機械学習によるデータの自動クラスタリング

杉山 将
〒 152-8552 東京都目黒区大岡山 2-12-1
東京工業大学 大学院情報理工学研究科 計算工学専攻
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi/

概 要

インターネットの普及や計算機性能の飛躍的な向上と相まって，膨大な量のデータから有用な知識を自動的に発見するための**機械学習**技術が近年注目されている．機械学習の標準的なアプローチは，データを生成するモデルの学習を通してデータ生成過程をシミュレーションし，所望のデータ解析を行うという方法である．このような生成モデル推定のアプローチは非常に汎用的であるが，反面，精度良く学習を行うことが難しい．一方，データ生成過程のモデル化を介さずに，直接的にデータ解析を行うアプローチが近年盛んに研究されている．本稿では，与えられたデータをグループ化するデータクラスタリングを例に取り，生成モデルの推定を介さない最新の機械学習技術を紹介する．

キーワード

機械学習，クラスタリング，相互情報量，生成モデル，密度比

1 はじめに

クラスタリングの目的は，与えられたデータ標本を教師なしでグループ化することである．**K 平均法** [3] は標準的なクラスタリング手法であるが，線形分離可能なクラスタしか抽出できないという弱点があった．

この弱点を克服すべく，様々な非線形クラスタリング手法が開発されてきた．**カーネル K 平均法** [2] では，カーネル関数によってデータを非線形変換した後，K 平均法を実行する．**スペクトルクラスタリング法** [4] では，スペクトル次元削減法によってデータ標本のもつ非線形多様体構造を展開した後，K 平均法を適用する．**識別的クラスタリング法** [9] では，クラスラベルもパラメータとみなして，サポートベクトルマシンなどの識別的なパターン認識器を学習する．**従属性最大化クラスタリング法** [5] では，データ標本との従属性が最大になるようクラスタラベルを決定する．

これらの非線形クラスタリング手法は，複雑な形状のクラスタを表現できるという優れた特徴を持っている．しかしこれらの手法には，非線形性を調整するパラメータが含まれ

ており，これらのパラメータの値は手動で定める必要がある．そのため，客観的なクラスタリング結果を得ることは困難である．

一方，**情報量最大化クラスタリング法** [1] では，データ標本とクラスタラベルの**相互情報量**を最大にするように，ロジスティック回帰モデルなどの確率的なパターン認識器を学習する．このアプローチでは，相互情報量最大化の原理に基づいて，調整パラメータを客観的に決定することができる．また，最適化問題が連続になるという特徴がある．しかし，最適化問題は非凸であり，良い局所解を求めることは容易ではない．

この問題を解決すべく，**二乗損失相互情報量** (squared-loss mutual information; SMI) を用いた新たなクラスタリング手法が提案された [7]．この手法は **SMI クラスタリング法**と呼ばれ，クラスタリング解が解析的に求まるという優れた特徴を持つ．また，カーネル関数のパラメータは，SMI のノンパラメトリック推定量である**最小二乗相互情報量** (least-squares MI; LSMI)[8] を用いることにより，客観的に決定できる．本稿では，SMI クラスタリング法を紹介する．

2 二乗損失相互情報量を用いた情報量最大化クラスタリング

本節では，文献 [7] で提案された SMI クラスタリング法を紹介する．

2.1 情報量最大化クラスタリングの定式化

未知の d 次元の独立同一分布に従う標本が n 個与えられる場合を考える：

$$\{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$$

この未知の確率分布の確率密度関数を $p^*(\boldsymbol{x})$ で表す．クラスタリングの目的は，これらの標本に対してクラスタラベル

$$\{y_i \mid y_i \in \{1, \ldots, c\}\}_{i=1}^n,$$

を割り当てることである．ここで，c はクラスタ数であり，本稿では既知であると仮定する．

情報量最大化クラスタリング [1] では，クラスタリングを教師なしのパターン認識問題として捉え，クラスタラベルの事後確率 $p^*(y|\boldsymbol{x})$ を $\boldsymbol{x}$ と y の情報量が最大になるように学習する．

2.2 二乗損失相互情報量

情報量の尺度として，**二乗損失相互情報量** (squared-loss mutual information; SMI) を用いる．$\boldsymbol{x}$ と y に対する SMI は

$$\mathrm{SMI} := \frac{1}{2}\int \sum_{y=1}^{c} p^*(\boldsymbol{x})p^*(y)\left(\frac{p^*(\boldsymbol{x},y)}{p^*(\boldsymbol{x})p^*(y)} - 1\right)^2 \mathrm{d}\boldsymbol{x} \tag{1}$$

で定義される．ここで，$p^*(\boldsymbol{x},y)$ は $\boldsymbol{x}$ と y の同時確率密度であり，$p^*(y)$ は y の周辺確率を表す．SMI は常に非負の値を取り，$\boldsymbol{x}$ と y が統計的に独立であるときだけゼロを取る．

以下では，SMI を情報量の尺度としたクラスタリング手法を紹介する．

2.3 SMI 最大化クラスタリング

式 (1) の二乗の項を展開すると，SMI は次式のように表現できる：

$$\mathrm{SMI} = \frac{1}{2}\int \sum_{y=1}^{c} p^*(y|\boldsymbol{x})p^*(\boldsymbol{x})\frac{p^*(y|\boldsymbol{x})}{p^*(y)}\mathrm{d}\boldsymbol{x} - \frac{1}{2} \tag{2}$$

クラスタラベルの事前確率 $p^*(y)$ を π_y (for $y = 1,\ldots,c$) に定め，一般性を失うことなく，$\pi_1 \leq \cdots \leq \pi_c$ と仮定する．もし $\{\pi_y\}_{y=1}^c$ が未知であれば，一様分布に設定する：

$$p^*(y) = \frac{1}{c} \text{ for } y = 1,\ldots,c$$

π_y を $p^*(y)$ に代入すれば，式 (2) は以下のように表現できる：

$$\frac{1}{2}\int \sum_{y=1}^{c} \frac{1}{\pi_y} p^*(y|\boldsymbol{x})p^*(\boldsymbol{x})p^*(y|\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{1}{2} \tag{3}$$

一方，クラスタラベルの事後確率 $p^*(y|\boldsymbol{x})$ を以下のカーネルモデルで近似することにする：

$$p(y|\boldsymbol{x};\boldsymbol{\alpha}) := \sum_{i=1}^{n} \alpha_{y,i} K(\boldsymbol{x},\boldsymbol{x}_i) \tag{4}$$

ここで，$\boldsymbol{\alpha} = (\alpha_{1,1},\ldots,\alpha_{c,n})^\top$ はパラメータベクトルであり，$^\top$ は転置を表す．$K(\boldsymbol{x},\boldsymbol{x}')$ は，パラメータ t を持つカーネル関数である．後に示す実験では，**スパース局所スケールカーネル**を用いる：

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \begin{cases} \exp\left(-\dfrac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma_i\sigma_j}\right) & \text{if } \boldsymbol{x}_i \in \mathcal{N}_t(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in \mathcal{N}_t(\boldsymbol{x}_i) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

ここで，$\mathcal{N}_t(\boldsymbol{x})$ は $\boldsymbol{x}$ の t 近傍標本の集合を表し，t がカーネルのパラメータである．また，$\sigma_i = \|\boldsymbol{x}_i - \boldsymbol{x}_i^{(t)}\|$ であり，$\boldsymbol{x}_i^{(t)}$ は $\boldsymbol{x}_i$ の t 番目の近傍標本である．

式 (3) に含まれる $p^*(\boldsymbol{x})$ に関する期待値を，標本 $\{\boldsymbol{x}_i\}_{i=1}^n$ の平均で近似することにより，次の SMI 推定量が得られる：

$$\widehat{\mathrm{SMI}} := \frac{1}{2n}\sum_{y=1}^{c}\frac{1}{\pi_y}\boldsymbol{\alpha}_y^\top \boldsymbol{K}^2\boldsymbol{\alpha}_y - \frac{1}{2} \tag{6}$$

ただし，$\boldsymbol{\alpha}_y := (\alpha_{y,1},\ldots,\alpha_{y,n})^\top$，$K_{i,j} := K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ と置いた．.

各々のクラスタ y に対して，$\|\boldsymbol{\alpha}_y\| = 1$ という拘束条件のもとで $\boldsymbol{\alpha}_y^\top \boldsymbol{K}^2\boldsymbol{\alpha}_y$ を最大化することにする[1]．この最適化問題は**レイリー商**と呼ばれ，最大解が $\boldsymbol{K}$ の主成分（最大固有値に対応する固有ベクトル）で与えられることが知られている．ここで，すべての解 $\{\boldsymbol{\alpha}_y\}_{y=1}^c$ が同じ主成分に縮退するのを防ぐため，解同士の直交性を仮定する：

$$\boldsymbol{\alpha}_y^\top\boldsymbol{\alpha}_{y'} = 0 \quad \text{for } y \neq y'$$

このとき，解は $\boldsymbol{K}$ の上位 c 個の主成分（上位 c 個の固有値に対応する固有ベクトル）$\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_c$ で与えられる．主成分 $\boldsymbol{\phi}_y$ の符号は任意であるため，以下のように符号を定める：

$$\widetilde{\boldsymbol{\phi}}_y = \boldsymbol{\phi}_y \times \mathrm{sign}(\boldsymbol{\phi}_y^\top \mathbf{1}_n)$$

ただし，$\mathrm{sign}(\cdot)$ は符号を表し，$\mathbf{1}_n$ はすべての要素が 1 の n 次元ベクトルを表す.

一方,

$$p^*(y) = \int p^*(y|\boldsymbol{x})p^*(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \approx \frac{1}{n}\sum_{i=1}^{n} p(y|\boldsymbol{x}_i;\boldsymbol{\alpha}) = \boldsymbol{\alpha}_y^\top\boldsymbol{K}\mathbf{1}_n$$

が成り立ち，またクラスタラベルの事前確率 $p^*(y)$ を π_y に定めたことにより，以下の正規化条件が得られる：

$$\boldsymbol{\alpha}_y^\top\boldsymbol{K}\mathbf{1}_n = \pi_y$$

更に，確率は非負であることから，負の推定値をゼロに切り上げることにする.

これらの正規化と非負条件を考慮することにより，$\boldsymbol{x}_i$ のクラスタラベル y_i は，最終的に次式で与えられる：

$$y_i = \operatorname*{argmax}_y \frac{[\max(\mathbf{0}_n, \boldsymbol{K}\widetilde{\boldsymbol{\phi}}_y)]_i}{\pi_y^{-1}\max(\mathbf{0}_n, \boldsymbol{K}\widetilde{\boldsymbol{\phi}}_y)^\top\mathbf{1}_n} = \operatorname*{argmax}_y \frac{\pi_y[\max(\mathbf{0}_n, \widetilde{\boldsymbol{\phi}}_y)]_i}{\max(\mathbf{0}_n, \widetilde{\boldsymbol{\phi}}_y)^\top\mathbf{1}_n}$$

[1]後に再正規化するため，この拘束条件は本質的でない.

ここで，ベクトルに対する max は成分毎の最大値を表す．$[\cdot]_i$ はベクトルの第 i 成分を表す．上記の式変形で $\boldsymbol{K}\widetilde{\boldsymbol{\phi}}_y = \lambda_y \widetilde{\boldsymbol{\phi}}_y$ を用いていることに注意されたい．また，与えられたデータ標本以外の点 $\boldsymbol{x}'$ のクラスタラベル y' は，次式で予測できる：

$$y' := \operatorname*{argmax}_y \frac{\pi_y \max\left(0, \sum_{i=1}^n K(\boldsymbol{x}', \boldsymbol{x}_i)[\widetilde{\boldsymbol{\phi}}_y]_i\right)}{\lambda_y \max(\boldsymbol{0}_n, \widetilde{\boldsymbol{\phi}}_y)^\top \boldsymbol{1}_n}$$

上記のクラスタリング手法は **SMI クラスタリング法**と呼ばれる [7].

2.4 SMI 最大化によるカーネルパラメータの選択

SMI クラスタリング法の解は，カーネル関数 $K(\boldsymbol{x}, \boldsymbol{x}')$ に含まれるパラメータ t に依存する．SMI クラスタリング法は SMI の最大化に基づいているため，SMI を最大にするように t を決定するのが自然であろう．その際，式 (6) で与えられる SMI 推定量 $\widehat{\mathrm{SMI}}$ を用いることができるが，$\widehat{\mathrm{SMI}}$ は $\{\boldsymbol{x}_i\}_{i=1}^n$ だけから求められた教師なしの SMI 推定量であるため，必ずしも推定精度が良いとは言えない．一方，クラスタリングにおけるカーネルパラメータの選択では，すでにラベル付きのデータ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ が得られているため，教師付き学習によってより良い SMI 推定量を得ることが可能である．

教師付き SMI 推定法の一つとして，**最小二乗相互情報量** (least-squares MI; LSMI)[8] と呼ばれる密度比推定に基づくノンパラメトリック推定法が提案されている．LSMI は，推定量が解析的に計算できるという特徴を持ち，さらに漸近的な収束率の意味で最適な推定精度を達成することが示されている．そこで，SMI クラスタリング法のカーネルパラメータの選択に，$\widehat{\mathrm{SMI}}$ でなく LSMI を用いることにする．

LSMI は教師付きデータ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ に基づく SMI 推定法である．LSMI は，$p^*(\boldsymbol{x}, y)$，$p^*(\boldsymbol{x})$，$p^*(y)$ の確率推定を経由せず，次の**密度比関数**

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

を直接するという考え方に基づいている．密度比関数の近似のために，次のカーネル密度比モデルを用いる：

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

ここで，$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ はパラメータベクトルであり，$L(\boldsymbol{x}, \boldsymbol{x}')$ はパラメータ γ をもつカーネル関数である．後に示す実験では，次のガウスカーネルを用いる：

$$L(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\gamma^2}\right) \tag{7}$$

ガウス幅 γ がカーネルパラメータである．

密度比モデルのパラメータ $\boldsymbol{\theta}$ は，次の二乗誤差を最小にするように学習する：

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \int \sum_{y=1}^{c} \Big(r(\boldsymbol{x}, y; \boldsymbol{\theta}) - r^*(\boldsymbol{x}, y) \Big)^2 p^*(\boldsymbol{x}) p^*(y) \mathrm{d}\boldsymbol{x} \tag{8}$$

基底関数 $\{L(\boldsymbol{x}, \boldsymbol{x}_\ell)\}_{\ell: y_\ell = y}$ に対応するパラメータベクトルを $\boldsymbol{\theta}_y$ で表す．すなわち，$\boldsymbol{\theta}_y$ は $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ の $\{\ell \mid y_\ell = y\}$ の要素に対応する部分ベクトルである．クラスタ y に属するデータ数を n_y で表す．この時，各 y に対する (8) の標本近似は次式で与えられる：

$$\min_{\boldsymbol{\theta}_y} \left[\frac{1}{2} \boldsymbol{\theta}_y^\top \widehat{\boldsymbol{H}}^{(y)} \boldsymbol{\theta}_y - \boldsymbol{\theta}_y^\top \widehat{\boldsymbol{h}}^{(y)} + \frac{\delta}{2} \boldsymbol{\theta}_y^\top \boldsymbol{\theta}_y \right] \tag{9}$$

ただし，第３項目は正則化項であり，$\delta\ (\geq 0)$ は正則化パラメータである．$\widehat{\boldsymbol{H}}^{(y)}$ と $\widehat{\boldsymbol{h}}^{(y)}$ は，以下のように定義される $n_y \times n_y$ 行列と n_y 次元ベクトルである：

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

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

ただし，$\boldsymbol{x}_\ell^{(y)}$ はクラスタ y に属する ℓ 番目の標本（$\widehat{\theta}_\ell^{(y)}$ に対応）である．

上記の最適化問題の解 $\widehat{\boldsymbol{\theta}}^{(y)}$ は，解析的に計算できる．

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

これを用いれば，密度比推定量も解析的に求められる：

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

上記の密度比推定量の精度は，カーネル関数 $L(\boldsymbol{x}, \boldsymbol{x}')$ に含まれるパラメータ γ と式 (9) に含まれる正則化パラメータ δ に依存する．これらのパラメータは**クロスバリデーション**を用いることにより，系統的に決定できる．まず，データ標本 $\mathcal{Z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ を M 個の（ほぼ）同じ大きさの重なりのない部分集合 $\{\mathcal{Z}_m\}_{m=1}^M$ に分割する（後の実験では $M = 5$ と設定する）．そして，$\mathcal{Z} \backslash \mathcal{Z}_m$（$\mathcal{Z}_m$ 以外の全ての標本）を用いて密度比推定量 $\widehat{r}_m(\boldsymbol{x}, y)$ を求め，$\mathcal{Z}_m$ を用いて $\widehat{r}_m(\boldsymbol{x}, y)$ の推定誤差を評価する：

$$\mathrm{CV}_m := \frac{1}{2|\mathcal{Z}_m|^2} \sum_{\boldsymbol{x}, y \in \mathcal{Z}_m} \widehat{r}_m(\boldsymbol{x}, y)^2 - \frac{1}{|\mathcal{Z}_m|} \sum_{(\boldsymbol{x}, y) \in \mathcal{Z}_m} \widehat{r}_m(\boldsymbol{x}, y)$$

これを $m = 1, \ldots, M$ に対して繰り返し，それらの平均値を出力する：

$$\mathrm{CV} := \frac{1}{M} \sum_{m=1}^{M} \mathrm{CV}_m$$

そして，このCVの値が最小になるように，カーネルパラメータγと正則化パラメータδを決定する．

こうして得られた密度比推定量$\widehat{r}(\boldsymbol{x},y)$を，式(1)で定義されるSMIの等価表現

$$\mathrm{SMI} = -\frac{1}{2}\int\sum_{y=1}^{c} r^*(\boldsymbol{x},y)^2 p^*(\boldsymbol{x})p^*(y)\mathrm{d}\boldsymbol{x} + \int\sum_{y=1}^{c} r^*(\boldsymbol{x},y)p^*(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x} - \frac{1}{2}$$

に代入すれば，次式のSMI推定量が得られる：

$$\mathrm{LSMI} := -\frac{1}{2n^2}\sum_{i,j=1}^{n}\widehat{r}(\boldsymbol{x}_i,y_j)^2 + \frac{1}{n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i,y_i) - \frac{1}{2}$$

これを，**最小二乗相互情報量** (least-squares MI; LSMI)[8] と呼ぶ．

このLSMIをSMIクラスタリング法のパラメータ決定に用いる．すなわち，式(4)に含まれるカーネル関数$K(\boldsymbol{x},\boldsymbol{x}')$のパラメータ$t$の関数としてLSMIを計算し，LSMIを最大にするようにtを決定する．SMIクラスタリングのMATLABの実装が，以下のウェブページから無償でダウンロードできる．

‘http://sugiyama-www.cs.titech.ac.jp/~sugi/software/SMIC’.

2.5 数値例

最後に，SMIクラスタリング法の実行例を示す．

カーネル関数$K(\boldsymbol{x},\boldsymbol{x}')$として，式(5)に示すスパース局所スケールカーネルを用いる．また，カーネル関数$L(\boldsymbol{x},\boldsymbol{x}')$として，式(7)に示すガウスカーネルを用いる．クラスタラベルの事前確率は一様分布に設定する．

図1の上段にクラスタリング結果，下段にモデル選択結果（カーネル関数$K(\boldsymbol{x},\boldsymbol{x}')$に含まれるパラメータ$t$に対するLSMIの値）を示す．この結果からわかるように，SMIクラスタリング法によって自然なクラスタリング結果が得られている．

3 まとめ

従来のクラスタリング法には，初期値を適切に決定しなければならないという問題，及び，調整パラメータを主観的に決定しなければならないという問題があった．一方，本稿で紹介した二乗損失相互情報量を用いた情報量最大化クラスタリング法では，クラスタリング解が解析に求められ，かつ，調整パラメータを情報量最大化により客観的に決定することができる．

本稿で紹介した技術は，困難なデータ生成過程のモデル化を回避しつつ，直接的にデータのクラスタリングを行うというアプローチであった．近年，このような考え方がより一般化され，第2.4節で示した密度比推定に基づき，様々な機械学習タスクを統一的に解決するという研究が行われている [10, 6]．今後の更なる発展が期待される．

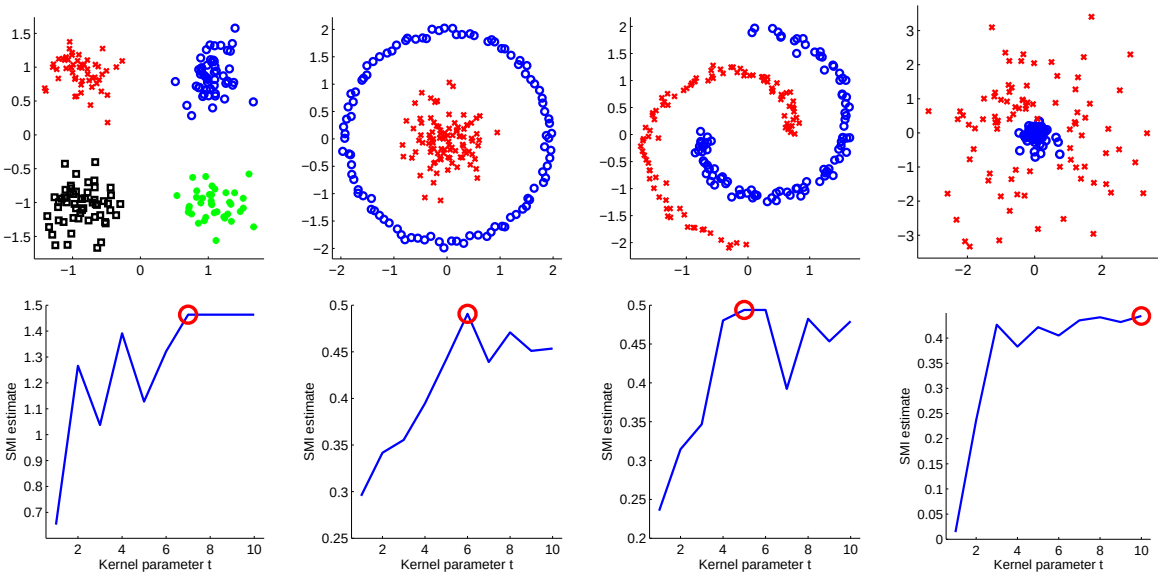

図 1: SMI クラスタリングによるクラスタリング結果（上段）．LSMI によるモデル選択結果（下段）．

参考文献

[1] F. Agakov and D. Barber. Kernelized infomax clustering. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 17–24. MIT Press, Cambridge, MA, USA, 2006.

[2] M. Girolami. Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*, 13(3):780–784, 2002.

[3] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, Berkeley, CA, USA, 1967.

[4] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[5] L. Song, A. Smola, A. Gretton, and K. Borgwardt. A dependence maximization view of clustering. In Z. Ghahramani, editor, *Proceedings of the 24th Annual International Conference on Machine Learning (ICML2007)*, pages 815–822, 2007.

[6] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, Cambridge, UK, 2012.

[7] M. Sugiyama, M. Yamada, M. Kimura, and H. Hachiya. On information-maximization clustering: Tuning parameter selection and analytic solution. In L. Getoor and T. Scheffer, editors, *Proceedings of 28th International Conference on Machine Learning (ICML2011)*, pages 65–72, Bellevue, Washington, USA, Jun. 28–Jul. 2 2011.

[8] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1):S52, 2009.

[9] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1537–1544. MIT Press, Cambridge, MA, USA, 2005.

[10] 杉山 将. 機械学習入門. **オペレーションズ・リサーチ**, 57(7):353–359, 2012.