カーネル多変量解析
第3章
3.3 クラスタリング

佐々木研究室
06T4073R 三上 健太

クラスタリング

- 高次元空間データを縮約する方法
  - 低次元空間に射影する ex.主成分分析
  - いくつかの離散点で代表させる
- クラスタリング
  - 高次元データをまとまりごとに集めてグループ分けすること
  - グループ分けした高次元中のいくつかの領域をまとめて1つの点で代表させる
- 本項では
  - カーネルk-means法(k-平均法)
    - k-means法をカーネルを用いて一般化
  - スペクトラルクラスタリング

  について述べる

カーネルk-平均法

- グループ分け対象
  - サンプル点集合： $x^{(1)}, x^{(2)}, \cdots, x^{(n)}$

    特徴ベクトル： $\phi(x^{(1)}), \phi(x^{(2)}), \cdots, \phi(x^{(n)})$
- 代表点の個数：c

  代表点： $\mu_1, \mu_2, \ldots, \mu_c$
- それぞれのサンプル点はその点に最も近い代表点のグループに入る
- 代表点はグループに属するサンプル点の重心
- N_i ：代表点 μ_i に対応するグループのメンバー集合

$$N_i = \left\{ x^{(l)} \middle| \mu_i = \arg\min \left\| \phi(x^{(l)}) - \mu_j \right\|^2 \right\} \qquad \mu_i = \frac{1}{|N_i|} \sum_{x^{(j)} \in N_i} \phi(x^{(j)})$$

カーネルk-平均法(2)

- このとき，代表点からメンバーへの二乗距離の総和ができるだけ小さくなるように代表点を決める
  - 目的関数 $L=\sum_{i=1}^{c}\sum_{x^{(j)}\in N_i}\left\|\phi(x^{(j)})-\mu_i\right\|^2$ を最小化するように N_i, μ_i を決める問題
- k-平均法では，代表点とグループの両方を一度に最適化することは難しい
  - 適当な初期値からスタートし，一方を固定して他方を最適化するという交互最適化を行い，局所最適解を求める
- まず，どのグループに属しているかを判定する
  - 特徴ベクトルと代表ベクトルの距離を計算する

カーネルk-平均法(3)

$$\left\|\phi(x^{(j)})-\mu_i\right\|^2=\left\|\phi(x^{(j)})-\frac{1}{|N_i|}\sum_{x^{(l)}\in N_i}\phi(x^{(l)})\right\|^2$$

$$=k(x^{(j)},x^{(j)})-\frac{2}{|N_i|}\sum_{x^{(l)}\in N_i}k(x^{(j)},x^{(l)})+\frac{1}{|N_i|^2}\sum_{x^{(l)}\in N_i}\sum_{x^{(m)}\in N_i}k(x^{(l)},x^{(m)})$$

- カーネル関数だけを使って書き表せる
  これをもとに新たなグループ分け N_i が求められる

- カーネルk-平均法のアルゴリズム

(1)サンプルを適当にc個のグループに分け N_i を初期化する
(2)上式に基づいて N_i を更新する
(3)グループ分けが収束するまでステップ(2)を繰り返す

スペクトラルクラスタリング

- k-平均法の欠点
  - 反復演算を必要とする点
  - 収束解が必ずしも目的関数を最適にするものではない
    → クラスタリングの問題を固有値問題として定式化することで問題点の解決を図る

- スペクトラルクラスタリング
  - 整数計画問題を緩和した固有値問題を解くことによってクラスタリングを行う手法

スペクトラルクラスタリング(2)

- 簡単のため，2つのグループだけの場合を考える
- それぞれのグループに1,-1という離散値を割り当てると，クラスタリングは各サンプル $x^{(i)}$ に対してグループに対応するラベル $\beta_i = \pm 1$ を割り当てる問題と捉えることができる

- ここで，ラプラシアン固有マップ法で導入したグラフ構造を考える

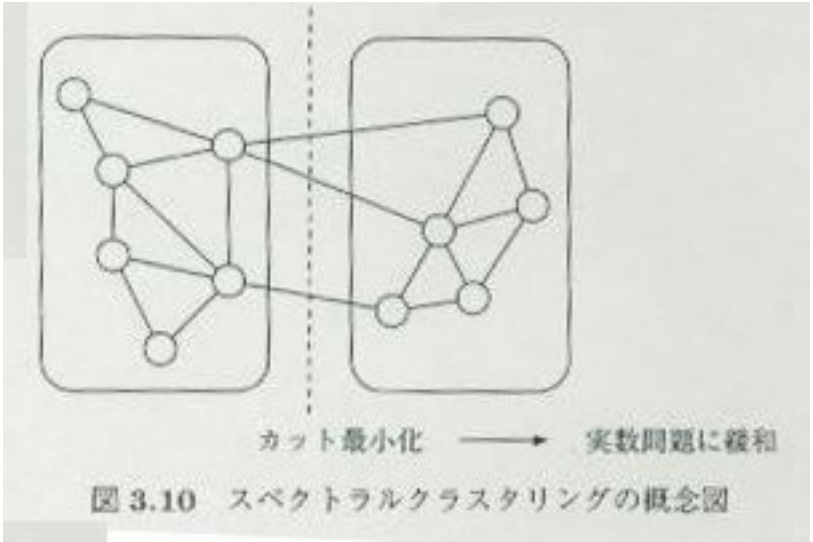

図 3.10　スペクトラルクラスタリングの概念図

  - 各頂点がサンプル点
  - 枝にはサンプル同士の近さを表す重みがついている

スペクトラルクラスタリング(3)

- グループ内はできるだけ近いもの同士が集まり，グループ間は遠く離れていることが望ましい
  - カットの重みの合計は小さいほどよい

$$\min_{\beta} \sum_{i,j} K_{ij}(\beta_i - \beta_j)^2 = 2\beta^T P\,\beta, \qquad \beta = \pm 1$$

- ここで対角行列Λを $\Lambda_{ii} = \sum_{j=1}^{n} K_{ij}$ と定義すると $P = \Lambda - K$ とかける

スペクトラルクラスタリング(4)

- β は2値ベクトルという制約がある
  - 整数計画問題と呼ばれ一般に解くのが困難
  - 整数という制約を取り払い任意の実数ベクトルに制約を緩める
  - ただし β の大きさは制約($\beta^T \Lambda \ \beta = 1$)
    → 制約しないといくらでも小さい値をとり得るから
    → ラプラシアン固有マップ法と完全に等価な最小化問題

- この場合も最小固有値0が存在し，それに対応する固有ベクトルは $\beta \propto 1$ である
  - 全てのサンプルを1つのグループにまとめてしまうという意味のない解
  - 実際には2番目以降の固有ベクトルを使ってクラスタリング

クラスタリング

- クラスタリング
  -最終的には実数ベクトルで得られたベクトルを離散化する

- そのための手法は様々
  - クラスタリング結果にどんな性質を期待するかによって変わる
  ex. 固有ベクトルの成分 $\beta_1,\cdots,\beta_n$ を並べて，これをある閾値で切ってグループ分けする方法
  - この場合閾値の決め方が重要となる
  - 必要に応じて規準を作りそれを最適化するように閾値を決定する方法がいろいろと提案されている

クラスタリング実行結果