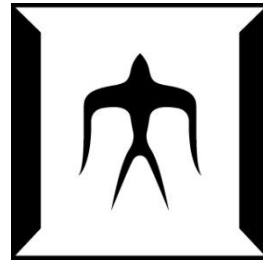


確率的パターン認識と クラスタリングの新手法



東京工業大学 計算工学専攻

杉山 将

本講演では、講演者のグループで最近開発した確率的パターン認識とクラスタリングの新手法を紹介する。

確率的パターン認識は、パターン x がクラス y に属する確率 $p(y|x)$ を推定し、この確率を最大するクラス y にパターン x を割り当てるパターン認識法である。確率的パターン認識の代表的な手法は(カーネル)ロジスティック回帰であるが、非線形最適化問題を準ニュートン法などで解く必要があるため、大規模なデータに適用するのが困難であった。そこで我々は、最適解が解析的に求められる手法「最小二乗確率的分類器(LSPC)」を提案した[1]。LSPCは、カーネルロジスティック回帰と同等の認識精度を維持しつつ、数百～千倍高速に学習できることを実験的に示す。

クラスタリングの目的は、似たパターンに同じクラスラベルを、似ていないパターンに異なるクラスラベルを割り当てることである。これまでに、様々なクラスタリング手法が提案されてきたが、アルゴリズムの適切な初期化が困難であったり、カーネル幅や近傍数などのパラメータの決定が困難であるという問題があった。そこで我々は、初期値に依存せず解析的に解を求められるクラスタリング法「二乗損失相互情報量クラスタリング(SMIC)」を提案する共に、カーネル幅や近傍数などのパラメータを客観的に決定するための規準「最小二乗相互情報量(LSMI)」を提案した[2]。提案手法の有効性を計算機実験により示す。

[1] Sugiyama, M.

Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting.
IEICE Transactions on Information and Systems, vol.E93-D, no.10, pp.2690-2701, 2010.
<http://sugiyama-www.cs.titech.ac.jp/~sugi/2010/LSPC.pdf>

[2] Sugiyama, M., Yamada, M., Kimura, M., & Hachiya, H.

On information-maximization clustering: tuning parameter selection and analytic solution.
International Conference on Machine Learning (ICML2011), pp.65-72, 2011.
<http://sugiyama-www.cs.titech.ac.jp/~sugi/2011/ICML2011a.pdf>

本発表の構成

1. 確率的パターン認識

- A) 条件付き確率推定
- B) 提案手法
- C) 高速化の理由
- D) 実験

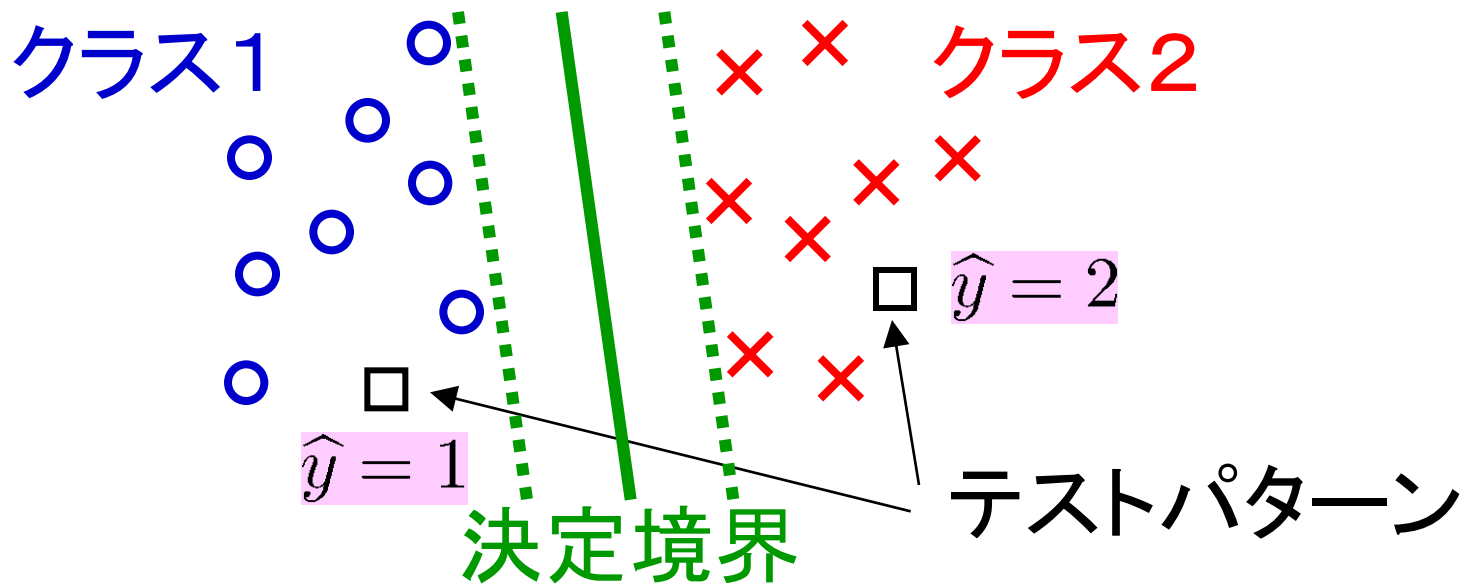
2. クラスタリング



決定的パターン認識の標準手法: 4

サポートベクトルマシン

- 訓練標本 $\{(x_i, y_i)\}_{i=1}^n$ から, マージンを最大にする**決定境界**を求める



- テストパターン x の属するクラス y を予測

確率的パターン認識

- 応用場面によっては、予測したクラスの**信頼度**を知りたいことがある
- テストパターン x が属するクラス y を予測するだけでなく、その**確率** $p(y|x)$ も求める
- 確率が最大のクラスにテストパターンを分類

$$\hat{y} = \underset{y}{\operatorname{argmin}} p(y|x)$$

- 確率が低いときは「**棄却**」という選択肢を選ぶこともできる（応用上は重要！）

確率的パターン認識の標準手法: ⁶ カーネルロジスティック回帰 (KLR)

- 多クラスカーネルロジスティックモデル:

$$q(y|\mathbf{x}; \boldsymbol{\theta}) \propto \exp \left(\sum_{i=1}^n \theta_i^{(y)} K(\mathbf{x}, \mathbf{x}_i) \right)$$

$$K(\mathbf{x}, \mathbf{x}') = \exp \left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2} \right)$$

- (罰則付き) 最尤推定により学習:

$$\min_{\boldsymbol{\theta}} \left[-\sum_{i=1}^n \log q(y_i|\mathbf{x}_i; \boldsymbol{\theta}) + \frac{\lambda}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} \right]$$

学習アルゴリズム

7

	サポートベクトルマシン (決定的分類)	ロジスティック回帰 (確率的分類)
非線形	高速	低速
線形で 入力が疎	超高速	超高速

- 超高速に学習できる非線形確率的分類器を紹介する

本発表の構成

1. 確率的パターン認識

- A) 条件付き確率推定
- B) 提案手法
- C) 高速化の理由
- D) 実験

2. クラスタリング



定式化

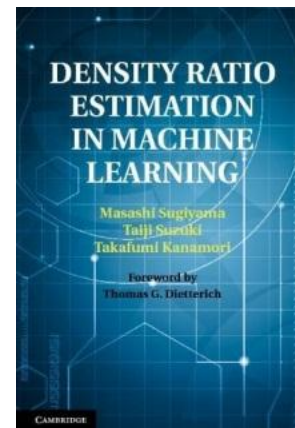
- パターン: $\boldsymbol{x} \in \mathbb{R}^d$
- クラス: $y \in \{1, \dots, c\}$
- 訓練標本: パターンとクラスの組

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} p(\boldsymbol{x}, y)$$

- 方針: 条件付き確率を密度比の形で推定する

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})}$$

Sugiyama, Suzuki & Kanamori,
Density Ratio Estimation
in Machine Learning,
Cambridge University Press, 2012



条件付き確率の カーネル最小二乗推定

■ モデル:
$$q(y|\mathbf{x}) = \sum_{i=1}^n \alpha_i^{(y)} K(\mathbf{x}, \mathbf{x}_i)$$

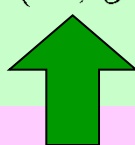
$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

■ クラス y のパラメータの学習規準:
$$\min_{\boldsymbol{\alpha}} J_y(\boldsymbol{\alpha})$$

$$J_y(\boldsymbol{\alpha}) = \frac{1}{2} \int \left(q(y|\mathbf{x}) - p(y|\mathbf{x}) \right)^2 p(\mathbf{x}) d\mathbf{x}$$

二乗損失の分解

$$\begin{aligned}
 J_y(\boldsymbol{\alpha}) &= \frac{1}{2} \int \left(q(y|\boldsymbol{x}) - p(y|\boldsymbol{x}) \right)^2 p(\boldsymbol{x}) d\boldsymbol{x} \\
 &= \frac{1}{2} \int \left(q(y|\boldsymbol{x}) \right)^2 p(\boldsymbol{x}) d\boldsymbol{x} - \int q(y|\boldsymbol{x}) p(\boldsymbol{x}, y) d\boldsymbol{x} \\
 &\quad + \frac{1}{2} \int \left(p(y|\boldsymbol{x}) \right)^2 p(\boldsymbol{x}) d\boldsymbol{x}
 \end{aligned}$$

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})}$$


- **1項目** : $(q(y|\boldsymbol{x}))^2$ の $p(\boldsymbol{x})$ に関する期待値
- **2項目** : $q(y|\boldsymbol{x})$ の $p(\boldsymbol{x}, y)$ に関する期待値
- **3項目** : 定数なので無視
- 期待値は標本平均で近似できる

 **密度推定不要**

最小二乗確率的分類器(LSPC) ¹²

Sugiyama (IEICE-ED 2010)

- 期待値を標本平均で近似し, ℓ_2 -正則化項を加える:

$$\hat{\alpha}^{(y)} = \operatorname{argmin}_{\alpha} \left[\frac{1}{n} \alpha^\top K \alpha - \frac{2}{n} \alpha^\top K \mathbf{1}^{(y)} + \frac{\lambda}{n} \alpha^\top \alpha \right]$$

$$K_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j) \quad \mathbf{1}_i^{(y)} = \begin{cases} 1 & \text{if } y_i = y \\ 0 & \text{if } y_i \neq y \end{cases}$$

- 解は**解析的**に計算できる:

$$\hat{\alpha}^{(y)} = (K^2 + \lambda I)^{-1} K \mathbf{1}^{(y)}$$

- $\mathbf{1}^{(y)}$ を出力とする**カーネルリッジ学習**と等価

本発表の構成



1. 確率的パターン認識

- A) 条件付き確率推定
- B) 提案手法
- C) 高速化の理由
- D) 実験

2. クラスタリング

$$q(y|\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left(\sum_{i=1}^n \theta_i^{(y)} K(\mathbf{x}, \mathbf{x}_i) \right)$$

- **正規化定数** $Z(\boldsymbol{\theta})$ の計算が煩雑:

$$Z(\boldsymbol{\theta}) = \sum_{y=1}^c \exp \left(\sum_{i=1}^n \theta_i^{(y)} K(\mathbf{x}, \mathbf{x}_i) \right) \quad \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}^{(1)} \\ \vdots \\ \boldsymbol{\theta}^{(c)} \end{pmatrix}$$

- $Z(\boldsymbol{\theta})$ は**全クラス**のパラメータを含む: $\dim \boldsymbol{\theta} = nc$
- 正規化なしでは尤度が無限大に**発散**:

$$\min_{\boldsymbol{\theta}} \left[- \sum_{i=1}^n \log q(y_i | \mathbf{x}_i; \boldsymbol{\theta}) + \frac{\lambda}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} \right]$$

- 最適化を高速化するための良い「**構造**」がない

なぜLSPCは速いか？

- LSPCは正規化定数を含まない:

$$\hat{\alpha}^{(y)} = \operatorname{argmin}_{\alpha} \left[\frac{1}{n} \alpha^{\top} K \alpha - \frac{2}{n} \alpha^{\top} K \mathbf{1}^{(y)} + \frac{\lambda}{n} \alpha^{\top} \alpha \right]$$

$$K_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j) \quad \mathbf{1}_i^{(y)} = \begin{cases} 1 & \text{if } y_i = y \\ 0 & \text{if } y_i \neq y \end{cases}$$

- LSPCは正規化なしでも**一**致性を持つ
(実用上は事後に正規化をしても良い)
 - 各クラス**別々**に学習できる: $\dim \alpha = n$
- 単純なカーネル最小二乗の定式化のおかげで、
解が解析的に求められる！

$$\hat{\alpha}^{(y)} = (K^2 + \lambda I)^{-1} K \mathbf{1}^{(y)}$$

$$\hat{\alpha}^{(y)} = \underset{\alpha}{\operatorname{argmin}} \left[\frac{1}{n} \alpha^\top K \alpha - \frac{2}{n} \alpha^\top K \mathbf{1}^{(y)} + \frac{\lambda}{n} \alpha^\top \alpha \right]$$

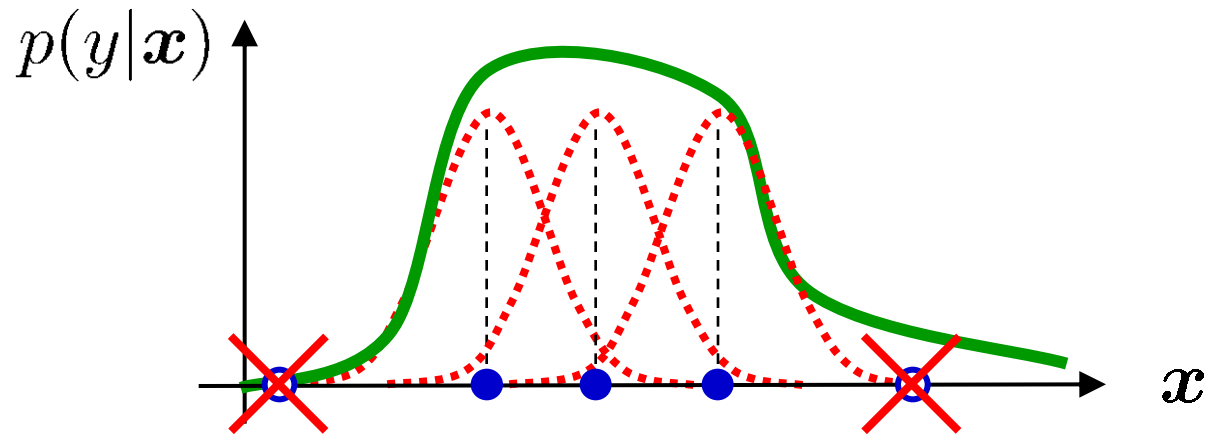
- 学習したパラメータ $\hat{\alpha}^{(y)}$ は負の値を取る可能性がある。
 - 非負拘束を加えても良いが、解を解析的に求められなくなる(二次計画問題を数値的に解く必要がある)

$$\hat{\alpha}^{(y)} = (K^2 + \lambda I)^{-1} K \mathbf{1}^{(y)}$$

- 単純に負の出力を0に切り上げることにする:

$$\hat{p}(y|\mathbf{x}) \propto \max \left(0, \sum_{\ell=1}^n \hat{\alpha}_\ell^{(y)} K(\mathbf{x}, \mathbf{x}_\ell) \right)$$

- カーネルを対象クラスの標本のみ配置する:
 - 対象クラスの標本がある場所は事後確率が大きい
 - それ以外では事後確率はゼロに近い



$$q(y|\mathbf{x}; \boldsymbol{\alpha}^{(y)}) = \sum_{\ell=1}^{n_y} \alpha_{\ell}^{(y)} \exp \left(-\frac{\|\mathbf{x} - \mathbf{x}_{\ell}^{(y)}\|^2}{2\sigma^2} \right)$$

$$\dim \boldsymbol{\alpha}^{(y)} = n_y$$

n_y : クラス y の標本数

本発表の構成



1. 確率的パターン認識

- A) 条件付き確率推定
- B) 提案手法
- C) 高速化の理由
- D) 実験

2. クラスタリング

実験結果

19

Dataset	LSPC	KLR
Aeroplane	82.6 (1.0)	83.0 (1.3)
Bicycle	77.7 (1.7)	76.6(3.4)
Bird	68.7(2.0)	70.8 (2.2)
Boat	74.4 (2.0)	72.8(2.6)
Bottle	65.4 (1.8)	62.1(4.3)
Bus	85.4 (1.4)	85.6 (1.4)
Car	73.0 (0.8)	72.1(1.2)
Cat	73.6 (1.4)	74.1 (1.7)
Chair	71.0 (1.0)	70.5 (1.0)
Cow	71.7 (3.2)	69.3(3.6)
Diningtable	75.0 (1.6)	71.4(2.7)
Dog	69.6 (1.0)	69.4 (1.8)
Horse	64.4 (2.5)	61.2(3.2)
Motorbike	77.0 (1.7)	75.9(3.3)
Person	67.6 (0.9)	67.0(0.8)
Pottedplant	66.2 (2.6)	61.9(3.2)
Sheep	77.8 (1.6)	74.0(3.8)
Sofa	67.4 (2.7)	65.4(4.6)
Train	79.2 (1.3)	78.4 (3.0)
Tvmonitor	76.7 (2.2)	76.6 (2.3)
Average	73.2	71.9
Train time [sec]	0.7	24.6

画像からの一般物体認識
Pascal VOC
(AUC)

オーディオタグ付け
Freesound (AUC)

	LSPC	KLR
AUC	70.1 (9.6)	66.7 (10.3)
Train time [sec]	0.005	0.612

■ LSPCは速くて高精度！



■ KLR: 対数線形モデルの最尤推定

- 正規化項の計算が煩雑
- 最適化を高速化するための良い「構造」がない

■ LSPC: 線形モデルの最小二乗推定

- 正規化が不要⇒クラスごとに学習できる
- 解が解析的に計算できる

<http://sugiyama-www.cs.titech.ac.jp/~sugi/software/LSPC/>

■ 関連研究:

- マルチタスク, マルチラベル, 転移学習への拡張
- 条件付き確率密度推定
- 顔画像からの年齢認証への応用



1. 確率的パターン認識

2. クラスタリング

A) 従来法の紹介

B) 提案法

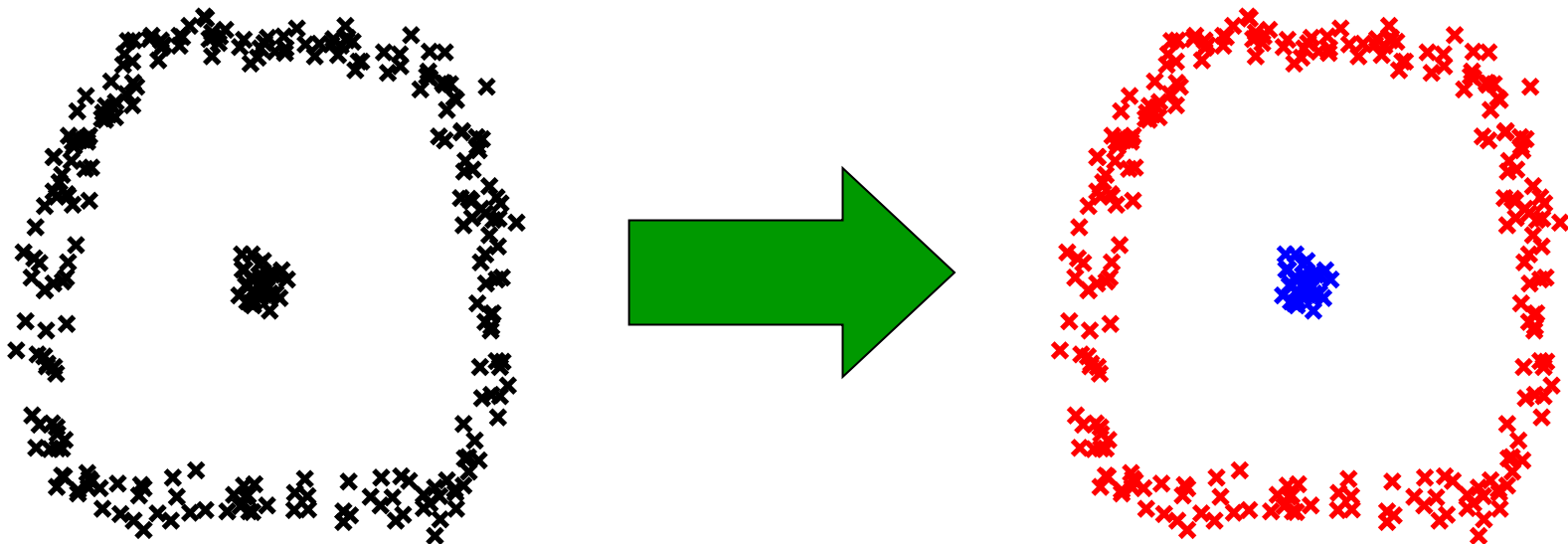
i. クラスタリング

ii. チューニングパラメータの自動選択

C) 実験

クラスタリング

- ラベルなし標本 $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$ に
クラスタラベル $\{y_i \in \{1, \dots, c\}\}_{i=1}^n$ を付与する:
 - 似た標本は同じクラスタに割り当てる
 - 似ていない標本は違うクラスタに割り当てる
- 本研究では, クラスタ数 c は既知と仮定する



本発表の構成



1. 確率的パターン認識

2. クラスタリング

A) 従来法の紹介

B) 提案法

i. クラスタリング

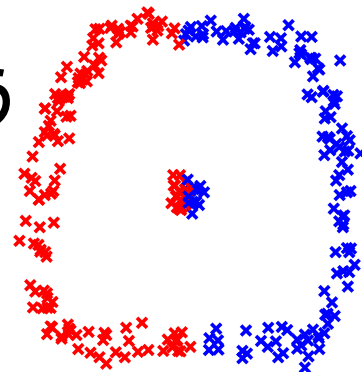
ii. チューニングパラメータの自動選択

C) 実験

- **混合モデル**を最尤推定やベイズ推定で学習

$$p(\boldsymbol{x}) = \sum_{y=1}^c p(\boldsymbol{x}|y)p(y)$$

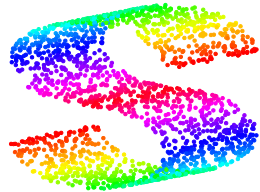
- **K平均法** (MacQueen, 1967)
 - **ディリクレ過程混合法** (Ferguson, 1973)
- **利点と欠点:**
 - ☺ チューニングパラメータがない
 - ☹ クラスタの形状があらかじめ定めたモデル(≒ガウシアン)で決まってしまう
 - ☹ アルゴリズムの初期化が困難



■ クラスターの形状に関して仮定を置かない:

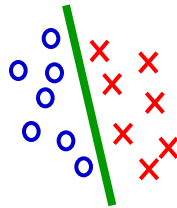
- **スペクトルクラスタリング**: 非線形次元削減を行ったあとにK平均法を適用する

(Shi & Malik, 2000; Ng *et al.*, 2002)



- **識別的クラスタリング**: 識別器とクラスラベルを同時に学習する

(Xu *et al.*, 2005; Bach & Harchaoui, 2008)



- **従属性最大化クラスタリング**: 標本との従属性が最大になるようにクラスラベルを決定する

(Song *et al.*, 2007; Faivishevsky & Goldberger, 2010)

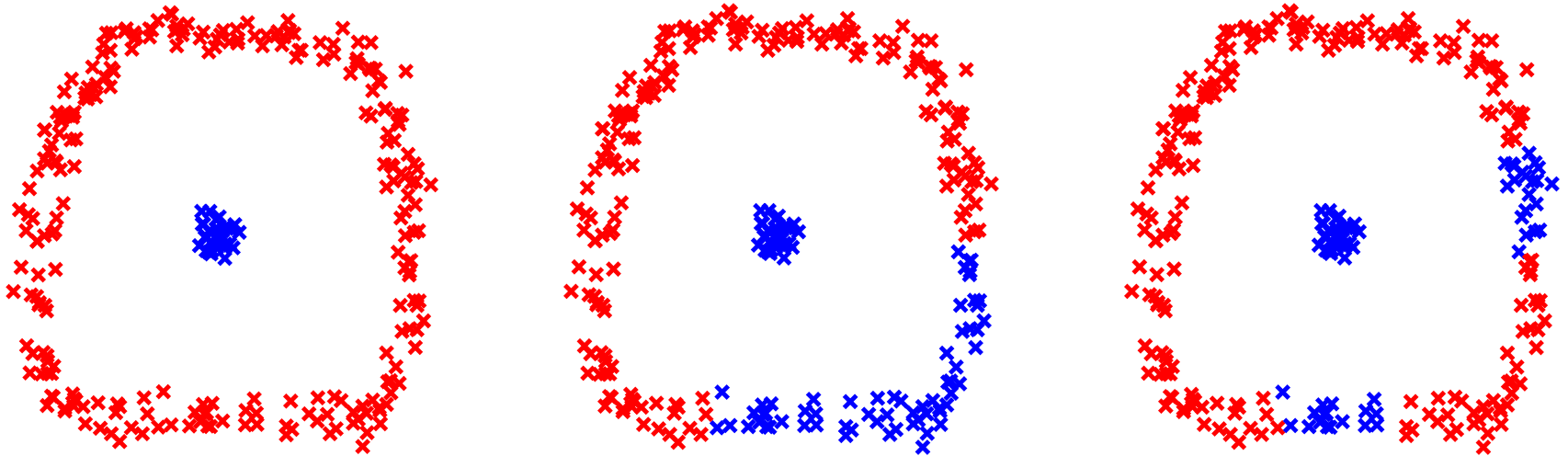
- **情報量最大化クラスタリング**: 情報量が最大になるように識別器を学習

(Agakov & Barberu, 2006; Gomes *et al.*, 2010)

モデルなしクラスタリング

■ 利点と欠点:

- 😊 クラスタの形状が柔軟
- 😞 カーネル・類似度のパラメータの決定が困難
- 😞 アルゴリズムの初期化が困難





1. 確率的パターン認識

2. クラスタリング

A) 従来法の紹介

B) 提案法

i. クラスタリング

ii. チューニングパラメータの自動選択

C) 実験

本研究の目的

■ 新たな情報量最大化クラスタリング法を提案する:

- ☺ 最適解が解析的に求められる
- ☺ チューニングパラメータを客観的に決定できる

■ 提案法の流れ:

- 情報量を最大にするようにノンパラメトリック・カーネル分類器を学習する
- 情報量を最大にするようにチューニングパラメータを決定する

- 情報量の尺度として, SMIを用いる:

$$\text{SMI} = \frac{1}{2} \int \sum_{y=1}^c p(\mathbf{x})p(y) \left(\frac{p(\mathbf{x}, y)}{p(\mathbf{x})p(y)} - 1 \right)^2 d\mathbf{x}$$

- 通常の相互情報量はカルバック・ライブラー (KL) ダイバージェンス
- SMIはピアソン (PE) ダイバージェンス
- KLもPEも, f-ダイバージェンスというクラスに属する (即ち, 似た性質を有する)
- 実際, 通常の相互情報量と同様に, SMIは

$$\text{SMI} = 0 \iff \mathbf{x} \perp\!\!\!\perp y$$



本発表の構成

1. 確率的パターン認識

2. クラスタリング

A) 従来法の紹介

B) 提案法

i. クラスタリング

ii. チューニングパラメータの自動選択

C) 実験

■ カーネル確率的分類器:

$$p(y|\mathbf{x}; \alpha) = \sum_{i=1}^n \alpha_{y,i} K(\mathbf{x}, \mathbf{x}_i)$$

■ SMIが最大になるように分類器を学習する:

$$\text{SMI} = \frac{1}{2} \int \sum_{y=1}^c p(\mathbf{x}) p(y) \left(\frac{p(\mathbf{x}, y)}{p(\mathbf{x}) p(y)} - 1 \right)^2 d\mathbf{x}$$

■ ただし、ラベルなしデータ $\{\mathbf{x}_i\}_{i=1}^n$ しか与えられない!

SMIの近似

- クラスタ事後確率をカーネルモデルで近似:

$$p(y|\mathbf{x}) \approx \sum_{i=1}^n \alpha_{y,i} K(\mathbf{x}, \mathbf{x}_i)$$

- 期待値を標本平均で近似:

$$\int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \approx \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i) \quad \{\mathbf{x}_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} p(\mathbf{x})$$

- クラスタ事前確率は一様と仮定:

$$p(y) = 1/c$$

c : クラスタ数

- これらにより, 次のSMIの推定量が得られる:

$$\widehat{\text{SMI}} = \frac{c}{2n} \sum_{y=1}^c \boldsymbol{\alpha}_y^\top \mathbf{K}^2 \boldsymbol{\alpha}_y - \frac{1}{2}$$

$$\boldsymbol{\alpha}_y = (\alpha_{y,1}, \dots, \alpha_{y,n})^\top$$

$$K_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$$

SMI推定量の最大化

$$\max_{\{\alpha_y\}_{y=1}^c} \widehat{\text{SMI}}$$

$$\widehat{\text{SMI}} = \frac{c}{2n} \sum_{y=1}^c \alpha_y^\top K^2 \alpha_y - \frac{1}{2}$$

- $\{\alpha_y\}_{y=1}^c$ が正規直交だという仮定のもと、大域的最適解はカーネル行列 K の主成分によって与えられる
 - Cf. Ding & He (ICML2004)

SMIに基づくクラスタリング (SMIC) ³⁴

■ 事後処理:

- 主成分 $\{\phi_y\}_{y=1}^c$ の符号を定める:

$$\tilde{\phi}_y = \phi_y \times \text{sign}(\phi_y^\top \mathbf{1}) \quad \mathbf{1} : \text{要素が全て1のベクトル}$$

- $p(y) = 1/c$ に基づいて正規化
- 負の出力をゼロに切り上げる

■ 最終的な解 (解析的に計算可能):

$$y_i = \underset{y}{\operatorname{argmax}} \frac{[\max(\mathbf{0}, \tilde{\phi}_y)]_i}{\max(\mathbf{0}, \tilde{\phi}_y)^\top \mathbf{1}}$$

$[\cdot]_i$: ベクトルの第 i 要素

$\mathbf{0}$: 要素が全て0のベクトル

本発表の構成

1. 確率的パターン認識

2. クラスタリング

A) 従来法の紹介

B) 提案法

i. クラスタリング

ii. チューニングパラメータの自動選択

C) 実験



チューニングパラメータの決定 36

- SMICの解はカーネル関数に依存する
- SMIを最大にするようにカーネルを決める
- クラスタリングに用いた $\widehat{\text{SMI}}$ を使えば良い？

$$\widehat{\text{SMI}} = \frac{c}{2n} \sum_{y=1}^c \alpha_y^\top \mathbf{K}^2 \alpha_y - \frac{1}{2}$$

- しかし、 $\widehat{\text{SMI}}$ はSMIの教師なし推定量のため、精度が良くない
- チューニングパラメータの決定の段階ではクラスタレベルの推定値が得られているため、**SMIの教師付き推定が可能！**

$$\text{SMI} = \frac{1}{2} \int \sum_{y=1}^c p(\mathbf{x})p(y) \left(\frac{p(\mathbf{x}, y)}{p(\mathbf{x})p(y)} - 1 \right)^2 d\mathbf{x}$$

■ 最小二乗相互情報量 (LSMI):

- 密度比を直接推定する

Suzuki & Sugiyama
(AISTATS2010)

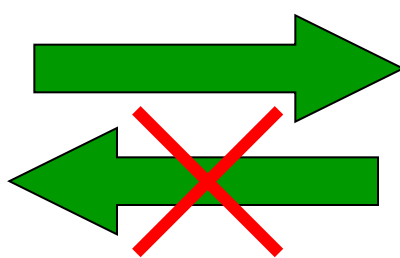
$$r(\mathbf{x}, y) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})p(y)}$$

※各確率密度は推定しない

- 密度比推定は密度推定よりも優しい
(Vapnikの原理)

各密度がわかる

$$p(\mathbf{x}, y), p(\mathbf{x}), p(y)$$



密度比がわかる

$$r(\mathbf{x}, y) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})p(y)}$$

密度比の直接推定

■ カーネル密度比モデル:

$$\hat{r}(\boldsymbol{x}, y) = \sum_{\ell: y_\ell = y} \theta_\ell L(\boldsymbol{x}, \boldsymbol{x}_\ell)$$

$L(\boldsymbol{x}, \boldsymbol{x}')$: カーネル関数

(実験ではガウスカーネルを用いる)

■ 最小二乗推定:

$$\frac{1}{2} \int \sum_{y=1}^c \left(\hat{r}(\boldsymbol{x}, y) - r(\boldsymbol{x}, y) \right)^2 p(\boldsymbol{x}) p(y) d\boldsymbol{x}$$

$$r(\boldsymbol{x}, y) = \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})p(y)}$$

密度比の直接推定

- 期待値を標本平均で近似し，正則化する：

$$\min_{\boldsymbol{\theta}_y} \left[\frac{1}{2} \boldsymbol{\theta}_y^\top \widehat{\mathbf{H}}^{(y)} \boldsymbol{\theta}_y - \boldsymbol{\theta}_y^\top \widehat{\mathbf{h}}^{(y)} + \frac{\delta}{2} \boldsymbol{\theta}_y^\top \boldsymbol{\theta}_y \right]$$

$$\widehat{H}_{\ell, \ell'}^{(y)} = \frac{n_y}{n^2} \sum_{i=1}^n L(\mathbf{x}_i, \mathbf{x}_\ell^{(y)}) L(\mathbf{x}_i, \mathbf{x}_{\ell'}^{(y)})$$

$$\widehat{h}_\ell^{(y)} = \frac{1}{n} \sum_{i: y_i=y} L(\mathbf{x}_i, \mathbf{x}_\ell^{(y)})$$

- 大域的最適解が解析的に計算できる：

$$\widehat{\boldsymbol{\theta}}^{(y)} = (\widehat{\mathbf{H}}^{(y)} + \delta \mathbf{I})^{-1} \widehat{\mathbf{h}}^{(y)}$$

$$\widehat{r}(\mathbf{x}, y) = \sum_{\ell=1}^{n_y} \widehat{\theta}_\ell^{(y)} L(\mathbf{x}, \mathbf{x}_\ell^{(y)})$$

- カーネル関数と正則化パラメータはクロスバリデーションで決定できる

- SMIの推定量が**解析的**に計算できる:

$$\text{LSMI} = -\frac{1}{2n^2} \sum_{i,j=1}^n \hat{r}(\mathbf{x}_i, y_j)^2 + \frac{1}{n} \sum_{i=1}^n \hat{r}(\mathbf{x}_i, y_i) - \frac{1}{2}$$

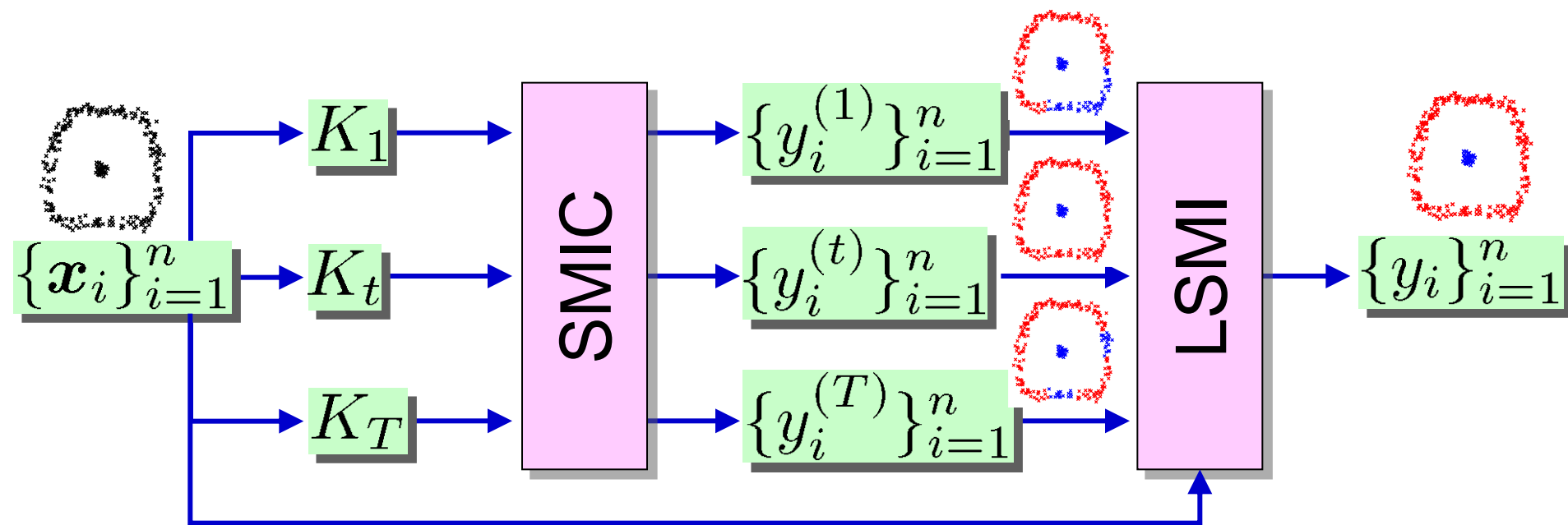
- LSMIは優れた理論的収束特性を有する！

Suzuki & Sugiyama (AISTATS2010)

- LSMIを最大にするように, SMICのカーネル関数を決定する

■ SMIクラスタリング + LSMIによるモデル選択:

- 入力: ラベルなし標本 $\{\mathbf{x}_i\}_{i=1}^n$
カーネルの候補 $\{K_t(\mathbf{x}, \mathbf{x}')\}_{t=1}^T$
- 出力: クラスタラベル $\{y_i\}_{i=1}^n$





本発表の構成

1. 確率的パターン認識

2. クラスタリング

A) 従来法の紹介

B) 提案法

i. クラスタリング

ii. チューニングパラメータの自動選択

C) 実験

■ 局所スケールカーネルのスパース版を用いる

(Zelnik-Manor & Perona, NIPS2004)

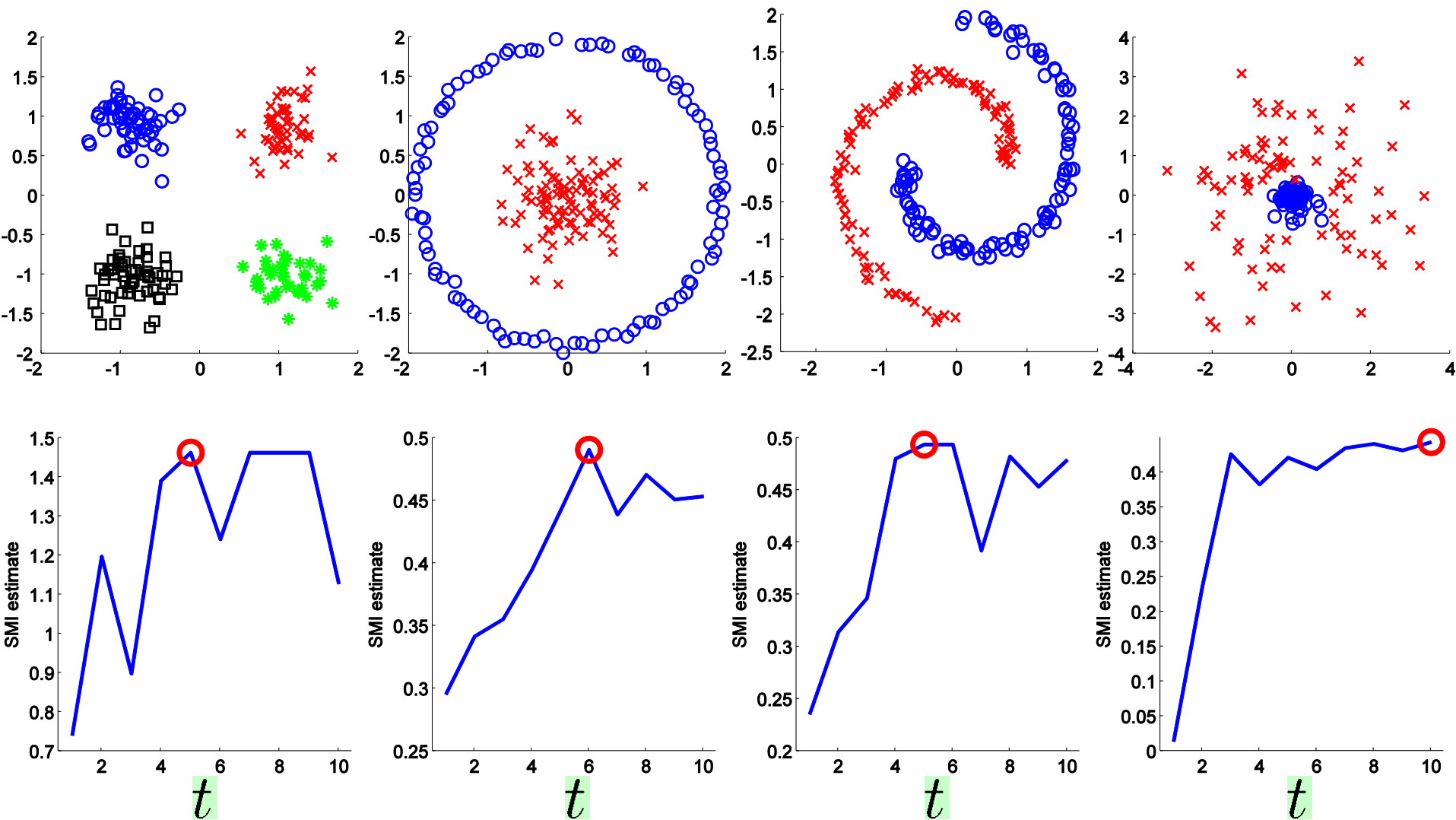
$$K(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_i\sigma_j}\right) & \text{if } \mathbf{x}_i, \mathbf{x}_j \text{ are } t \text{ neighbors} \\ 0 & \text{otherwise} \end{cases}$$

$$\sigma_i = \|\mathbf{x}_i - \mathbf{x}_i^{(t)}\|$$

$\mathbf{x}_i^{(t)}$: t -th neighbor of \mathbf{x}_i

- チューニングパラメータ t は, LSMIを最大にするように決定する

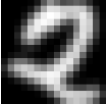
SMICの実行例



■ 提案法により自然なクラスタリング結果が得られた


- **KM**: K平均法 (MacQueen, 1967)
- **SC**: 自動調整スペクトルクラスタリング (Zelnik-Manor & Perona, NIPS2004)
- **MNN**: 平均最近傍近似に基づく従属性最大化クラスタリング (Faivishevsky & Goldberger, ICML2010)
- **MIC**: ロジスティックモデルを用いた情報量最大化クラスタリング + 最尤相互情報量に基づくモデル選択 (Gomes, Krause & Perona, NIPS2010)
(Suzuki, Sugiyama, Sese & Kanamori, FSDM2008)

実験結果




Digit ($d = 256, n = 5000$, and $c = 10$)

	KM	SC	MNN	MIC	SMIC
ARI	0.42(0.01)	0.24(0.02)	0.44(0.03)	0.63(0.08)	0.63(0.05)
Time	835.9	973.3	318.5	84.4[3631.7]	14.4[359.5]



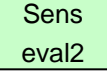
Face ($d = 4096, n = 100$, and $c = 10$)

	KM	SC	MNN	MIC	SMIC
ARI	0.60(0.11)	0.62(0.11)	0.47(0.10)	0.64(0.12)	0.65(0.11)
Time	93.3	2.1	1.0	1.4[30.8]	0.0[19.3]



Document ($d = 50, n = 700$, and $c = 7$)

	KM	SC	MNN	MIC	SMIC
ARI	0.00(0.00)	0.09(0.02)	0.09(0.02)	0.01(0.02)	0.19(0.03)
Time	77.8	9.7	6.4	3.4[530.5]	0.3[115.3]




Word ($d = 50, n = 300$, and $c = 3$)

	KM	SC	MNN	MIC	SMIC
ARI	0.04(0.05)	0.02(0.01)	0.02(0.02)	0.04(0.04)	0.08(0.05)
Time	6.5	5.9	2.2	1.0[369.6]	0.2[203.9]

Accelerometry ($d = 5, n = 300$, and $c = 3$)

	KM	SC	MNN	MIC	SMIC
ARI	0.49(0.04)	0.58(0.14)	0.71(0.05)	0.57(0.23)	0.68(0.12)
Time	0.4	3.3	1.9	0.8[410.6]	0.2[92.6]



Speech ($d = 50, n = 400$, and $c = 2$)

	KM	SC	MNN	MIC	SMIC
ARI	0.00(0.00)	0.00(0.00)	0.04(0.15)	0.18(0.16)	0.21(0.25)
Time	0.9	4.2	1.8	0.7[413.4]	0.3[179.7]

- Adjusted Rand index (ARI):
大きい方が良い
- 赤: 1%のt検定で最適か最適タイ
- SMICは精度が良く、計算速度も速い!

クラスタリングのまとめ



■ 従来のクラスタリング法の欠点:

- アルゴリズムの初期化が困難
- チューニングパラメータの初期化が困難

■ SMIC: 二乗損失相互情報量 (SMI) に基づく 新たな情報量最大化クラスタリング法

- 大域的最適解が解析的に求められる
- チューニングパラメータを客観的に決定できる

<http://sugiyama-www.cs.titech.ac.jp/~sugi/software/SMIC/>