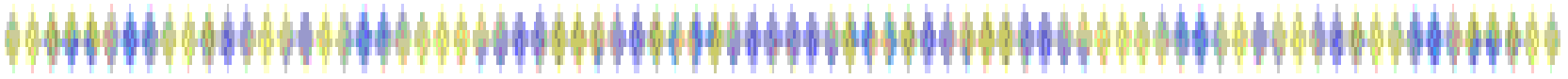
文書クラスタリングの技法ゼミ

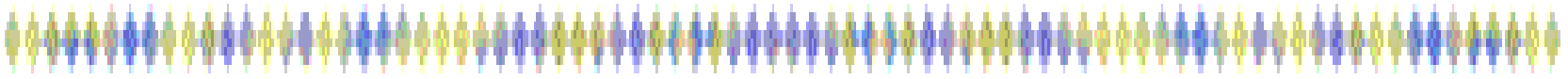


D. 確率モデルに基づくクラスタリング

発表日 : 2007/6/6

発表者 : 阿部 竜之介

確率モデルに基づくクラスタリング

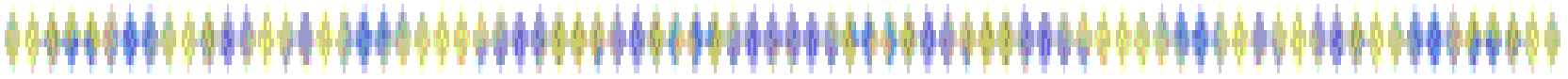


- 何らかの確率分布を仮定して、文書 d_i に対する各クラスタの確率 $P(C_k \mid \mathrm{d}_i, \Theta)$ （ $k = 1,\ldots,L$ ）を推計し、その値が最大であるクラスタにその文書を含める

  Θ は確率分布のパラメータ（のベクトル）

- Liuらによって提案された方法では、Θ をEMアルゴリズムを用いて推計し、確率分布として多次元ガウス分布を仮定している

Liuらによるクラスタリング法(1)



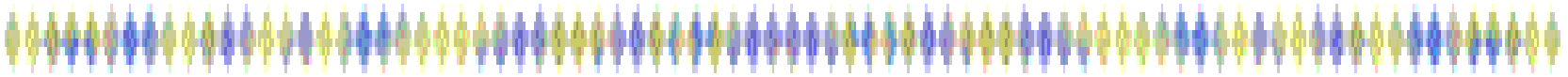
アルゴリズム

(1)初期設定として、クラスタ個数 L と $\beta(\mathrm{f}_j)$ に対する閾値とを設定する。

(2)GMM+EMアルゴリズムによって、初期的なクラスタ群を生成する。

(3)与えられたクラスタ群に対する各特長の $\beta(\mathrm{f}_j)$ の値を計算し、閾値を超える特徴の集合 Γ を求める。

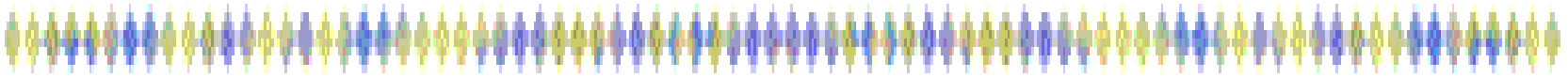
Liuらによるクラスタリング法(2)



(4)各文書に含まれる「識別力のある特徴」$\mathrm{f}_j(\in\Gamma)$に対して、それが最頻出するクラスタをそれぞれ調べ、その中で、最頻出のクラスタとして最も多く挙げられたものを、その文書が属する新たなクラスタとして割り当てる。

(5)新たに得られたクラスタ群とその前段階のクラスタ群とを比較し、変化がなければ処理を終了する。そうでなければ(3)に戻る。

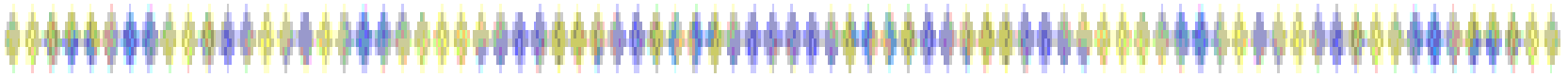
Liuらによるクラスタリング法(3)



(1)の詳細解説

- クラスタの個数はLiuらでは、モデル選択の手法を使って最適な L を見つける方法も考案されている
- $\beta(\mathrm{f}_j)$ は特定のクラスタのみに出現し、他のクラスタには出現しないような識別力のある特徴 f_j がクラスタ C_k に出現する回数に基づいて定義された指標
- この指標に関する詳細は元の文献を参照

Liuらによるクラスタリング法(4)



(2)の詳細解説

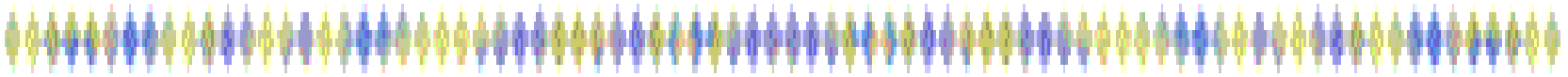
■GMM（ここでは、重心ベクトル m_k と共分散行列 Σ_k をパラメータとする M 次元ガウス分布）

$$P(\mathrm{d}_i \mid C_k) = \frac{1}{(2\pi)^{M/2}\left|\Sigma_k\right|^{1/2}} \times \exp\left(-\frac{1}{2}(\mathrm{d}_i - \mathrm{m}_k)^T \Sigma_k^{-1}(\mathrm{d}_i - \mathrm{m}_k) \right) \quad (33)$$

■各文書ベクトルは L 個のクラスタから構成されるモデル M から確率的に生成されると考える。すなわち、

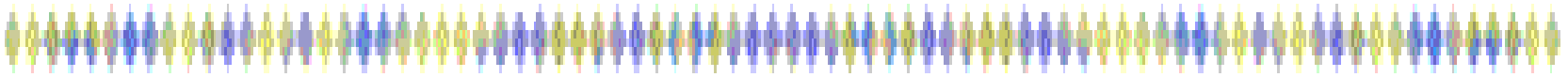
$$P(\mathrm{d}_i \mid M) = \sum_{k=1}^{L} P(C_k) P(\mathrm{d}_i \mid C_k), \quad i = 1, \ldots, N \quad (32)$$

Liuらによるクラスタリング法(5)



- L 個の重心ベクトルの初期値は、$\mathrm{m}_0 = N^{-1}\sum_{i=1}^{N}\mathrm{d}_i$ および $\Sigma_0 = N^{-1}\sum_{i=1}^{N}(\mathrm{d}_i - \mathrm{m}_0)(\mathrm{d}_i - \mathrm{m}_0)^T$ をパラメータとする正規分布からの無作為抽出によって設定し、また、Σ_k については、すべて等しい初期値とし、Σ_0 を使い、(32)式が最大になるようなパラメータ群をEMアルゴリズムを適用して求める
- パラメータが求められたら、各文書の属するクラスタを (33)式を使って決定
- EMアルゴリズムの詳細は記述なし（EMアルゴリズムは反復法によって局所最適解を求めるアルゴリズム。この場合どのように用いるか分からなかった）

Liuらによるクラスタリング法(6)



特徴

- 確率的なクラスタリングの結果を事後的に反復計算によって精緻化している点
- これが必要になるのは学習用データのない状況において確率的な方法を適用することの難しさによるものと推察される