Recent Developments in Document Clustering 2. Good, Bad, and Ugly Clustering

茨城大学工学部情報工学科
佐々木稔

クラスタリングの良し悪し

- クラスタリングを評価するとは？
  - 処理結果と正解データと比較を行う
- クラスタリングを評価する手順
  - 解答付きデータ集合を用意する
    - 自作、もしくは既存研究で使われた公開データ集合
  - 解答を見ずにデータを分類する
    - 新たな観点による自動的な分類を試行
  - 結果を解答と比較する
    - 実験結果と人手による分類結果の相違を調査
    - その観点が有効かどうか判断

クラスタリングの評価方法

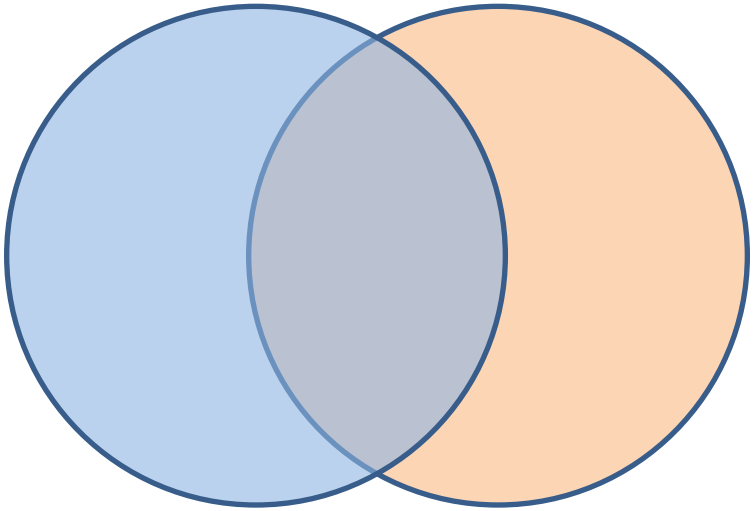
- クラスタリング結果の評価方法
  - 「これが最良」という方法はない
- 目的に応じて評価尺度を選ぶのが実状
  - 一般的に使われるのは「適合率」と「再現率」
  - クラスタリングでは「Purity」と「エントロピー」
  - 人工知能では「相互情報量」
  - 情報検索では「F値」
- 本節の内容
  - 上記の様々な評価尺度を紹介

再現率

- 全関連文書に対する検索された関連文書の割合

$$\mathrm{Recall} = \frac{|A \cap S|}{|A|}$$

検索された
文書集合 S



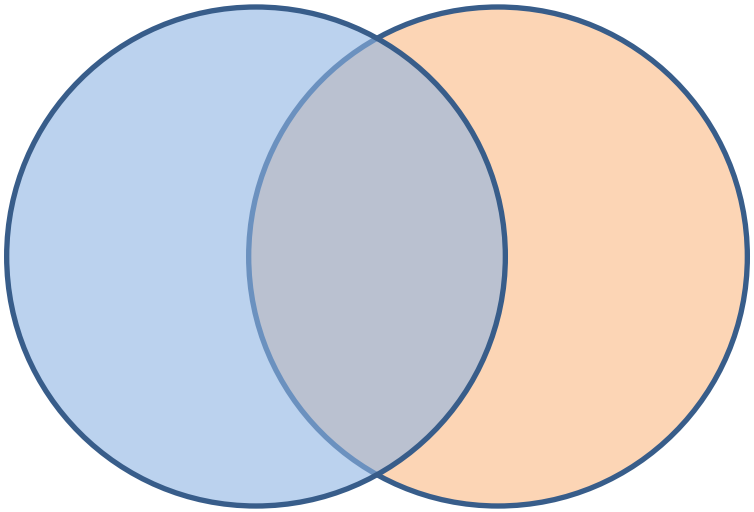
正解の関連
文書集合 A

適合率

- 検索された文書に対する関連文書の割合

$$\mathrm{Precision} = \frac{|A \cap S|}{|S|}$$

検索された
文書集合 S



正解の関連
文書集合 A

F-measure（F値）

- 検索結果に全文書を取り出す
  - 再現率は必ず1になって無意味
- F-measure
  - 再現率 R と適合率 P を組み合わせた指標

$$F_\alpha = \frac{(1+\alpha)RP}{\alpha P + R}$$

  - α=1 のとき、R と P を同じ割合で重み付け
  - α=2 のとき、R と P を 2:1 の割合で重み付け
  - α=0.5 のとき、R と P を 1:2 の割合で重み付け

再現率、適合率のクラスタリングへの拡張

- 複数の正解文書集合を用意する
  - 検索の場合は正解文書集合はひとつ
- 複数集合の文書に対するクラスタリング結果
  - 正解となる集合 *i*
  - 得られたクラスタ集合 *j*
- クラスタリング結果の F 値
  - *n* は全文書数

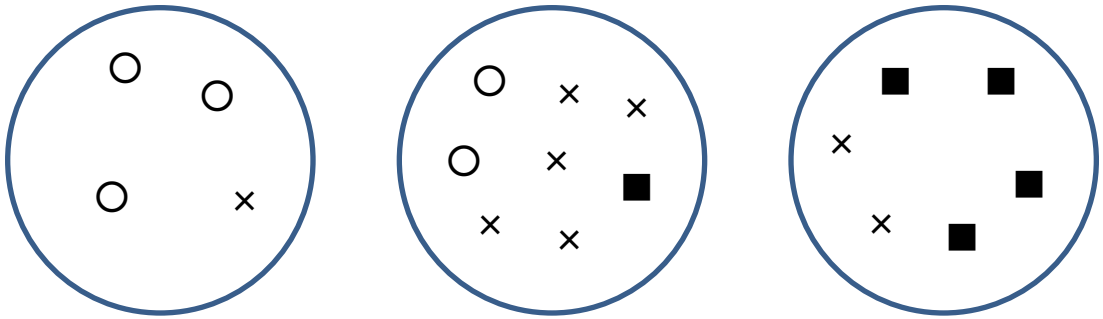
$$F = \sum_i \frac{n_i}{n} \max_j F(i, j)$$

Purity

- クラスタ内で最大の正解集合となる文書数の割合の重み付き和

$$\mathrm{Purity} = \sum_j \frac{n_j}{n} \max_i P(i, j)$$

- 下図の場合
  - クラスタ1は3、クラスタ2は5、クラスタ3は4
  - Purity=4/18×3/4+8/18×5/8+6/18×4/6≒0.67



エントロピー

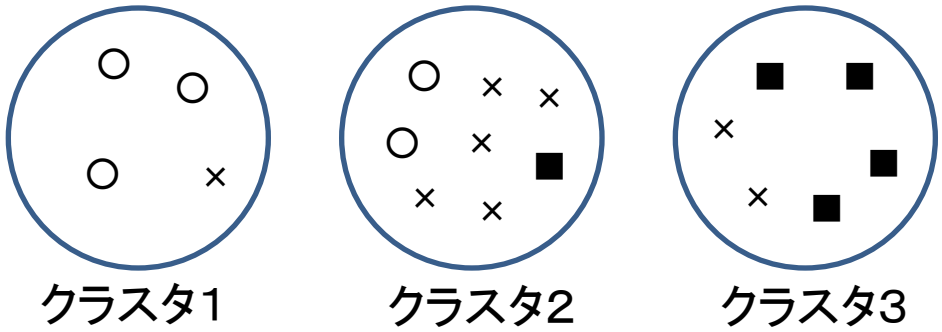
- 各クラスタにおける正解集合の分布度合
  - 0～1の値を取り、小さいほど良い

$$\mathrm{Entropy} = -\frac{1}{\log k}\sum_j \frac{n_j}{n}\sum_i P(i,j)\log P(i,j)$$

- 下図の場合、

$$-\frac{1}{\log 3}\left(\frac{4}{18}\times\left(\frac{3}{4}\log\frac{3}{4}+\frac{1}{4}\log\frac{1}{4}\right)+\frac{8}{18}\times\left(\frac{2}{8}\log\frac{2}{8}+\frac{5}{8}\log\frac{5}{8}+\frac{1}{8}\log\frac{1}{8}\right)+\frac{6}{18}\times\left(\frac{2}{6}\log\frac{2}{6}+\frac{4}{6}\log\frac{4}{6}\right)\right)$$

$$\approx 0.67$$



Normalized Mutual Information（正規化相互情報量）

- 正解集合の数とクラスタ数が異なる場合
  - PurityやEntropyは良い指標ではない場合がある
  - 正規化相互情報量を指標として使う方が良い
- 正規化相互情報量
  - クラスタリング結果のラベル集合 Y
  - 正解文書のラベル集合 Y’

$$\mathrm{NMI} = \frac{I(Y;Y')}{\sqrt{H(Y)}\sqrt{H(Y')}}$$

Normalized Mutual Information（正規化相互情報量）

- 詳細な定義
  - n_h : 正解の集合 h の文書数
  - n_l : クラスタ l の文書数
  - $n_{h,l}$: クラスタ l 中で正解集合 h に属する文書数

$$NMI = \frac{\sum_{h,l} n_{h,l} \log \frac{n \cdot n_{h,l}}{n_h n_l}}{\sqrt{\sum_h n_h \log \frac{n_h}{n}} \sqrt{\sum_l n_l \log \frac{n_l}{n}}}$$

- NMIの値は、0～1 の範囲
  - 完全に一致すれば、NMIの値は 1 となる

クラスタリング結果の安定性

- クラスタリング処理
  - 実行ごとに変化は少ない方が良い
  - 複数回の実行における評価値の平均を取る
- 正規化相互情報量での例
  - r 回のクラスタリング結果集合 Λ
  - 1 回のクラスタリング結果 λ

$$\varphi(\Lambda, \lambda) = \frac{1}{r} \sum_i \mathrm{NMI}(\lambda, \lambda_i)$$

Confusion Matrix（混同行列）

- 正解データとクラスタリング結果の対応表
  - 行はクラスタリング結果に与えたラベル
  - 列は正解文書に割り当てられていたラベル

	A	B	C
A	2	2	0
B	2	2	0
C	0	0	8

ソフトクラスタリングの評価

- 複数クラスタへの割り当てを認める場合
  - ファジィクラスタリング
  - データに各クラスタへの「帰属度」が存在

	A	B	C
a	0.7	0.3	O
b	0.2	0.4	0.4
c	O	O	1

ソフトクラスタリングの評価

- 複数クラスタに分布すると評価が難しい
  - 複数クラスタへの分布を一か所に限定する
    - *“hardening the clusters”* と呼ばれる
- 評価方法
  - 帰属度にしきい値を設定
    - しきい値を超えるクラスタに文書を割り当てる
  - 帰属度が最大のクラスタに限定

おわりに

- クラスタリングの一般的な評価方法
  - 「Purity」と「エントロピー」
  - クラスタ数が分かっている場合には有効
- クラスタ数が未知の場合の評価方法
  - 「正規化相互情報量」
- 同じ処理でも結果が異なる場合の評価方法
  - 評価値の平均値を計算する
- 分類先のラベルを調べたい場合
  - 「混同行列」を作成する