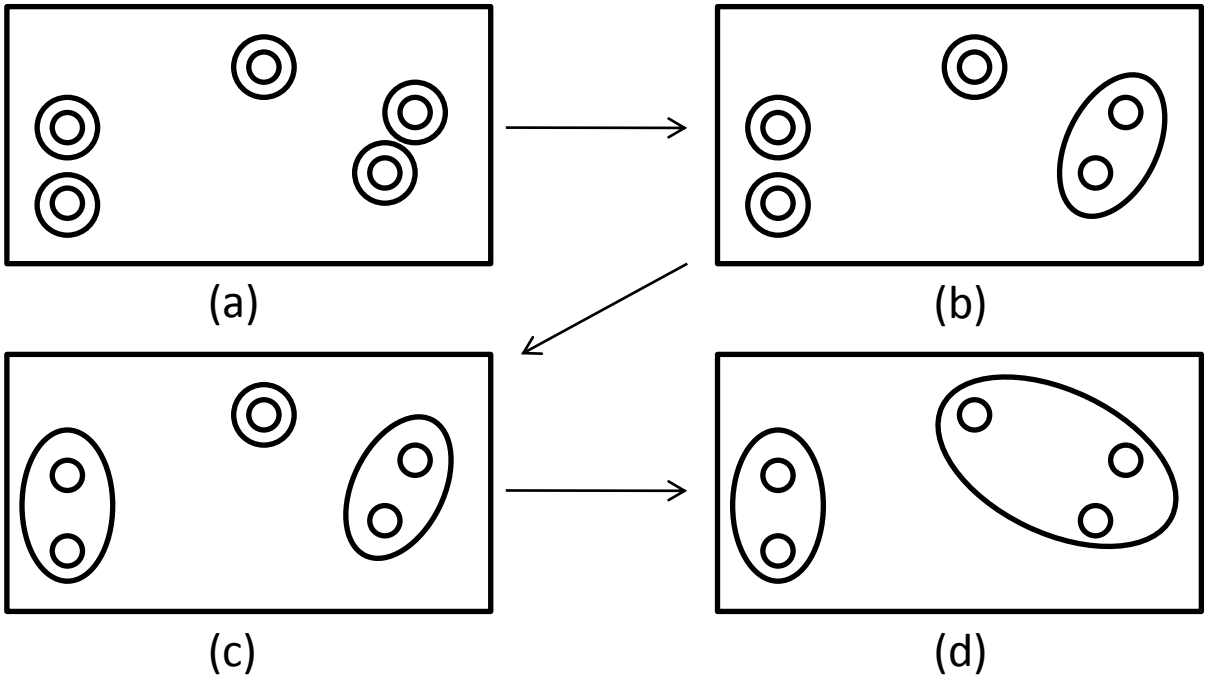
3.1　準　備

- クラスタリング（clustering）
  似ている文書同士や単語同士を一つのグループにまとめる作業。
- クラスタ（cluster）
  クラスタリングによってできあがったグループ。
- 学習（learning）、訓練（training）
  データが与えられ、そこからなんらかのモデルや処理手段を導くこと。
  クラスタリングは学習の一種。
- 学習データ（learning data）、訓練データ（training data）
  学習に用いるデータ。
- 訓練事例（training instance）
  訓練データ中の事例。

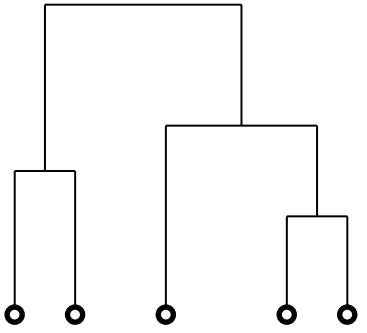
3.2　凝集型クラスタリング

- 単純に最も似ているもの同士をくっつけていくという直感的なクラスタリングの方法。
- ボトムアップクラスタリングなどとも呼ばれる。

例



樹形図で表すと



凝集型クラスタリングのアルゴリズム

入力：事例集合 $D = \{x^{(1)}, x^{(2)}, \cdots, x^{(|D|)}\}$

$C = \{c_1, c_2, \cdots, c_{|D|}\}$

１つのクラスタに１つの事例を割り当てる

$c_1 = \{x^{(1)}\}, c_2 = \{x^{(2)}\}, \cdots, c_{|D|} = \{x^{(|D|)}\}$

while $|C| \geq 2$　# 停止条件

　# もっとも似ているクラスタ対をみつける

　$(c_m, c_n) = \arg\max_{c_i, c_j \in C} sim(c_i, c_j)$

　# みつかったクラスタ対を融合する

　merge (c_m, c_n)

end while

（ただし、*sim*は二つのクラスタの類似度を表す）

クラスタの融合

- はじめは似ている事例同士を融合する。
- クラスタリングが進むとクラスタと事例、または似ているクラスタ同士を融合する。

↓

- 事例同士の場合
  例えば、それぞれの事例ベクトルの予言類似度を測って最も似ているクラスタ同士を見つける。

- 一般にクラスタ同士の場合
  どのように計算したらよいか自明ではない。

クラスタ同士の類似度を測る方法

- 単連結法（single-link method）
- 完全連結法（complete-link method）
- 重心法（centroid method）

事例同士 (d_k, d_l) の類似度 $sim(d_k, d_l)$ は
定義されているとする。

単連結法（single-link method）

- 二つのクラスタが与えられたとき、その中で最も近い事例対の類似度を、その二つのクラスタの類似度とする方法である。

$$sim(c_i, c_j) = \max_{x_k \in c_i, x_l \in c_j} sim(x_k, x_l).$$

- 特徴

  クラスタが長く伸びてしまおうがお構いなしに融合していく。

完全連結法（complete-link method）

- 二つのクラスタが与えられたとき、その中で最も遠い事例同士の類似度を、その二つのクラスタの類似度とする方法である。

$$sim(c_i, c_j) = \min_{x_k \in c_i, x_l \in c_j} sim(x_k, x_l).$$

- 特徴

  鎖のように長く伸びたクラスタを嫌う。

  （そのようなクラスタができにくい）

重心法（centroid method）

- 各クラスタは、それが含む事例の重心ベクトルにより代表されているとすると、与えられた二つのクラスタに対し、それらの重心間の類似度をこれらのクラスタ間の類似度とする方法。

$$sim(c_i, c_j) = sim\left(\frac{1}{|c_i|}\sum_{x \in c_i} x, \frac{1}{|c_j|}\sum_{x \in c_j} x\right).$$

- 特徴

  二つの方法の中間的な性質を持つ。

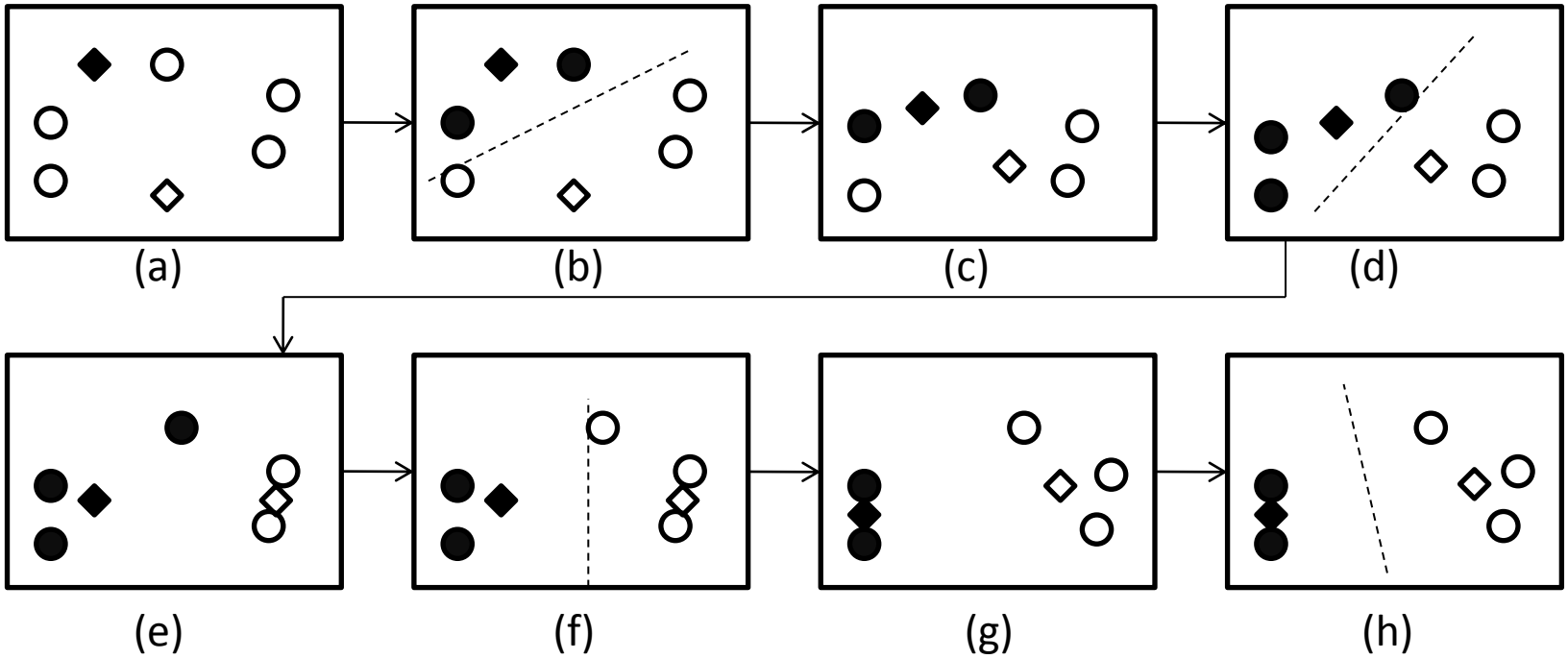
3.3 k-平均法

- とりあえず適当に分けてしまって、それからよりうまく分かれるように調整していくことによって、クラスタリングを行う方法。

- クラスタ数k（いくつに分けたいか）はユーザが前もって決定する。

- 各クラスタは平均ベクトルなどの代表ベクトルで表現される。

例（各ベクトルが2次元でクラスタ数が2、つまりk=2のとき）



1. 無作為に代表ベクトルを決める。
2. どちらの代表ベクトルに近いかという基準に従って各事例ベクトルをどちらかのクラスタに帰属させる。
3. 各クラスタに含まれている事例ベクトルの平均を計算し、これを新たな代表ベクトルとし、この代表ベクトルに従って事例ベクトルを再び二つに分ける。
4. これを収束するまで繰り返す。

k-平均法のアルゴリズム

入力：事例ベクトル集合 $D = \{x^{(1)}, x^{(2)}, \cdots, x^{(|D|)}\}$

　　：クラスタ数 k

無作為に $m_1, m_2, \cdots, m_k$ を決定。

until 収束

　foreach $x^{(i)} \in D$

　　$\forall c, c_{\max} = \arg\max_{c} sim(x^{(i)}, m_c)$ # 事例ベクトル集合の分割

　　insert $x^{(i)}$ into $c_{\max}$

　end foreach

　$\forall c, m_c = \frac{1}{|c|} \sum_{x^{(i)} \in c} x^{(i)}$ #代表ベクトルを再計算

end until