# 多クラスに対するスペクトラルクラスタリング

06T4029T 斉藤優

# 研究内容

- 多クラスに対するスペクトラルクラスタリング
  Mcutを用いたスペクトラルクラスタリングを用いてデータを多クラスに分類
  $\Longrightarrow$ 固有値問題を複数回も解くのは大変

  そこで1回の固有値問題で解くスペクトラルクラスタリングの手法を考える

- 多クラスの分類を以下の手法を使って比較

  $\left\{\begin{array}{l}\text{Mcutを用いたスペクトラルクラスタリング}\\ \text{1回の固有値問題で解くスペクトラルクラスタリング}\\ \text{k-means}\end{array}\right.$

# スペクトラルクラスタリング

- 個々のデータを頂点としたグラフの分割問題を解いてデータを分類する
- データ間の辺にはあらかじめ類似度の重みを設定しておく

# Mcut(1)

- 以下の評価関数を最小にするようグラフを分割

$$Mcut = \frac{cut(A,B)}{\text{Aの類似度の総和}} + \frac{cut(A,B)}{\text{Bの類似度の総和}}$$

$cut(A,B)$はAB間の類似度の総和
目的のクラス数になるまで再帰的に$Mcut$で2分割する

- $Mcut$は以下の行列$L$の固有値問題に解くことに等しい

$$L = I - D^{-1/2} W D^{-1/2}$$

$W$は類似度を要素とした隣接行列
$D$は$i$番目のデータとその他のデータとの類似度の和を$i$番目の対角要素とする行列

# Mcut(2)

- $L$の2番目に小さい固有値に対応する固有ベクトル
  この固有ベクトルの$i$番目の要素が正であれば$i$番目のデータをAへ,負であればBに分類する

- 再帰的に2分割
  AやBのように形成されたサブグラフのうち類似度平均の低い方を固有値問題を解いて2分割する
  これを目的のクラス数が得られるまで繰り返す

# 1回の固有値問題で解くスペクトラルクラスタリング

- アルゴリズム
  入力：デース数$n$の$n \times m$行列$S$, 目標のクラスタ数$k$

  1. $Mcut$と同じように$S$から$L$を算出
  2. $L$から固有値問題を解いて,固有値が小さい方から$k$個取り出し対応する固有ベクトル$v_1, \cdots, v_k$を得る
  3. $v_1, \cdots, v_k$を列とする$n \times k$行列$V$の作成
  4. $V$の行を1に正規化した行列$U$を作成

     $$u_{ij} = v_{ij}/(\sum_k v_{ik}^2)^{1/2}$$

  5. $U$の行をデータに対応させ,k-meansを用いて$k$個のクラスタにデータを分類

  出力：$k$個に分類されたクラリスタリング結果

# 比較

- 比較に用いた行列データ

| | tr23 | tr31 | tr41 | tr45 |
|---|---|---|---|---|
| 行数(データ数) | 204 | 927 | 878 | 690 |
| 列数(次元数) | 5832 | 10128 | 7454 | 8261 |
| 非0の要素数 | 78609 | 248903 | 171509 | 193605 |
| 目標のクラスタ数 | 6 | 7 | 10 | 10 |

全てのデータは1に正規化されている

# 比較結果

1回の固有値問題で解くスペクトラルクラスタリングをspc2002と呼ぶ事にする

- エントロピー

| | tr23 | tr31 | tr41 | tr45 |
|---|---|---|---|---|
| Mcut | 0.4895708 | 0.2945980 | 0.3375775 | 0.4509916 |
| spc2002 | 0.4455226 | 0.2858442 | 0.2467648 | 0.244881 |
| k-means | 0.4859781 | 0.4081418 | 0.2286196 | 0.313689 |

- 純度

| | tr23 | tr31 | tr41 | tr45 |
|---|---|---|---|---|
| Mcut | 0.6666667 | 0.8036677 | 0.6879271 | 0.642029 |
| spc2002 | 0.6764706 | 0.7820928 | 0.8075171 | 0.8289855 |
| k-means | 0.6470588 | 0.6623517 | 0.8405467 | 0.7536232 |

# 比較実験からわかったこと

- spc2002が他の手法より良い結果が得られる
- 1回の固有値問題だけで済むのでMcutよりspc2002やk-meansのほうが実行時間が短い
- Mcutは比較的大きいデータのほうが良好な結果が得られる傾向にある
- Mcutの結果は不変だがspc2002とk-meansは実行毎に結果が多少変わる