LF-013

事後確率構造の可視化
Visualization via Posterior Preserving Embedding

岩田 具治 † 斉藤 和巳 † 上田 修功 †
Tomoharu Iwata Kazumi Saito Naonori Ueda

1. まえがき

近年，膨大な情報が電子的に蓄積されつつあり，価値のある情報を見つけることや，情報の全体的な構造を理解することがますます困難になってきている．この問題に対する 1 つ有効な解決法は，情報を可視化することである．可視化により，データに内在する構造的特徴が浮き彫りになり，知識発見のための重要な手がかりが得られる．本稿で，事後確率を保存しデータを埋め込む (Posterior Preserving Embedding: PPE) 新たな多次元データ可視化法を提案する．提案法では，与えられたデータに対する確率モデルを仮定し各データの事後確率を推定した後，事後確率をできるだけ保存するように 2 次元または 3 次元のユークリッド空間に埋め込むことによりデータを可視化する．

従来の可視化法の多くはデータ間の類似度 (距離) を基に埋め込みを行うが (例えば MDS[1] や Isomap[4])，提案法はデータの事後確率に基づいて埋め込む点で従来法とは本質的に異なる．事後確率を用いることにより，クラス情報を持つデータのクラス構造をより直接的に可視化できる．加えて，提案法は，データの全ての 2 点間の類似度 (距離) の算出が不要故，計算量の観点でも極めて効率的であるという利点を持つ．

2. 事後確率を保存する埋め込み

提案法は，K クラスの事前確率 $P(k)$，$k=1,\ldots,K$ と N データの事後確率をベクトルにした事後確率ベクトル $\mathbf{q}_n = (P(1|d_n),\ldots,P(K|d_n))$，$n=1,\ldots,N$ が与えられたときに，事後確率をできるだけ保存するように全データを低次元に埋め込む．ここで $P(k|d_n)$ はデータ d_n が与えられたときの第 k クラスの事後確率を表す．

まず，第 k クラスに属するデータの中心となる可視化空間における座標 $\boldsymbol{\phi}_k = (\phi_{k1},\ldots,\phi_{kD})$ を考える．D は可視化空間の次元 (通常 $D=2$ または 3) を表す．また，第 n データの座標を $\mathbf{r}_n = (r_{n1},\ldots,r_{nD})$ とする．ここで可視化空間における，第 n データが第 k クラスに属す事後確率を

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

と考える．つまり，$\mathbf{r}_n$ は平均 $\boldsymbol{\phi}_k$，分散共分散行列 I_D の D 次元正規分布に基づくとし，平均である $\boldsymbol{\phi}_k$ に近ければ第 k クラスに属す確率が高いとする．$P(k|\mathbf{r}_n)$ をベクトルにしたものを $\mathbf{s}_n = (P(1|\mathbf{r}_n),\ldots,P(K|\mathbf{r}_n))$ とする．可視化空間における事後確率ベクトル $\mathbf{s}_n$ が与えられた事後確率ベクトル $\mathbf{q}_n$ の十分よい近似となっていればよい．$\mathbf{q}_n$ と $\mathbf{s}_n$ を離散確率分布と考え，両確率分布間の距離を表すカルバック擬距離を全データに対して最小にすることで，近似を実現することができる．このとき $\mathbf{q}_1,\ldots,\mathbf{q}_N$ は既知であるため，全データの座標 $\mathbf{R} = \{\mathbf{r}_1,\ldots,\mathbf{r}_N\}$ を得るための，最大化すべき目的関数は $L = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log s_{nk}$ となる．適当な $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_K\}$ の初期値を与え，$\mathbf{R}$ に関する L の最大化と $\boldsymbol{\Phi}$ に関する L の最大化を収束するまで繰り返すことにより，$\mathbf{R}$ を推定することができる．ここで，$\boldsymbol{\Phi}$ が与えられたとき，L は $\mathbf{R}$ に関して厳密に上に凸となり，収束値が大域的最適解を保証するという好ましい性質を持つ．

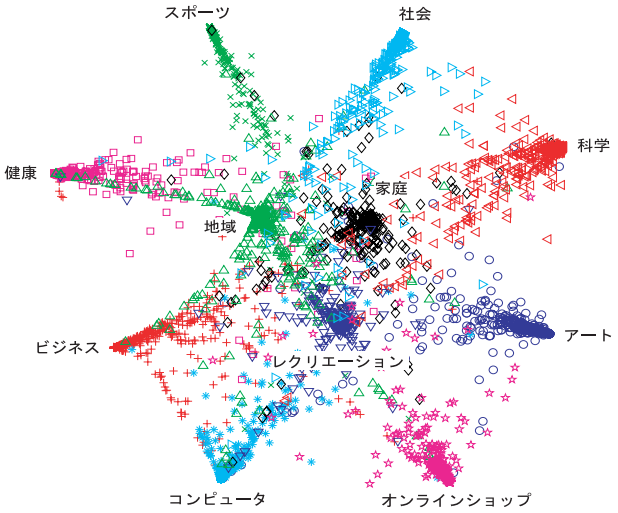
3. 提案法の評価

提案法を評価するため，Open Directory Project (http://dmoz.org/) によって分類された日本語 Web ページ群を提案法および既存法を用い可視化した．使用したデータは，11 クラスの中から各クラス 500 ページ，全 5500 ページをサンプリングしたものである．

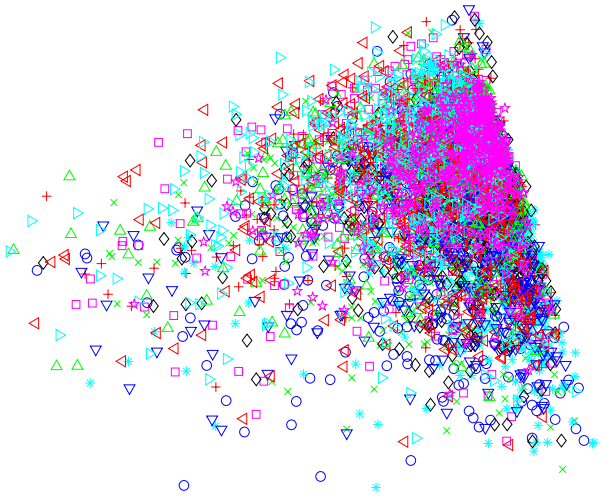
提案法を適用するためには事後確率および事前確率が必要であるが，ここではテキストモデルとして広く用いられているナイーブベイズモデル [3] によってこれらを推定した．図 1(a) は提案法による可視化結果である．同じクラスに属すページがクラスタを形成しており，スポーツと健康など関連の強いクラスは近くに配置されていることが分かる．またクラスタ間に位置するページは複数のクラスに属す可能性があるページを表す．ビジネスとコンピュータの間には多くのページが配置されており，この分類の境界が曖昧であることを意味する．さらに，クラスタのなかに異なるクラスのページが含まれていることがあるが，このページは誤って分類されたものである可能性を示唆する．このように，提案法の可視化結果よりクラス構造やページの特性を一目瞭然に理解することが可能である．

文書群可視化の既存の代表的手法として，単語頻度ベクトルを MDS 用い次元圧縮するものがある [1]．MDS は文書間の類似度を保存するようにデータを埋め込む手法であるが，クラス情報を用いることができない．それ故，図 1(b) に示すように，クラス構造が不明瞭な可視化結果となる．クラス情報を考慮したデータ可視化法として，Fisher 線形判別法を応用した class-preserving projections(CPP)[2] も提案されている．CPP はデータのクラス間分散を最大にするようにデータを低次元空間に埋め込む．図 1(c) は単語頻度ベクトルを基に CPP によって可視化した結果であるが，全クラスのデータが重なり，MDS と同様にデータの特性を抽出することができていない．これは，CPP が線形射影に基づく手法故，データが線形判別不能な場合，クラス構造の抽出法として限界があることを意味する．また，事後確率を MDS によって可視化するという方法が考えられ，図 1(d) はその結果である．同じクラスに属すページが近くに配置

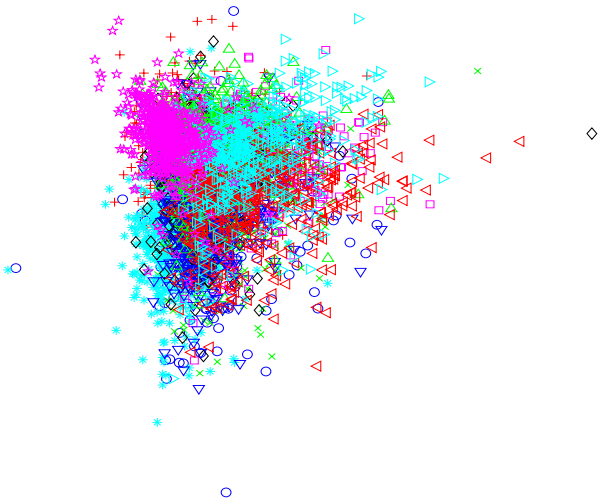
† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所



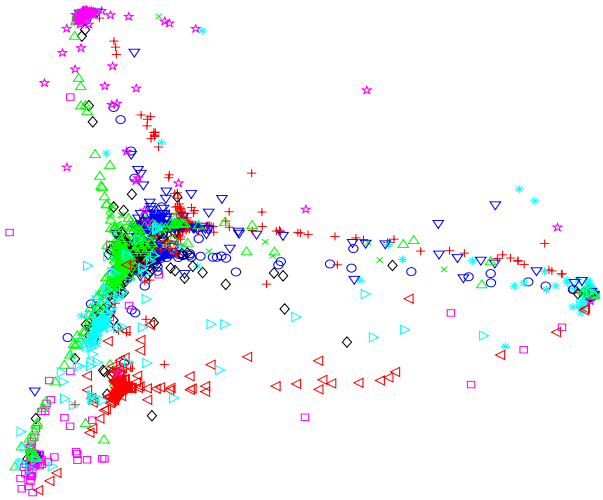
(a) 提案法



(b) 単語頻度ベクトルに基づく MDS



(c) 単語頻度ベクトルに基づく CPP



(d) 事後確率ベクトルに基づく MDS

図 1: 分類された Web ページ群の可視化結果．同じ色形の点は同じクラスに属するページを表す．

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，提案法に比べクラス構造を適切に抽出できていない．

4. あとがき

本稿で，事後確率を保存し埋め込むという新たな多次元データ可視化法を提案し，分類された Web ページ群に適用し提案法の有効性を示した．実験ではクラスラベルつきのデータの可視化を行ったが，クラスラベルのないデータであっても，潜在クラスを導入し事後確率を推定することで提案法を適用することが可能である．提案法は，データの特性だけでなく確率モデルの特性をも可視化していると考えられる．これは可視化結果がモデル選択の際に有益な情報を与える可能性を示唆するものであり，今後，この観点での考察を進める予定である．

参考文献

[1] M.Chalmers and P.Chison, BEAD: Explorations in information visualization, SIGIR'92, ACM Press, pp.330–337, 1992

[2] I.Dhillon, D.Modha and W.Spangler, Class visualization of high-dimensional data with applications, Computational Statistics and Data Analysis 41(1) pp.59–90, 2002.

[3] R.Duda, P.Hart and D.Stork, Pattern classification (2nd ed.), John Wiley & Sons, New York, 2002.

[4] J.Tenenbaum, V.de Silva and J.Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 pp.2319–2323, 2000.