パラメトリック埋め込み法によるクラス構造の可視化

岩田具治† 斉藤和巳† 上田修功†

データをそのクラス構造とともに低次元の可視化空間へ埋め込む非線形手法，パラメトリック埋め込み法（PE）を提案する．PE は，可視化空間において混合正規分布を仮定し，原空間における事後確率と可視化空間における事後確率の Kullback-Leibler ダイバージェンスができるだけ小さくなるようにデータとクラスを埋め込む．PE は，データ間距離（非類似度）を直接計算しないため，従来法に比べ計算効率が良い．分類済み Web ページを用いたクラスラベル付きデータの可視化，手書き文字を用いた分類モデルの可視化，単語群を用いたラベルなしデータの可視化を行い，PE の有効性を示す．

Class Structure Visualization by Parametric Embedding

Tomoharu Iwata,† Kazumi Saito† and Naonori Ueda†

We propose a nonlinear method, Parametric Embedding (PE), that embeds objects with the class structure into a low dimensional visualization space. We assume a spherical Gaussian mixture in the embedding space, and embeds objects and classes simultaneously by minimizing the Kullback-Leibler divergences between posteriors in the original space and posteriors in the embedding space. PE does not calculate the objects' pair-wise distance directly, therefore, PE has a computational advantage over the conventional embedding methods. In the experiments, we show the validity of PE by visualizing classified web pages, classifiers of handwritten digits, and latent topics of words.

1. はじめに

近年，膨大な電子情報が蓄積されつつある．データ可視化は，こうしたデータの構造を直感的に理解することを可能にし，さらに知識発見のツールとして有用である．古典的な可視化法として，多次元尺度法（MDS）[19] や主成分分析（PCA）[12] が著名であるが，線形射影に基づく手法ゆえの限界がある．近年，非線形構造が抽出可能なより高度な埋め込み法もいくつか提案されている（たとえば文献 11)，18)，20)）．

本論文では，これらの従来法とは異なり，データをそのクラス構造とともに埋め込む問題を考える．すなわち，データ間の関係ではなく，データのクラス帰属確率（事後確率）を保存する新たな可視化手法を提案する．具体的には，可視化空間において等方等分散の混合正規分布（各正規分布の平均ベクトルがクラス情報に相当）を仮定し，原空間におけるクラス事後確率と可視化空間におけるクラス事後確率の Kullback-Leibler（KL）ダイバージェンスができるだけ小さくなるようにデータとクラスを同時に埋め込む．ここで原空間における事後確率は，与えられたデータに適した確率モデルを仮定し得られる．

直感的には，事後確率が高ければそのデータとクラスの可視化空間における距離を近くし，低ければ距離を遠くするように埋め込む．これにより，クラスラベル付きデータを直接的に可視化することができ，そのクラス構造が明らかになる．また，クラスラベルが与えられていない場合でも，応用例で示すように，潜在クラスを導入することにより可視化することが可能である．

提案法は，原空間で仮定された確率モデルを低次元の混合正規分布に変換するものともいえる．従来の可視化法は，可視化空間においてクラス構造を持つ確率モデルを明示的に仮定することなく，データ間の関係（距離や近傍関係など）を基にデータを埋め込むノンパラメトリックな方法であった．一方，提案法はデータのモデルを考えた可視化法ゆえ，パラメトリック埋め込み法（Parametric Embedding: PE）と呼ぶこととする．PE はモデル変換という枠組みを可視化に新たに組み込んだ手法であり，それゆえ，

- データとクラスの非線形埋め込み可能

† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

- 背景知識を確率モデルの形で取り入れ可能
- 計算効率が従来法に比べ高い（後述）

という特長を持つ．

以下の本文では，まず 2 章で提案法 PE について詳述する．3 章では，関連研究について述べる．4 章では，分類済み Web ページ群を用い，ラベル付きデータの可視化において PE が従来法に比べ有効であることを実証する．5 章では，学習サンプル数の異なる 2 つの分類モデルで手書き文字の可視化を行い，PE はデータとともに仮定したモデルの特徴も可視化することを示す．6 章では，単語群に教師なし学習モデルを適用することで，PE はラベルなしデータの潜在クラス構造を可視化することが可能であることを示す．7 章では，結論と今後の課題を述べる．

2. 提案法：Parametric Embedding

可視化対象のデータの集合を $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$，クラスの集合を $\boldsymbol{C} = \{c_k\}_{k=1}^K$ とする．我々の目的は，データのクラス帰属確率（事後確率）$\boldsymbol{P} = \{(P(c_1|\boldsymbol{x}_n), \ldots, P(c_K|\boldsymbol{x}_n))\}_{n=1}^N$ をできるだけ保存しているデータの座標の集合 $\boldsymbol{R} = \{\boldsymbol{r}_n\}_{n=1}^N$ およびクラスの座標の集合 $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_k\}_{k=1}^K$ を求めることである．以後，事後確率は与えられている，もしくは，推定されているものとする．$\boldsymbol{x}_n$ は連続値，離散値だけでなく非整数値でもよい．また $\boldsymbol{r}_n \in \mathcal{R}^D$，$\boldsymbol{\phi}_k \in \mathcal{R}^D$（$D$ は可視化空間の次元で通常 $D = 2$ or 3）である．

可視化空間においても確率モデルを仮定することにより，座標が与えられたときの事後確率 $P(c_k|\boldsymbol{r}_n)$ を推定することができる．ここで，各データの座標は，平均 $\boldsymbol{\phi}_k$，分散共分散行列 I の混合正規分布（混合数 K）に従うと仮定すると，座標 $\boldsymbol{R}$，$\boldsymbol{\Phi}$ から推定される事後確率は，ベイズの定理より

$$P(c_k|\mathbf{r}_n) = \frac{P(c_k)\exp(-\frac{1}{2} \parallel \mathbf{r}_n - \boldsymbol{\phi}_k \parallel^2)}{\sum_{l=1}^K P(c_l)\exp(-\frac{1}{2} \parallel \mathbf{r}_n - \boldsymbol{\phi}_l \parallel^2)} \quad (1)$$

となる．ここで $\parallel \cdot \parallel$ は可視化空間におけるユークリッドノルムを表す．事前確率 $P(c_k)$ が与えられていない場合は全クラス一様と仮定する．上式から明らかなように，この分布を仮定することにより，データの座標 $\boldsymbol{r}_n$ とクラスの座標 $\boldsymbol{\phi}_k$ のユークリッド距離が近ければその事後確率 $P(c_k|\boldsymbol{r}_n)$ は高くなり，直観と一致する可視化結果が得られると期待できる．

この可視化空間における事後確率 $P(c_k|\boldsymbol{r}_n)$ が原空間における事後確率 $P(c_k|\boldsymbol{x}_n)$ の良い近似となっていればよい．この近似は，全データに対する両確率分布間の KL ダイバージェンスを最小化するように $\boldsymbol{R}$ および $\boldsymbol{\Phi}$ を最適化することで実現することができる．$\boldsymbol{P}$ は既知であるため，最小化すべき目的関数 E は次式となる．

$$E(\boldsymbol{R}, \boldsymbol{\Phi}) = -\sum_{n=1}^N \sum_{k=1}^K P(c_k|\boldsymbol{x}_n) \log P(c_k|\boldsymbol{r}_n) \quad (2)$$

特筆すべきは，次式で表される E の $\boldsymbol{r}_n$ に関するヘシアンは，分散共分散行列の形になることから分かるように，半正定値行列となる．

$$\frac{\partial^2 E}{\partial \mathbf{r}_n \partial \mathbf{r}_n'} = \sum_{k=1}^K P(c_k|\mathbf{r}_n)\boldsymbol{\phi}_k\boldsymbol{\phi}_k' - \left(\sum_{k=1}^K P(c_k|\mathbf{r}_n)\boldsymbol{\phi}_k\right)\left(\sum_{k=1}^K P(c_k|\mathbf{r}_n)\boldsymbol{\phi}_k\right)' \quad (3)$$

ここで $'$ は転置を表す．E に正則化項を付け加えた

$$J = E + \eta_r \sum_{n=1}^N \parallel \mathbf{r}_n \parallel^2 + \eta_\phi \sum_{k=1}^K \parallel \boldsymbol{\phi}_k \parallel^2 \quad (4)$$

を目的関数とすると，J の $\boldsymbol{r}_n$ に関するヘシアンは正定値行列となる．ここで η_r，$\eta_\phi > 0$ である．それゆえ，$\boldsymbol{\Phi}$ が与えられたとき，J は $\boldsymbol{R}$ に関して厳密に下に凸となり，$\boldsymbol{\Phi}$ が与えられたとき $\boldsymbol{R}$ の大域的最適性が保証されるという好ましい性質を持つ[13]．

最適化は以下のアルゴリズムにより行う．

(1) $\boldsymbol{R}$ および $\boldsymbol{\Phi}$ を適当に初期化
(2) $\boldsymbol{\Phi}$ を固定し，$\boldsymbol{R}$ に関する J の最適化
(3) $\boldsymbol{R}$ を固定し，$\boldsymbol{\Phi}$ に関する J の最適化
(4) もし収束していれば終了，さもなければステップ (2) へ戻る

ステップ (1) における $\boldsymbol{\Phi}$ の初期値として，クラスに関する情報がある場合，MDS などの次元圧縮法の結果を用いることも可能である．情報がない場合は，$\boldsymbol{R}$，$\boldsymbol{\Phi}$ ともにランダムに初期化する．ステップ (2) および (3) における最適化はニュートン法[13] により行う．このとき，J が増大することはなく，（局所）最適解に収束する．ステップ (3) は複数の局所最適解が存在する非線形最適化問題であるが，ステップ (2) では上述のように唯一最適解が求まるため，最適化過程全体での実質的なパラメータ数は $\boldsymbol{\Phi}$ の要素数，つまりクラス数 K と可視化空間の次元 D の積である．一方，従来の非線形可視化法の場合のパラメータ数は ND となる．一般に $K \ll N$ であるため，従来法と比べ PE は安定した最適化が期待できる．

J の勾配ベクトルは，

$$\frac{\partial J}{\partial \mathbf{r}_n} = \sum_{k=1}^{K} \alpha_{nk}(\mathbf{r}_n - \boldsymbol{\phi}_k) + \eta_r \mathbf{r}_n, \quad (5)$$

$$\frac{\partial J}{\partial \boldsymbol{\phi}_k} = \sum_{n=1}^{N} \alpha_{nk}(\boldsymbol{\phi}_k - \mathbf{r}_n) + \eta_\phi \boldsymbol{\phi}_k, \quad (6)$$

となる．ここで $\alpha_{nk} = P(c_k|\mathbf{x}_n) - P(c_k|\mathbf{r}_n)$ である．興味深いことに，上記学習則は両空間の事後確率の差に応じて $\mathbf{r}_n$ ($\boldsymbol{\phi}_k$) を $\boldsymbol{\phi}_k$ ($\mathbf{r}_n$) に近づけたり遠ざけたりするという，直感的に妥当な学習則となっている．

PE ではデータとクラスの間の関係のみを見るので，その計算量は $O(NK)$ となる．一方，データ間類似度に基づく従来の可視化法の場合，その計算量は $O(N^2)$ となる．一般に $K \ll N$ であるため，PE は従来法に比べ計算量の観点でも優れており，スケーラビリティが高く大規模データにも適用可能である．

3. 関 連 研 究

代表的な線形可視化法として MDS と PCA がある．MDS はデータ間の距離をできるだけ保存するように埋め込み，PCA はデータの分散が最大になるように埋め込む．線形手法は大域的最適解が求まり計算効率が良いが，非線形構造を抽出することができないという欠点を持つ．非線形埋め込み法は Isomap [18]，local linear embedding [17]，stochastic neighbor embedding (SNE) [11]，connectivity preserving embedding [20] など近年数多く提案されている．しかしこれらの手法は，線形手法，PE と比べ計算効率が悪い．また従来の線形手法，非線形手法の多くはクラスの概念を取り入れていない点で PE と異なる．

原空間と可視化空間において確率モデルを仮定するという点で，PE は SNE と共通している．しかしながら SNE では，原空間において PE のようにクラスごとの事後確率 $P(c_k|\boldsymbol{x}_n)$ ではなくデータごとの事後確率 $P(\boldsymbol{x}_m|\boldsymbol{x}_n)$ を用い，可視化空間においてもデータごとの事後確率 $P(\boldsymbol{r}_m|\boldsymbol{r}_n)$ を用いる．つまり，PE はクラスとデータの関係を基にして，SNE はデータ間の関係を基にして埋め込みを行う．この相違点により，PE は計算効率の良いクラス構造可視化が可能となっている．また，全データ間の距離ではなく，データと代表点（クラス）間の距離を基に可視化を行うという点で，ランドマーク MDS (LMDS) [5] は PE と関連深い．LMDS は PE と同様に計算効率は良いが，全データの中からランダムに代表点を選ぶため，代表点の意味が明確でない．

クラス情報を扱うことができる著名な手法として，Fisher 線形判別法 (FLDA) [6]，および FLDA をカーネル法を利用し非線形埋め込み可能なもの拡張したカーネル判別法 (KDA) [1),16)] がある．FLDA は（KDA は特徴空間における）クラス間分散が最大かつクラス内分散が最小になるようにデータを埋め込む．FLDA，KDA ともに，データとクラスの対の集合 $\{\boldsymbol{x}_n, c(n)\}_{n=1}^N$ ($c(n)$ はデータ $\boldsymbol{x}_n$ のクラスラベル）が与えられたときの次元圧縮法であり，事後確率ベクトル集合 $\boldsymbol{P}$ が与えられたときの次元圧縮法である PE とは異なる．そのため，FLDA や KDA は，分類の前処理として高次元データを次元圧縮する場合にも用いられるが，PE は分類の結果である事後確率が必要であるため，分類の前処理としては適していない．一方，PE は分類モデルによって推定された事後確率の集合を可視化することによって，5 章で示すように分類モデルの特徴を可視化することが可能である．

4. ラベル付きデータの可視化

本章では，分類済み Web ページ群を用い，PE がラベル付きデータをどのように可視化する（2 次元空間へ埋め込む）か示す．また，PE を従来法（MDS, FLDA, SNE, KDA）と可視化結果，事後確率保存，計算量の観点で比較する．

4.1 実 験 設 定

使用したデータは，Open Directory Project ☆によって分類された日本語の Web ページで，10 クラスの中から各クラス 500 ページ，全 5,000 ページをサンプリングしたものである．ここで，50 単語以下のページ，複数のクラスに分類されているページは除いた．PE の入力として必要な事後確率ベクトル集合 $\boldsymbol{P}$ は，形態素解析し単語頻度ベクトルに置き換えた後，ナイーブベイズ (NB) モデル [14] を用い推定した（付録 A.1 参照）．また，クラス座標集合 $\boldsymbol{\Phi}$ の初期値は，NB モデルの各クラスのパラメータを MDS により 2 次元ベクトルに次元圧縮したものを用いた．2 章の表記との対応をとると，$\boldsymbol{x}_n$：単語頻度ベクトル，c_k：分類クラス，$P(c_k|\boldsymbol{x}_n)$：NB モデルで推定した事後確率，$N = 5{,}000$，$K = 10$ となる．なお，単語頻度ベクトルの次元は 34,248 である．

4.2 比 較 手 法

PE と関連強い以下の 5 手法を比較対象とした．

- MDS1：単語頻度ベクトル間のコサイン類似度を入力とした MDS（クラス情報を用いない線形手法）
- MDS2：事後確率ベクトルのユークリッド距離を

☆ http://dmoz.org/

入力とした MDS（事後確率を保存する線形手法）
- SNE：事後確率ベクトルの KL ダイバージェンスを入力とした SNE（事後確率を保存する非線形手法）
- FLDA：単語頻度ベクトルを PCA で 2,000 次元に圧縮したものを入力☆とした FLDA（クラス情報を用いる線形手法）
- KDA：単語頻度ベクトルを入力とした KDA☆☆（クラス情報を用いる非線形手法）

手法によって入力される情報が異なる（PE，MDS2，SNE は事後確率ベクトル，FLDA，KDA は単語頻度ベクトルおよびクラスラベル，MDS1 は単語頻度ベクトルのみが入力）ことに注意されたい．

4.3 可視化結果

図 1 (a) は PE での可視化結果である．各点は 1 つの Web ページを表し，その色形はページが分類されていたクラスを示す（図 1 最上部参照）．同じクラスに属するページはクラスタを形成しており，sports と health や，computers と online-shopping など，関連の強いクラスは近くに配置されていることが分かる．sports のように，他のクラスタとの間にデータがあまり配置されていないクラスは，他のクラスとの区別が容易なクラスである．一方，regional のように，中心にあり他のクラスタとの境界が曖昧なクラスは，他の複数のクラスのトピックを含んでいるクラスであると考えられる．また，このようなクラス間の関係だけでなく，可視化結果から特徴的なページも明らかになる．たとえば，クラスタの中心近くに配置されたページはそのクラスの典型的なページであり，クラスタ間に配置されたページは複数のクラスの属性を持つページである．さらに，クラスタの中に異なるクラスのページが含まれていることがあるが，このページは誤って分類されたものである可能性を示唆する．

PE の結果においてクラスごとにデータが分離されているが，その理由を次に述べる．あるデータ $\boldsymbol{x}_n$ の事後確率が，1 つのクラス c_l のみ 1（つまり $P(c_l|\boldsymbol{x}_n)=1$）で，他のクラスは 0（つまり $P(c_k|\boldsymbol{x}_n)=0$ $(k \neq l)$）であるとする．このとき，データ座標 $\boldsymbol{r}_n$ が $\boldsymbol{\phi}_k$ $(k \neq l)$ に近ければ目的関数の値が高くなるため，$\boldsymbol{r}_n$ は $\boldsymbol{\phi}_k$ $(k \neq l)$ から離れた位置に埋め込まれる．本章の実験で用いたデータでは，この $\boldsymbol{x}_n$ のように事後確率が 1 つのクラスのみ高いものが多かったために，このようにクラスごとに分離している．事後確率が一様に分布している場合は，図 1 (a) とは異なりクラスごとに分離することはない．

図 1 (b) は MDS1 の結果である．クラス情報を用いていないため，クラス構造が見えない結果になっている．図 1 (c) は MDS2 の結果である．MDS1 の結果に比べ同じクラスに属すページが近くに配置されているが，中心で複数のクラスが重なってしまっているなどの偏りがあり，PE と比べクラス構造を適切に抽出できていないといえる．図 1 (d) は SNE の結果である．PE と同じように各クラスが分かれており，クラス間の関係や特徴的なページが明らかになっている．なお，単語頻度ベクトルを入力とし SNE を適用した場合は，MDS1 と同様に，クラス構造の見えない可視化結果が得られる．図 1 (e) は FLDA の結果である．FLDA はクラス情報を用いることができるため，MDS1 の結果と比べ同じクラスのページが集まっているものの，多くクラスのページが重なっており，クラス間の関係が不明瞭である．図 1 (f) は KDA の結果である．全クラスが明確に分離されており，クラス間の関係を理解することができる．

4.4 事後確率保存の比較評価

前節では，各手法がどのようにデータを可視化するかを見た．本節では，我々の目的である事後確率の保存が達成されているか定量的に評価する．MDS1，FLDA，KDA は事後確率が入力として与えられないので，ここでは，PE，MDS2，SNE の 3 手法の比較を行う．

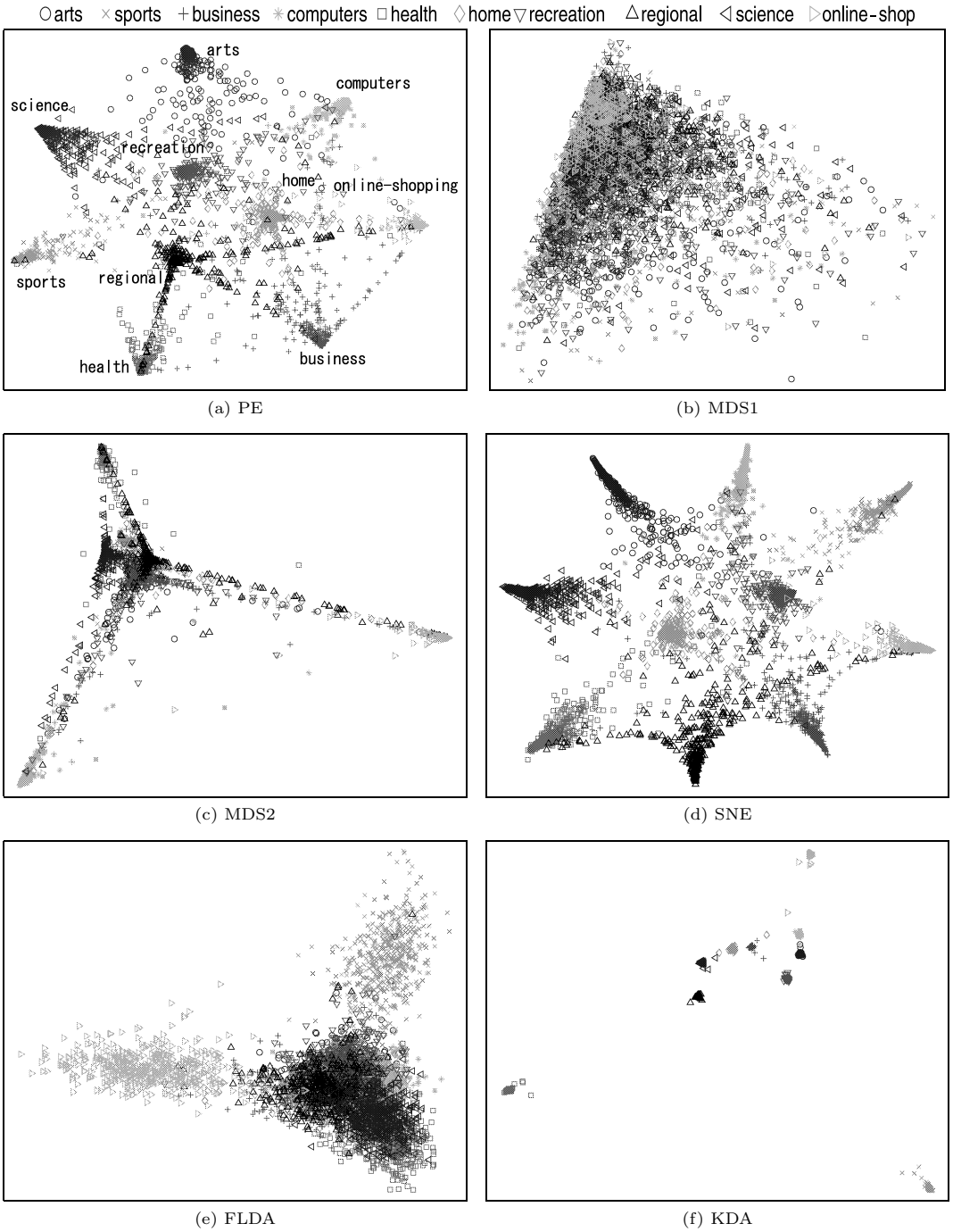
今，データ座標集合 $\boldsymbol{R}$ およびクラス座標集合 $\boldsymbol{\Phi}$ が得られたとする．完全に事後確率が保存されていれば，各クラス c_k において，クラス座標 $\boldsymbol{\phi_k}$ の近傍 h 個のデータの集合と，事後確率 $P(c_k|\boldsymbol{x}_n)$ の高い h 個のデータの集合は一致するはずである．そこで，この 2 つの集合の一致率（precision）を事後確率保存の指標として用いる．

PE 以外の手法ではクラス座標集合 $\boldsymbol{\Phi}$ は得られない．そこで，事後確率が最も高いデータをそのクラスの代表点と考え，各クラス c_k において，事後確率が最も高いデータが埋め込まれた座標 $\boldsymbol{r}_{\arg_n \max P(c_k|\boldsymbol{x}_n)}$ を（PE も含め）クラス座標 $\boldsymbol{\phi}_k$ とする．

近傍数 h を 10 から 500 まで変化させたときの各手法による一致率を**図 2** に示す．横軸は近傍数，縦軸は一致率を表す．PE がすべての近傍数において最も高い一致率を示しており，埋め込み後の事後確率が

☆ データ次元数がデータ数 − クラス数 (N − K) より大きい場合，クラス内分散行列が特異行列となり，FLDA が適用できない（small sample size problem [7]）ため，PCA で次元圧縮した[2]．

☆☆ ガウシアンカーネルを採用し，文献 16) に従い正則化することによって small sample size problem を回避した．



(a) PE (b) MDS1

(c) MDS2 (d) SNE

(e) FLDA (f) KDA

図 1 分類済み Web ページの可視化
Fig. 1 Visualization of classified web pages.

最も保存されているといえる．SNE は近傍数が少ないときやや一致率が低い．これは，SNE は，クラスとデータの関係ではなく，データ間のすべての近傍関係を保存するように埋め込むため，クラスとデータの関係を表す一致率が低くなっていると考えられる．また，MDS2 は近傍数が大きくなるにつれ一致率が低下している．これは図 1 (c) の可視化結果から分かるように，中心付近でクラスで重なっていることが原因である．

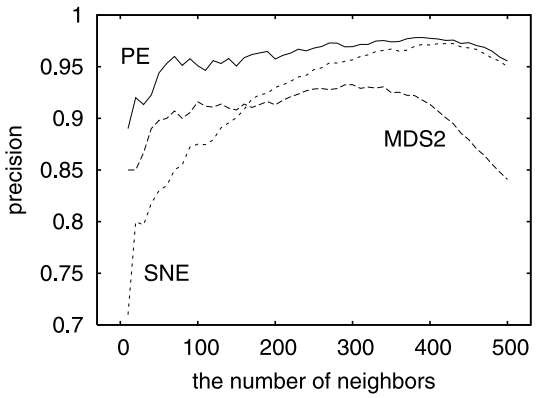


図 2 事後確率保存度の比較実験
Fig. 2 Experimental comparisons of the degree of posterior preservation.

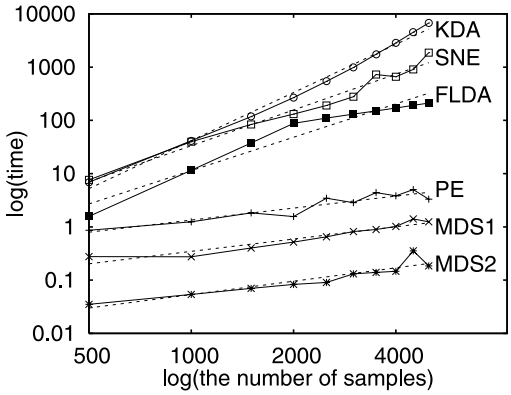


図 3 計算速度の比較実験
Fig. 3 Experimental comparisons of the computational complexity.

4.5 計算時間の比較評価

PE の大きな特長はその計算効率の高さである．2 章で述べたように，PE の計算量は $O(NK)$ である．つまりデータ数に関して線形のオーダでしか計算量は増加しない．MDS は距離行列の固有ベクトルを計算することによって解くことができる．固有値問題は，Lanczos 法[9] によって計算することで効率的に解くことができ，MDS の計算量は，PE と同様にデータ数に関して線形のオーダで増加する．SNE は，データ間の関係を見るためデータ数の 2 乗のオーダの計算量が必要である．FLDA，KDA は一般化固有値問題に帰着でき，一般化固有値問題の計算量は行列サイズの 3 乗のオーダである．

実際の計算量を他手法と比較するため，データ数を 500 から 5,000 まで 500 ずつ変化させたときの計算時間を，Xeon 3.2 GHz，メモリ 2 GB の計算機を用いて調べた．**図 3** はその結果である．横軸はデータ数 N，縦軸は計算時間（秒）であり，両対数プロットしている．点線は，両対数プロットにおける回帰直線である．なお，前処理（PE，MDS2，SNE における事後確率推定，PE における MDS によるクラス座標初期値計算，FLDA における PCA による次元圧縮）の計算時間は省いてある．また，**表 1** にその回帰直線の傾きを示す．この傾きは，データ数の何乗のオーダで計算時間が増加するかを表すものである．上に述べた各手法の理論的な計算量と実際の計算時間がほぼ対応がとれていることが（反復法によって解を求めているため，厳密にではないが）分かる☆．

データ数 5,000 の場合の PE の計算時間は 3.13 秒であり，前処理の時間（2.33 秒）を考慮しても，SNE，FLDA，KDA（それぞれ計算時間は 1,869 秒，211 秒，6,752 秒）に比べきわめて高速である．

表 1 図 3 における回帰直線の傾き
Table 1 The slopes of regression lines in Fig. 3.

PE	MDS1	MDS2	SNE	FLDA	KDA
0.749	0.770	0.834	2.232	2.080	2.998

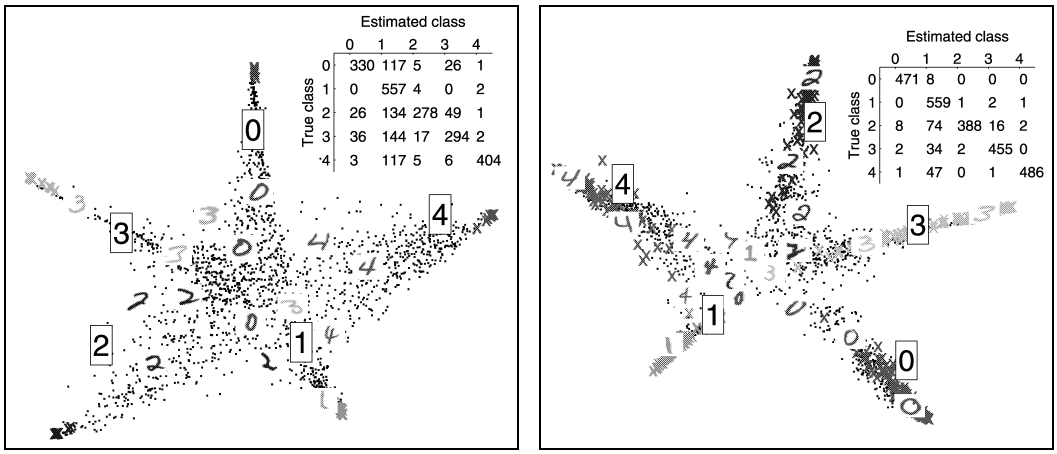
4.6 比較結果まとめ

以上の結果により，事後確率ベクトルが与えられたとき，既存手法に比べ，事後確率を保存し，計算効率が高いことが示された．MDS も PE と同様に計算効率が高いが，適切にクラス構造を抽出することはできない．SNE により PE と近い可視化結果が得られるが，データ数の 2 乗のオーダの計算量が必要であるという問題点がある．FLDA および KDA は，クラス情報を用いた可視化法という点では PE と共通しているが，事後確率保存という観点で可視化をする手法ではない点で PE とは異なり，またより多くの計算時間を必要とする．

5. 分類モデルの可視化

分類問題において基本となる統計的アプローチであるベイズ決定方式[4] では，事後確率の最も高いクラスを推定クラスとして選択する．分類問題を解く際にベイズ決定方式によって得られた事後確率を，できるだけ保存するように可視化することによって，分類モデルの特徴（たとえば，どのクラスの分類が困難か，どのようなデータが誤分類されるかなど）を直感的に理解

☆ FLDA の前処理として，データ数が 2,000 以下のときは $(N-K)$ 次元へ，2,500 以上の場合は 2,000 次元へ，PCA により圧縮を行っている．そのため，計算時間の傾向がデータ数 2,000 あたりで変わっている．2,000 までの計算時間の回帰直線の傾きは 2.898 であり，理論値と近い．



(a) 10 labels/class (b) 100 labels/class

図 4 手書き文字の PE による可視化
Fig. 4 Visualization of handwritten digits by PE.

することが可能になり，モデル選択において有益な情報を与えると考えられる．そこで本章では，ラベル付きデータとラベルなしデータを用いる semi-supervised 分類問題において，学習サンプル数の異なる分類モデルを用いた場合に，PE による可視化結果がどのように異なるかを見る．

実験に用いたデータは，MNIST データベース☆から得た 2,558 の手書き数字（0 から 4 の全 5 クラス）である．以下の手順で 2 種類の分類モデルを構築する．まず，10 もしくは 100 個の手書き数字を学習サンプルとして各クラスからランダムに選ぶ．次に，画素空間において，選ばれた学習サンプルを平均とする分散共分散が等しい正規分布を考える．そして，これらの正規分布の混合を分類モデルとする（付録 A.2 参照）．この分類モデルから得られた 2,558 のすべての数字の事後確率ベクトルを，PE の入力とする．以後，各クラス学習サンプル数 10 の分類モデルを M_{10}，学習サンプル数 100 の分類モデルを M_{100} と呼ぶ．

図 4 (a)，(b) は，それぞれ M_{10}，M_{100} を用いて推定した事後確率を入力としたときの，PE の可視化結果である．ここで，各点は 1 つの画像の座標 $\mathbf{r}_n$，枠付き数字は各クラスの平均の座標 $\boldsymbol{\phi}_k$，× は学習サンプルを表し，いくつかのテストサンプルの画像が表示されている．また，右上の表は confusion matrix である．どちらの場合も 5 つのクラスに対応する 5 つのクラスタが形成されいる．また全体的に見ると，各クラスタは中心に向かって延びており，星型になっていることが分かる．事後確率が 1 つのクラスのみ高いデータは中心からできるだけ遠くに配置され，中心付近にはその分類モデルにとってクラスの特定が困難なデータが配置されている．

図 4 (a) に比べ，図 4 (b) の結果は中心付近にテストサンプルが少なくクラスごとのデータが分かれており，これは M_{100} は分類がより明確にできていることを示している．また，図 4 (b) では学習サンプルがより各クラスタ内で散らばっており，M_{100} の汎化性能が高いことを反映している．一方，図 4 (b) における学習サンプルは中心から最も離れた位置に集中しており，M_{10} は学習サンプルを正確に分類することはできるが，テストサンプルを分類できない可能性が高いことを表している．また，どちらの場合も，「1」のクラスタは他の数字と比べ中心近くに配置されており，これは他のクラスのデータが「1」に誤分類されやすいことを反映している．「1」のクラスタ近くに配置されたデータは「1」に似た形状をしている傾向があることが分かる．図 4 (a) で，「0」と「3」は一部重なって配置されているが，これは，M_{10} ではこのクラス間の区別が困難なことを反映している．また，「2」のデータが「3」のクラスタ方向に広がっているが，これは多くの「2」のデータが「3」と誤分類されてしまうことを反映している．このように学習サンプル数の異なる 2 つの分類モデルの特徴を PE の可視化結果から見てとることができる．

6. 潜在クラスの可視化

上の 2 つの実験では，（少なくとも一部は）ラベル付きのデータの可視化を行ったが，PE は，ラベルなしデータの場合でも，潜在クラスを導入し事後確率を

☆ http://yann.lecun.com/exdb/mnist/

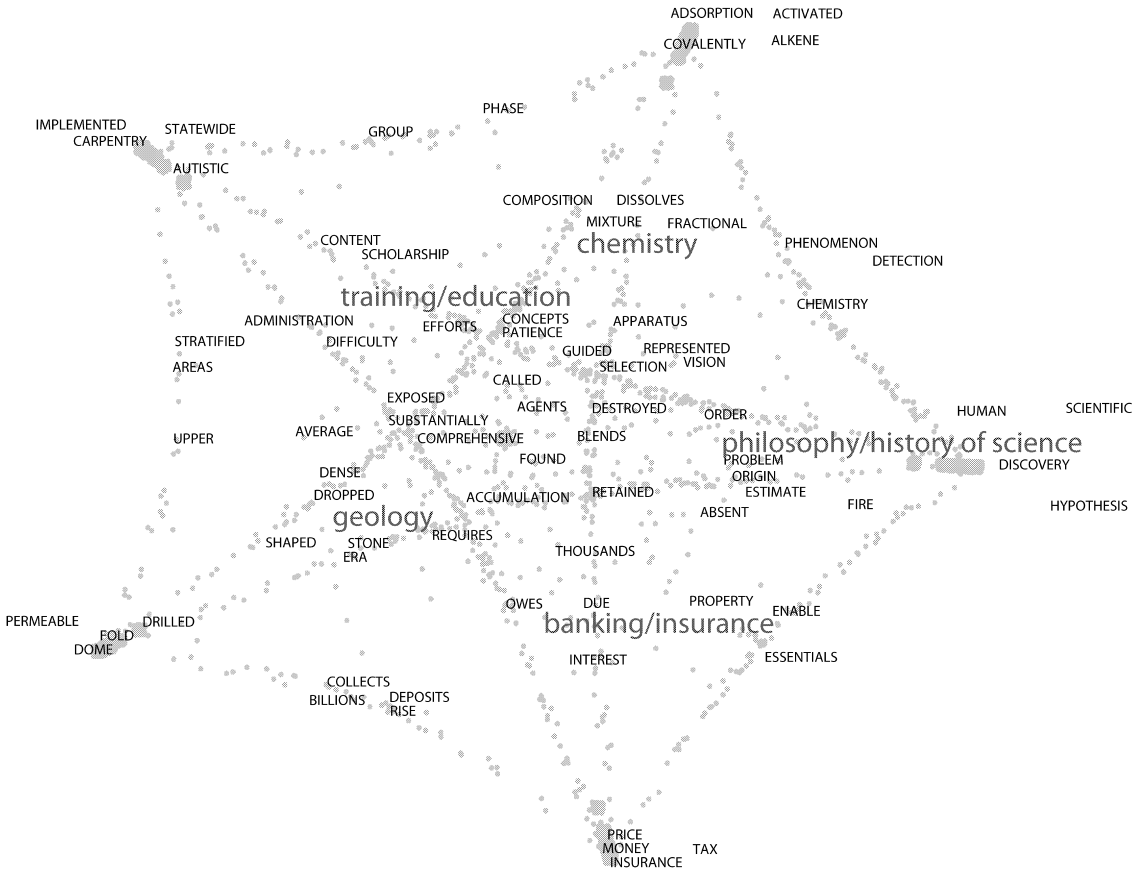


図 5 PE による単語とトピックの可視化
Fig. 5 Visualization of words and topics by PE.

推定することによって適用することが可能である．本章では，PE によるラベルなしデータ可視化の 1 例として，単語の潜在トピックに基づく可視化を行った．ここでデータは単語，クラスは潜在トピックに対応する．用いたデータは，約 37,000 の文書から作成された 26,243 語の単語群（TASA コーパス）である．全単語の潜在クラスに対する事後確率は Latent Dirichlet Allocation（LDA）[3] を用いて推定した（付録 A.3 参照）．LDA は，各文書は潜在クラス（トピック）の混合として生成され，混合比はディレクレ分布に従い，また，各クラスの単語分布は多項分布に従うとするモデルである．

図 5 は，50 の潜在クラスを持つ LDA を用いて TASA コーパスを学習した後，特徴的な 5 つのクラスの事後確率を推定し，PE で可視化した結果である．各点は 1 つの単語の座標 $\mathbf{r}_n$，大きな語句はトピックの平均の座標 $\boldsymbol{\phi}_k$ を表し，いくつかの単語の例を表示している．五角形とその頂点を結ぶ直線上に単語の多くが配置されているが，頂点に集まっている単語はそのトピックにおける典型的な単語であり，中心にいくに従ってより一般的な単語である傾向がある．たとえば，化学で使われる「activated」は chemistry の頂点に位置している．また，すべての 2 つのクラスタ間に曲線が張られている．このパターンは，クラスタを形成している 1 つのトピックのみを持つ単語に加え，2 つのトピックが関係している単語が数多く存在するという，与えられたデータおよびトピックモデルの確率的構造の特徴を表している．曲線上にある単語は 2 つのトピックを持った単語である．たとえば，「deposits」は「堆積」と「預金」という geology（地学）と banking（金融）に関する 2 つの異なる意味があり，geology と banking の頂点で結ばれた曲線上に位置している．

7. おわりに

本論文では，データをそのクラス構造とともに可視化する確率モデルに基づく非線形埋め込み法（PE）を

提案した．PE では原空間および可視化空間において確率モデルを仮定し，両空間の事後確率の KL ダイバージェンスが小さくなるように，データとクラスを埋め込む．PE は，原空間で仮定した確率モデルの特徴も可視化することができる．また，PE は，従来のデータ間類似度に基づく可視化法に比べスケーラビリティが高く，大規模データにも適用可能である．

5 章の実験では，学習サンプル数の異なる 2 つの分類モデルの可視化を行ったが，仮定したモデルは同一のものであった．異なる分類モデルを用いて同一のデータを可視化することで，PE の分類モデル可視化性能のさらなる評価が必要である．また，PE は，原空間における確率モデルを適切に選ぶことで，任意のデータに適用することができる．今後，遺伝子データなどの科学データの可視化を通して，科学知識発見ツールとして本手法をさらに発展させる予定である．

謝辞 4 章の実験における形態素解析では奈良先端科学技術大学院大学松本研究室で開発された「茶筌」[15] を使用した．5 章における手書き文字の可視化，6 章における単語群の可視化の実験は，MIT の S. Stromsten，T. Griffiths，J. Tenenbaum の協力によるものである．

参 考 文 献

1) Baudat, G. and Anouar, F.: Generalized discriminant analysis using a kernel approach, *Neural Computation*, Vol.12, pp.2385–2404 (2000).
2) Belhumeur, P.N., Hespanha, J.P. and Kriegman, D.J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.19, No.7, pp.711–720 (1997).
3) Blei, D., Ng, A. and Jordan, M.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
4) Duda, R.O., Hart, P.E. and Stork, D.G.: *Pattern classification*, 2nd edition, John Wiley & Sons, New York (2001).
5) de Silva, V. and Tenenbaum, J.: Global versus local methods in nonlinear dimensionality reduction, *Advances in Neural Information Processing Systems 15*, pp.705–712 (2002).
6) Fisher, R.: The use of multiple measurements in taxonomic problem, *Annuals of Eugenics*, Vol.7, pp.179–188 (1950).
7) Fukunaga, K.: *Introduction to statistical pattern recognition*, 2nd edition, Academic Press, New York (1990).
8) Gilks, W.R., Richardson, S. and Spiegelhalter, D.J.: *Markov chain Monte Carlo in practice*, Chapman & Hall, New York (1996).
9) Golub, G. and Van Loan, C.: *Matix Computation*, 3rd edition, John Hopkins University Press, Baltimore, Maryland (1996).
10) Griffiths, T.L. and Steyvers, M.: Finding scientific topics, *Proc. National Academy of Sciences*, Vol.101, pp.5228–5235 (2004).
11) Hinton, G. and Roweis, S.: Stochastic neighbor embedding, *Advances in Neural Information Processing Systems 15*, pp.833–840 (2002).
12) Jolliffe, I.T.: *Principal component analysis*, Springer-Verlag, New York (1980).
13) Luenberger, D.: *Linear and nonlinear programming*, 2nd edition, Kluwer Academic Publisher, Massachusetts (2003).
14) McCallum, A. and Nigam, K.: A comparison of event models for naive Bayes text classification, *Proc. AAAI Workshop on Learning for Text Categorization*, pp.41–48 (1998).
15) 松本裕治，北内 啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸：日本語形態素解析システム『茶筌』version 2.2.1 使用説明書 (2000).
16) Mika, S., Ratsch, G., Weston, J., Scholkopf, B. and Muller, K.: Fisher discriminant analysis with kernels, *Neural Networks for Signal Processing IX*, pp.41–48 (1999).
17) Roweis, S. and Saul, L.: Nonlinear dimensionality reduction by local linear embedding, *Science*, Vol.290, pp.2323–2326 (2000).
18) Tenenbaum, J., de Silva, V. and Langford, J.: A global geometric framework for nonlinear dimensionality reduction, *Science*, Vol.290, pp.2319–2323 (2000).
19) Torgerson, W.: *Theory and methods of scaling*, Wiley, New York (1958).
20) 山田武士，斉藤和巳，上田修功：クロスエントロピー最小化に基づくネットワークデータの埋め込み，情報処理学会論文誌，Vol.44, No.9, pp.2401–2408 (2003).

付 録

A.1 Web ページ分類モデル

4 章で事後確率推定に用いた NB モデルについて説明する．NB モデルでは，クラス c_k に属す Web ページ $\boldsymbol{x}_n$ の生成確率を多項分布

$$p(\boldsymbol{x}_n|c_k) \propto \prod_{j=1}^{V} {\theta_{kj}}^{x_{nj}} \tag{7}$$

と仮定する．ここで，V は総単語数，x_{nj} はページ $\boldsymbol{x}_n$ における単語 w_j の頻度，θ_{kj} はクラス c_k のペー

ジで単語 w_j が現れる確率（$\theta_{kj} > 0$, $\sum_{j=1}^{V} \theta_{kj} = 1$）を表す．$\theta_{kj}$ は，最大事後（MAP）推定により計算した．このとき，推定値は

$$\hat{\theta}_{kj} = \frac{\sum_{n \in \boldsymbol{C}_k} x_{nj} + \lambda_k}{N_k + \lambda_k V} \tag{8}$$

となる．ここで N_k はクラス c_k に属する総ページ数，$\boldsymbol{C}_k$ はクラス c_k に属するページ集合，λ_k はハイパーパラメータを表す．λ_k は leave-one-out クロスバリデーション法により推定した．

A.2 手書き文字分類モデル

5 章で事後確率推定に用いた手書き文字の分類モデルについて説明する．学習サンプルを平均とする共分散行列が単位行列の正規分布の混合モデル，つまり，

$$P(\boldsymbol{x}_n|c_k) \propto \sum_{m \in \boldsymbol{C}_k} \exp\left(-\frac{1}{2} \parallel \boldsymbol{x}_n - \boldsymbol{x}_m \parallel^2\right) \tag{9}$$

を分類モデルとした．ここで $\boldsymbol{x}_n$ は手書き文字の画素ベクトル（256 次元），$\boldsymbol{C}_k$ はクラス c_k の学習サンプル集合を表す．

A.3 単語潜在トピックモデル

6 章で事後確率推定に用いた Latent Dirichlet Allocation（LDA）について説明する．LDA は，各文書は潜在クラス（トピック）の混合として生成され，混合比はディレクレ分布に従い，また，各クラスの単語分布は多項分布に従うとするモデルである．具体的には，$\boldsymbol{x}$ を文書，w_m を文書中で m 番目に現れた単語，M を文書 d の総単語数，z_k を潜在トピック，K を潜在トピック数としたとき，文書生成モデルは

$$p(\boldsymbol{x}) = \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\psi}} \left(\prod_{m=1}^{M} \sum_{k=1}^{K} p(w_m|z_k;\boldsymbol{\psi}) p(z_k|\boldsymbol{\theta}) \right) \times p(\boldsymbol{\psi};\boldsymbol{\beta}) p(\boldsymbol{\theta};\boldsymbol{\alpha}) d\boldsymbol{\psi} d\boldsymbol{\theta} \tag{10}$$

となる．ここで $p(w_m|z_k;\boldsymbol{\psi})$，$p(z_k|\boldsymbol{\theta})$ は多項分布，$p(\boldsymbol{\psi};\boldsymbol{\beta})$，$p(\boldsymbol{\theta};\boldsymbol{\alpha})$ はディレクレ分布である．文献 10) に従い，潜在トピック z_k が与えられたときの単語 w_m 生起確率 ψ_{km} をギブスサンプリング[8] により推定した．そして，ψ_{km} を基にベイズ定理を用い，PE の入力となる事後確率（単語 w_m の潜在トピック z_k 帰属確率）を求めた．

(平成 16 年 12 月 9 日受付)
(平成 17 年 7 月 4 日採録)

岩田 具治

昭和 54 年生．平成 13 年慶應義塾大学環境情報学部環境情報学科卒業．平成 15 年東京大学大学院総合文化研究科広域科学専攻修了．同年 NTT 入社．機械学習，異常値検出，可視化の研究に従事．現在，NTT コミュニケーション科学基礎研究所社員．FIT 船井ベストペーパー賞受賞（平成 16 年）．電子情報通信学会会員．

斉藤 和巳（正会員）

昭和 38 年生．昭和 60 年慶應義塾大学理工学部数理科学科卒業．工学博士．同年 NTT 入社．平成 3 年より 1 年間オタワ大学客員研究員．神経回路網，機械学習の研究に従事．現在，NTT コミュニケーション科学基礎研究所主任研究員（特別研究員）．情報処理学会論文賞受賞（平成 9 年）．人工知能学会論文賞受賞（平成 11 年）．FIT 船井ベストペーパー賞受賞（平成 16 年）．人工知能学会研究会優秀賞受賞（平成 17 年）．電子情報通信学会，人工知能学会，日本神経回路学会，IEEE 各会員．

上田 修功（正会員）

昭和 33 年生．昭和 57 年大阪大学工学部通信工学科卒業．昭和 59 年同大学大学院修士課程修了．工学博士．同年 NTT 入社．平成 5 年より 1 年間 Purdue 大学客員研究員．画像処理，パターン認識・学習，ニューラルネットワーク，統計的学習，Web 統計解析の研究に従事．現在，NTT コミュニケーション科学基礎研究所知能情報研究部長創発学習研究グループリーダ兼務，奈良先端科学技術大学院大学客員助教授．日本神経回路学会研究奨励賞受賞（平成 4 年）．電気通信普及財団賞受賞（平成 9 年）．電子情報通信学会論文賞（平成 12，16 年）．FIT 船井ベストペーパー賞受賞（平成 16 年）．人工知能学会研究会優秀賞受賞（平成 17 年）．電子情報通信学会，日本神経回路学会，IEEE 各会員．