Chapter 2 Section 7
CLUSTERING ALGORITHMS

BOOK : Semi-Supervised Learning and Domain Adaptation in Natural Language Processing

Author : Anders Soegaard

茨城大学大学院理工学研究科情報工学専攻

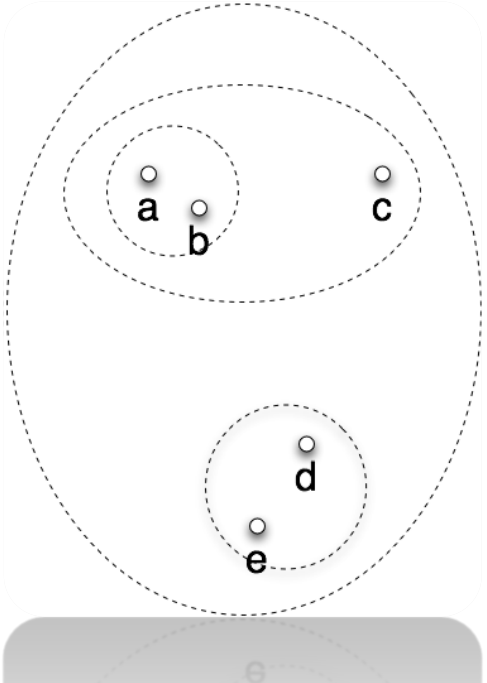
菊池裕紀

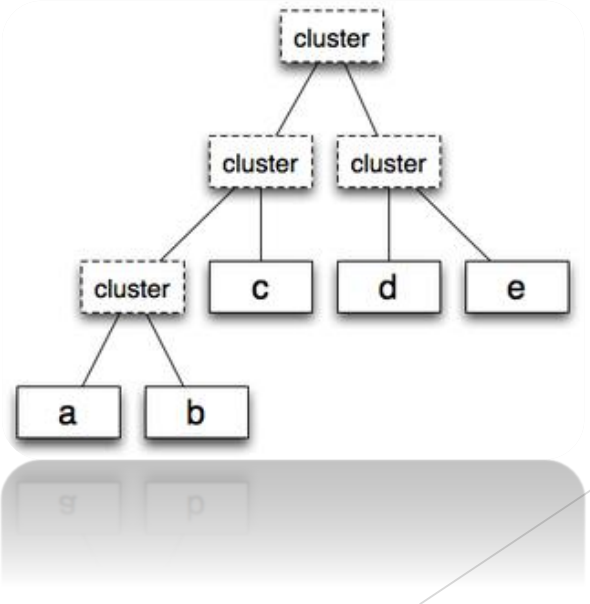
HIERARCHICAL CLUSTERING(1)

- What's the "HIERARCHICAL CLUSTERING" ?

  The method of building hierarchies of cluster

  Together two data points which have shortest distance one by one





HIERARCHIICAL CLUSTERING(2)

- The method of calculating distance

Single linkage $: d(c_j, c_k) = \min_{\boldsymbol{x} \in c_j, \boldsymbol{x}' \in c_k} d(\boldsymbol{x}, \boldsymbol{x}')$

All linkage $: d(c_j, c_k) = \max_{\boldsymbol{x} \in c_j, \boldsymbol{x}' \in c_k} d(\boldsymbol{x}, \boldsymbol{x}')$

Average linkage $: d(c_j, c_k) = \frac{1}{|C|} \sum_{\boldsymbol{x} \in c_j, \boldsymbol{x}' \in c_k} d(\boldsymbol{x}, \boldsymbol{x}')$

Ward's method $: \Delta(c_j, c_k) = \sum_{i \in c_j \cup c_k} d\left(x_i - m_{c_j \cup c_k}\right)^2 - \sum_{i \in c_j} d\left(x_i - m_{c_j}\right)^2 - \sum_{i \in c_k} d\left(x_i - m_{c_k}\right)^2$

k-MEANS

- What's the "k-means" ?
  1. randomly select k data points
  2. alternate all data points to their nearest centroid (E-step)
  3. Compute the actual centroids of these clusters (M-step)
  4. GO TO 2STEP

  SAMPLE PAGE → http://tech.nitoyon.com/ja/blog/2013/11/07/k-means/

EXPECTATION MAXIMIZATION

- EM for Gaussian mixture model

$$P(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

1. Set parameter($\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$) at random
2. Compute $P(y_j|x, \theta)$ with $j \in \{1,2\}$ (Baye's rule)
3. Reestimate $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ from distribution
4. Apply the model to data
5. Exit after convergence (if not GO TO STEP2)

EVALUATING CLUSTERING ALGORITHMS

- F-measure

  Defined as Harmonic means of precision and recall

- V-measure

  Defined as Harmonic means of homogeneity and completeness