潜在変数を用いたカーネルの確率変数化による類似度からのクラスタリング

竹岡 邦紘　岡留 剛（関西学院大）

概要

目的
データ間の類似度のみが与えられたときの，
クラスタリング手法の提案

アプローチ
類似度行列に対する確率的生成モデルの構築

ポイント
「カーネルで表現された」類似度行列
「事後確率最大」でクラス数と所属クラスを決定

確率的生成モデル

類似度行列の要素は以下の条件を満たす.

$0 \le k_{ij} \le k_{ii}$ かつ　$0 \le k_{ij} \le k_{jj}\ (i,j=1,\ldots,N)$

個体が持つM次元の潜在特徴ベクトル $\mathbf{z}_i$ の
内積 $k_{ij} = \mathbf{z}_i^{\mathrm{T}} \mathbf{z}_j$ で類似度行列Kを生成

$$\mathbf{z}_i = (z_{il})_i, z_{il} \in \{0,1\}, z_{il} \sim Bern(q_c)$$

尤度
この生成モデルにおける尤度関数は

$$p(\mathbf{K}|\mathbf{S},\boldsymbol{q},M) = \prod_{i=1}^{N} {}_M C_{k_{ii}} \prod_{c=1}^{C} \{\pi_c q_c^{k_{ii}} (1-q_c)^{M-k_{ii}}\}^{s_{ic}} \times \prod_{i=1}^{N}\prod_{j=1}^{N} p(k_{ij}|k_{11},\ldots,k_{NN})$$

Kの各要素が非独立のため最適化が難しい.

対角要素が与えられたとすると(k_{ii}: given)，
任意の第i行について他の要素が条件付き独立

任意の第m行について
対角要素が与えられたときの尤度は

$$p(\mathbf{k}_{-m} \mid k_{mm}, \mathbf{S}, \mathbf{q}) = \prod_{i=1}^{N} {}_{k_{mm}} C_{k_{mi}} \prod_{c=1}^{C} \left\{ q_c^{k_{mi}} (1-q_c)^{k_{mm}-k_{mi}} \right\}^{S_{ic}}$$

事後確率

$$p(\mathbf{S},\mathbf{q},\boldsymbol{\pi} \mid \mathbf{k}_m) \propto p(\mathbf{k}_{-m} \mid k_{mm}, \mathbf{S}, \mathbf{q}, \boldsymbol{\pi}) p(\mathbf{S} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\mathbf{q})$$

$$= \prod_{i=1}^{N} {}_{k_{mm}} C_{k_{mi}} \prod_{c=1}^{C} \left\{ \pi_c q_c^{k_{mi}} (1-q_c)^{k_{mm}-k_{mi}} \right\}^{S_{ic}} \times \prod_{c=1}^{C} q_c^{\alpha_c - 1} (1-q_c)^{\beta_c - 1} \pi_c^{\gamma_c - 1}$$

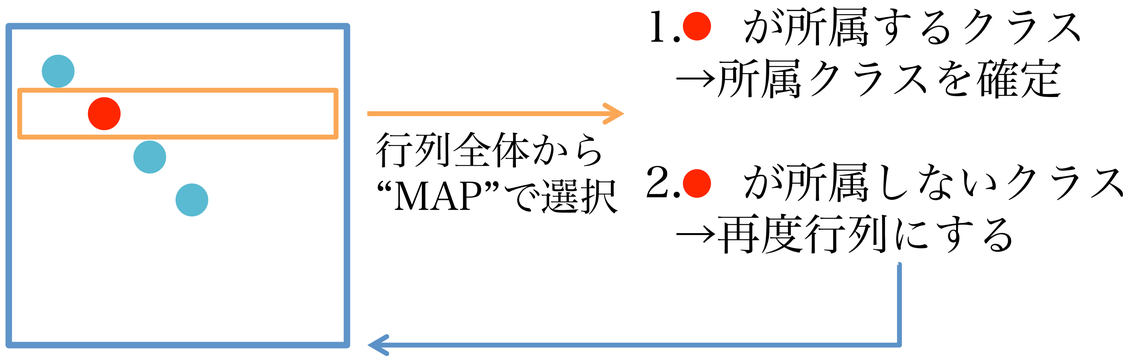
$$p(\mathbf{S} \mid \mathbf{k}_m) \propto \frac{\prod_{c=1}^{C} \Gamma\left(\alpha_c + \sum_{i=1}^{N} k_{mi} s_{ic}\right) \Gamma\left(\beta_c + \sum_{i=1}^{N} (k_{mm} - k_{mi}) s_{ic}\right) \Gamma\left(\gamma_c + \sum_{i=1}^{N} s_{ic}\right)}{\Gamma\left(N + \sum_{c=1}^{C} \gamma_c\right) \prod_{c=1}^{C} \Gamma\left(\alpha_c + \beta_c + \sum_{i=1}^{N} k_{mm} s_{ic}\right)}$$

$\mathbf{S}$：1-of-K符号化法によるデータの所属クラス行列
$\mathbf{q}$：潜在特徴ベクトルを生成するベルヌーイ分布のパラメータ
$\boldsymbol{\pi}$：$\mathbf{S}$ の事前確率のパラメータ

提案手法の流れ

各行ごとに，事後確率最大となる各データの
所属クラスを決め，
行列中で最大の事後確率のクラス割り当てを採用
そのクラス割り当てで「分割」を繰り返す

提案手法の概略図

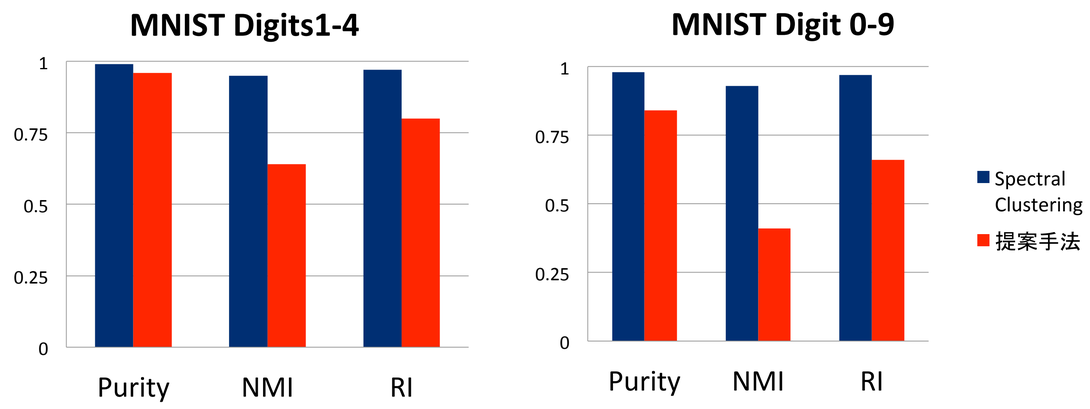


「事後確率最大化」によって，
所属クラスが確定していないデータがなくなると終了.

性能向上のため，K近傍法で後処理をする.
結果，クラス数と各個体の所属クラスが決定される.

実験と考察

条件
手書き数字データセットMNIST
（各数字100個を一つのクラスとして扱う）
類似度関数： 線形カーネル



SC+CRPと提案手法の比較

提案手法の特徴
- 提案手法はクラス数をよりよく推定できる
- スコアはSC+CRPに劣る
- 超パラメータを変えずにSC+CRPに近い性能を出せる

今後の課題
- K近傍法のKに対する依存性
- 他のデータ（文書など）に対する実験が必須
- 実行に時間がかかる