ピボットを用いた類似度データの可視化法

鈴木　紳吾† 　伏見　卓恭†† 　斉藤　和巳† 　池田　哲夫†

† 静岡県立大学経営情報学部経営情報学科　〒 422–8526 静岡県静岡市駿河区谷田 52-1
†† 静岡県立大学経営情報学研究科　〒 422–8526 静岡県静岡市駿河区谷田 52-1
E-mail: †{b08068,j09118,k-saito,t-ikeda}@u-shizuoka-ken.ac.jp

あらまし　類似度が定義されたオブジェクト集合を低次元ベクトルとして埋め込むことは，データの隠れた構造やオブジェクト間の関係を視覚的に把握するために重要である．そのためには，多次元尺度法やスペクトラル法など既存可視化法が利用できる．一方，Bustos らが提案した類似検索法が存在する．本論文では，この類似検索法を土台にピボット可視化法と呼ぶ新たな手法を提案し，既存可視化法と比較する．実験では 2 つのデータ集合を用い，距離を近似させた埋め込みが実現できているかをクラス情報の観点に基づき定性的に評価する．この実験より，ピボット可視化法を用いれば，代表的な既存可視化法に匹敵しつつ，固有の埋め込みが実現できることを示す．
キーワード　埋め込み，可視化，類似度検索法，ピボット

Visualization method of similarity data using pivot

Shingo SUZUKI†, Takayasu FUSHIMI††, Kazumi SAITO†, and Tetsuo IKEDA†

† School of Management and Information, University of Shizuoka
52-1 Yada, Suruga-ku, Shizuoka, 422–8526 Japan
†† Graduate School of Management and Information, University of Shizuoka
52-1 Yada, Suruga-ku, Shizuoka, 422–8526 Japan
E-mail: †{b08068,j09118,k-saito,t-ikeda}@u-shizuoka-ken.ac.jp

Abstract　Embedding a set of objects with some similarity measure into low dimension vectors is important for visually uncovering the latent data structure and the relations between objects. To this end, we can utilize conventional visualization methods such as multidimensional scaling, spectral embedding and so on. On the other hand, there exists a similarity search method proposed by Bustos et al. In this paper, based on this similarity search method, we propose a new method, called pivot visualization, and compare it with these conventional visualization methods. In our experiments using two real object data sets, we qualitatively evaluate the performance of distance preservation in terms of the classification information. From the experiments, we show that the proposed pivot visualization method produces particular embedding results comparable to those of conventional methods.
Key words　embedding, visualization, similarity search method, pivot

1. はじめに

情報爆発時代の到来により，構造化されたデータがインターネット上に大量に蓄積されている．そのため，データ間の関係や特性を把握することは一層重要になっている．複雑な対象を理解しようとするとき，本質的な構造を適切に可視化し，オブジェクト間の関係を知ることができれば極めて有用である．本稿では，多様な可視化技術が存在する中で頻繁に使われている，多次元尺度法，スペクトラル法に着目する．

また，大規模なデータの中から類似のオブジェクトを検索する問題（類似検索問題）がある．クエリと類似度の高いオブジェクトを検索するというタスクである．単純にクエリと全オブジェクト間の類似度（距離）を計算すると膨大な時間が掛かってしまう．計算時間を短縮する手法の 1 つとして，Bustos らが提案したインクリメンタル法がある [1]．これは，距離の三角不等式の性質を活かして，クエリと全オブジェクト間の類似度計算回数を削減する方法である．本研究では，この方法を可視化に応用した手法を提案する．

本論文では，各手法の基本アイデアと可視化アルゴリズムについて説明し，規模の異なる 2 種のデータでの可視化結果を示すことで，提案法の有効性に関して述べる．

2. 可視化手法

与えられたメトリック空間の全オブジェクト集合を $V = \{v_1, \cdots, v_N\}$ とし，オブジェクト v_i と v_j 間の距離（非類似度）が与えられ，全オブジェクト集合を対象とした $N \times N$ の非類似度行列を $\mathbf{G} = \{g_{i,j}\} = \{d(v_i, v_j)\}$ とする．以下では，オブジェクト集合を $\{x_1, x_2, \cdots, x_N\}$ として埋め込むことを考え，その距離行列を $\mathbf{D}$ とする．ここでは簡略化のため，x_i はオブジェクト v_i の 1 次元の埋め込み点を表すとする．なお，一般の m 次元空間上での可視化（埋め込み）への拡張は容易にできるため割愛する．また，各オブジェクトの座標点を要素とするベクトルを $\mathbf{x}^T = (x_1, \cdots, x_N)$ と表す．ここで，$\mathbf{x}^T$ はベクトル $\mathbf{x}$ の転置を表す．

2.1 多次元尺度法

古典的な多次元尺度法 [3] は，$\|\mathbf{x}\|^2 = 1$ の制約のもとで，次式を最大化する座標ベクトルを求める可視化法である．

$$\mathcal{M}(\mathbf{x}; \mathbf{G}) = \mathbf{x}^T(-\frac{1}{2}\mathbf{HGH})\mathbf{x}. \tag{1}$$

ここで，$\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}^T/N$ は，中心化行列またはハウスホルダー変換行列と呼ばれ，任意の N 次元ベクトルに対し，その要素の総和（重心）が 0 になるよう平行移動させる行列である．なお，$\mathbf{1} = (1, 1, \cdots, 1)^T$ を表す．

いま，点集合 $\{y_1, \cdots, y_N\}$ が与えられ，2 点間 y_i と y_j の非類似度をユークリッド距離の自乗 $d_{i,j} = (y_i - y_j)^2$ で定義すれば，上記と同様に非類似度行列 $\mathbf{D}$ を考えることができ次式を得る．

$$\mathcal{M}(\mathbf{x}; \mathbf{D}) = \mathbf{x}^T(\mathbf{H}\mathbf{y}\mathbf{y}^T\mathbf{H})\mathbf{x}. \tag{2}$$

明らかに，$\mathbf{x} \propto \mathbf{H}\mathbf{y}$ で $\mathcal{M}(\mathbf{x}; \mathbf{D})$ は最大化され，適当な平行移動とスケーリングを施せば元の点集合の復元が可能になる．

2.2 スペクトラル法

基本的なスペクトラル法は，$\|\mathbf{x}\|^2 = 1$ の制約のもとで，次式を最小化する座標ベクトルを求める可視化法である [2]．

$$\mathcal{S}(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}(x_i - x_j)^2. \tag{3}$$

ここで，類似度行列 $\mathbf{W}$ の要素を $w_{i,j} = \exp(-d(v_i, v_j)/\sigma)$ と定める．これはパラメータ σ でのヒートカーネルと呼ばれるもので，直感的には，類似度の高い（距離の小さい）オブジェクトペアに対応する要素の値を大きくするようになっている．

式 (3) を最小化するに際して，自明な解 $\mathbf{x}_0 \propto \mathbf{1} = (1, 1, \cdots, 1)^T$ は除外する．直感的には，類似度の高いオブジェクトペアの座標は近くするように可視化する目的関数である．いま，オブジェクト v_i に対応する γ_i を，$\gamma_i = \sum_{j=1}^{N} w_{i,j}$ とし，$\gamma_1, \cdots, \gamma_N$ の順に対角要素として並べて構成した対角行列 $\mathbf{\Gamma}$ を定義すれば，式 (3) は以下のように変形できる．

$$\mathcal{S}(\mathbf{x}) = \mathbf{x}^T(\mathbf{\Gamma} - \mathbf{W})\mathbf{x}. \tag{4}$$

自明な解 $\mathbf{x}_0 \propto \mathbf{1}$ は，行列 $(\mathbf{\Gamma} - \mathbf{A})$ における固有値 $\lambda_0 = 0$ の固有ベクトルに他ならない．よって，次に小さな固有値 $\lambda_1 \leqq \lambda_j$ $(j > 1)$ に対する固有ベクトル $\mathbf{x}_1$ を求めれば，スペクトラル法での可視化結果を得ることができる．

2.3 ピボット法

検索対象となるオブジェクト集合 V に対してレンジクエリを考える．レンジクエリとは，与えられたレンジ $r \in R$ とクエリ q に対し，以下のオブジェクト集合 $R(q, r) \subseteqq V$ を検索するものである．

$$R(q, r) = \{v_n \in V : d(q, v_n) \leqq r\}. \tag{5}$$

出来るだけ少ない距離計算回数で $R(q, r)$ を見つけることが望ましい．他のオブジェクトとの距離を既に計算済みのピボット p_k というオブジェクトを用意する．距離の公理（三角不等式・対称性）より，$d(q, v_n) \geqq |d(q, p_k) - d(v_n, p_k)|$ が得られ，クエリとピボット間の距離 $d(q, p_k)$ からクエリとオブジェクト間の距離 $d(q, v_n)$ に対する下界値を導出することができる．上述したように，ピボットと他のオブジェクトとの距離 $d(v_n, p_k)$ は既に計算済みであることを仮定する．K 個のピボット p_k の集合を P とすると，クエリ q に対して，ピボット集合 P による最大の距離下界値を以下のように定義できる．

$$D(q, v_n; P) = \max_{1 \leqq k \leqq K} |d(q, p_k) - d(v_n, p_k)|. \tag{6}$$

任意のオブジェクトペア v_i と v_j に対し，実際の距離とピボット集合による下界値との関係は $d(v_i, v_j) \geqq D(v_i, v_j; P)$ のようになる．この $D(v_i, v_j; P)$ で $d(v_i, v_j)$ を近似する（下限距離）．

式 (5) および式 (6) より下界値が r 以上のオブジェクト集合 $L = \{v_n; D(q, v_n; P) > r\}$ に対し，距離計算が不要になることがわかる．従って，$|L|$ の期待値が最大になるようなピボット集合 P を求める必要がある．すなわち下界値の総和 $F(P) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} D(v_i, v_j; P)$ を最大化するピボット集合を選択する必要がある．ピボット選択の効率的な手法として，Bustos らのインクリメンタル法がある [1]．この方法は，$k = 1$ から $k = K$ になるまで順に，既に求めたピボット集合を固定し，そこに最良のピボットを逐次追加していく近似解法である（貪欲法）．

（1） 初期化： $P \leftarrow \phi, k \leftarrow 1$

（2） 目的関数 $F(P \cup \{v_n\})$ を最大化するピボットを 1 つ選択： $p_k^* = \arg\max_{v_n \in V} F(P \cup \{v_n\})$

（3） （ 2 ）で選択された新たなピボットを追加： $P \leftarrow P \cup \{p_k^*\}$

（4） $k = K$ ならば終了，otherwise $k \leftarrow k + 1$ とし（ 2 ）へ

選択された最良ピボットは，オブジェクト集合の Outlier に近いことが知られている．しかし，Outlier を選択しても，最良ピボットになるという保証はない．

この方法により求められたピボット集合 P からの距離を用いて，全オブジェクトを K 次元空間で可視化する．

3. 可視化結果と考察

各手法での可視化結果を示すとともに，それぞれの特性を考

0	479	red	#FF0000
1	563	black	
2	488	aqua	#00FFFF
3	493	purple	#800080
4	535	olive	#808000
5	434	green	#008000
6	501	yellow	#FFFF00
7	550	lime	#00FF00
8	462	blue	#0000FF
9	495	navy	#000080

図 1　MNIST の凡例

Non-REM-1	919	black	
Non-REM-2	1479	red	#FF0000
Non-REM-3&4	1658	blue	#0000FF
REM	890	yellow	#FFFF00
wakefulness	1517	green	#008000
indeterminate	74	gray	#808080

図 2　sleep の凡例

察するため 2 種のデータを用いる．MNIST というデータ [4] (オブジェクト数 5,000) と，sleep というデータ [5] (オブジェクト数 6,537)，である．MNIST は手書き文字認識用データベースのことであり，”0,1,2,3,4,5,6,7,8,9”の 10digits の手書き文字の 1 つが，28 × 28=784 画素に，各画素 0–255 の 256 階調グレースケールで表されている．画像の特徴量は，各画素値を要素とする 784 次元ベクトルである．sleep は，睡眠障害の診断に用いられる検査のデータベースのことであり，6 つの特徴により分類されている．(”Non-REM-1,Non-REM-2,Non-REM-3and4,REM,wakefulness,indeterminate”)．各データの詳細は，それぞれ参考文献など参照されたい．

3.1　MNIST の可視化結果

図 3 から図 5 は MNIST への適用で，2 次元平面での多次元尺度法，スペクトラル法，ピボット法の可視化結果を示す．スペクトラル法のパラメータ σ は 0.5 に設定している．図 3 と図 4 を見ると，距離が小さい (類似度が高い) と考えられる同じクラスのオブジェクトは重なったり近くに配置されている，また，距離が大きい (類似度が低い) と考えられる異なるクラスのオブジェクトは遠い位置に配置されている．特に，”3,7”や”0,1”が離れていることがわかる．”4,7,9”は近くに配置されていて類似度が高いことがわかる．図 3 と図 4 の結果の形が似ていることも見て取れる．また図示はしていないが，スペクトラル法のパラメータ σ を 0.1 に設定した場合，多くのオブジェクトが密集してオブジェクト間の関係がわからないような可視化結果となった．また，σ を 1.0, 2.0 と大きく設定するにしたがって，多次元尺度法に近い可視化結果となり，パラメータの変化に対する可視化結果の変化が小さくなっていくことも確認した．図 5 では，”0,1”がピボットとして選ばれ，そのピボットとの距離に基づいて可視化している．そのため，”6”は”1”と類似度が高い，”7”は”1,2”と類似度が低いということが，一目でわかるような可視化結果となっている．Bustos らの主張のように，ピボットとして選択されたオブジェクトは，図 3 と図 4 においても Outlier に近いことがわかる．

3.2　sleep の可視化結果

図 6 から図 8 は sleep への適用で，2 次元平面での多次元尺度法，スペクトラル法，ピボット法の可視化結果を示す．図 4 と同様に，スペクトラル法のパラメータ σ は 0.5 に設定している．MNIST と同じように図 6 と図 7 を見ると，似ている (類似度が高い) と考えられる同じクラスのオブジェクトは重なったり近くに配置されており，似ていない（類似度が低い）と考えられる異なるクラスのオブジェクトは遠い位置に配置されている．また，”Non-REM-2”と”Non-REM-3and4”は互いに同じような配置をしている．図 8 は，”Non-REM-3and4,wakefulness”がピ

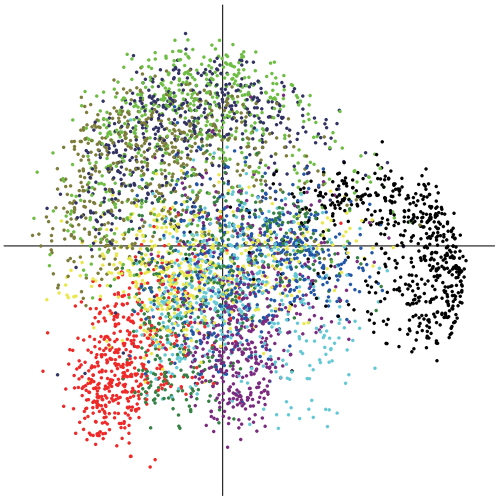

図 3　多次元尺度法 (MNIST)

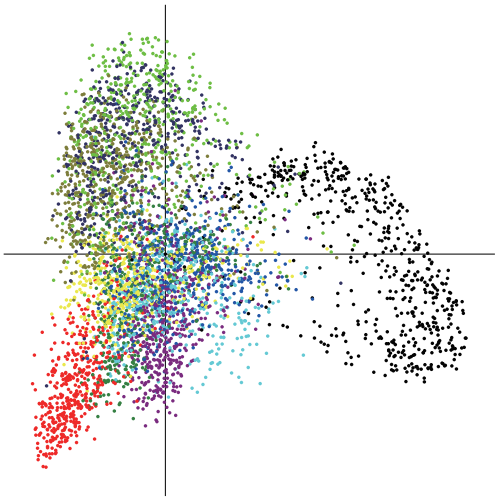

図 4　スペクトラル法 (MNIST)

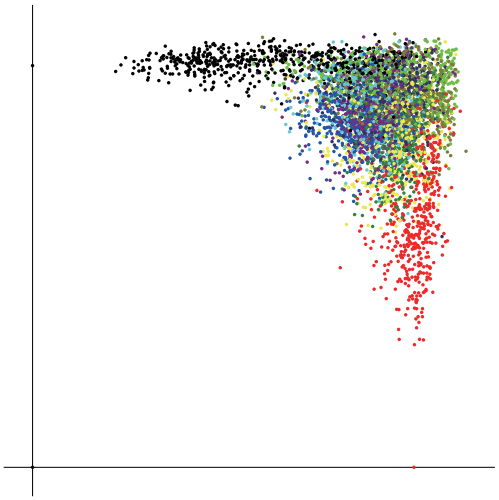

図 5　ピボット法 (MNIST)

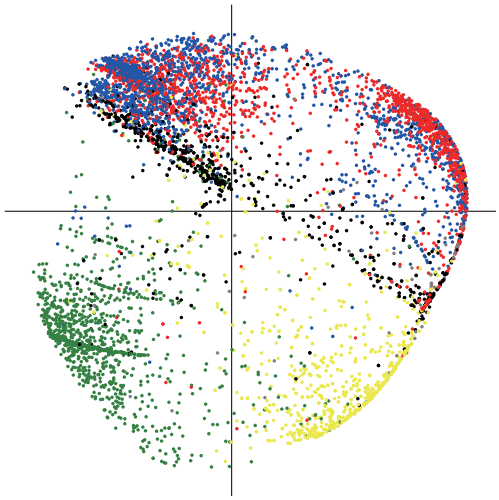

図 6　多次元尺度法 (sleep)

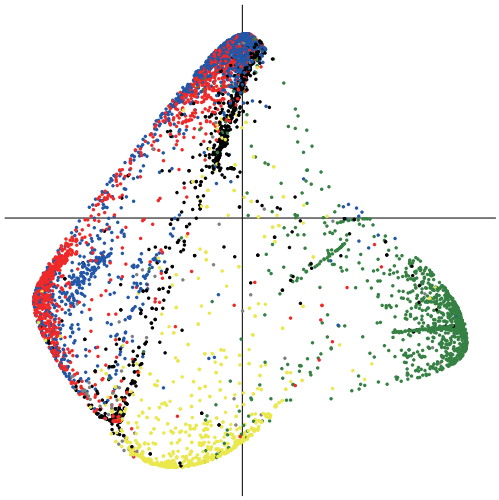

図 7　スペクトラル法 (sleep)

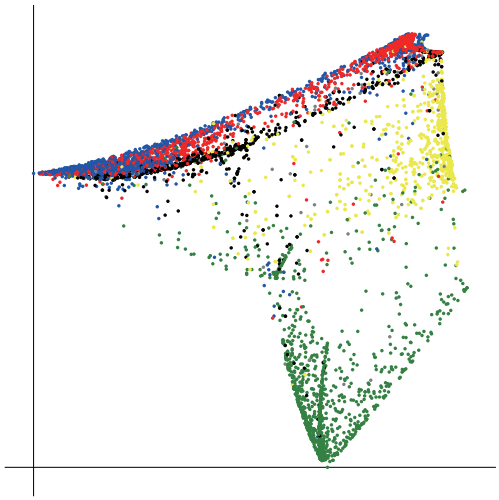

図 8　ピボット法 (sleep)

ボットとして選ばれ，そのピボットとの距離に基づいて可視化している．そのため，”REM”は”Non-REM-3and4,wakefulness”と類似度が低いということが，一目でわかるような図となっている．また，図 4 と同様に，スペクトラル法のパラメータ σ を 0.1 に設定した場合，多くのオブジェクトが密集してオブジェクト間の関係がわからないような可視化結果となった．σ を 1.0, 2.0 と大きく設定するにしたがって，多次元尺度法に近い可視化結果となり，パラメータの変化に対する可視化結果の変化が小さくなっていくことも確認した．

3.3　考　　察

2 つの高次元データに対する実験に関して，多次元尺度法とスペクトラル法では，オブジェクト間の距離が同様に保存され，似たような配置になっていることがわかる．全オブジェクトの可視化結果を見ても，全体として似たような形に配置されている．提案するピボット法でも同様に，オブジェクト間の距離は保たれ，同じクラスのオブジェクトは近くに配置されている．しかし，多次元尺度法とスペクトラル法に比べるとオブジェクト間の距離が縮小して埋め込まれていることが見て取れる．全オブジェクトの可視化結果を見ても，全体としての外観は既存手法と異なる配置になっている．これは，既存手法では各オブジェクト間の距離に関する固有値問題を解いて適切な配置を決定しているが，ピボット法では類似検索における最適なピボットを選択し，そのピボットからの距離に基づいて配置していることが起因していると考えられる．

これら 3 つの可視化手法の比較から，提案するピボット法は従来の手法とは異なる埋め込みを実現する．新たな可視化手法としての有効性を表していることが示唆される．また，類似検索で用いられる手法を可視化に適用しているため，この可視化結果を，類似検索問題だけでなく多くの分野へ応用できることが期待される．

4.　お わ り に

本論文では，類似検索法を土台にピボット可視化法と呼ぶ新たな手法を提案し，既存可視化法と比較した．実験より，ピボット可視化法を用いれば，代表的な既存可視化法に匹敵しつつ，固有の埋め込みが実現できることを示した．今後は，さらに多様なデータに提案法を適用し，その有効性を評価する．

謝辞　本研究は，科学研究費補助金基盤研究 (C) (No. 22500133) の補助を受けた．

文　　献

[1] B. Bustos, G. Navarro, and E. Chavez, “Pivot Selection Techniques for Proximity Searching in Metric Spaces”, Proc. of Pattern Recognition Lettes, Vol.24, No.14, pp. 2357-2366, 2003.

[2] F. R. K. Chung, “Spectral Graph Theory”, Number 92 in CBMS Regional Conference Series in Mathematics, American Mathematical Society, 1997.

[3] W. Torgerson, “Theory and methods of scaling”, Proc. of Wiley New York, 1958.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haner, “Gradient-based learning applied to document recognition”, Proc. of the IEEE, Vol.86, pp. 2278-2324, 1998.

[5] Pablo A. Estevez, Cristian J. Figueroa, and K. Saito, “Cross-entropy Embedding of High-dimensional Data Using the Neural Gas Model”, Neural Networks, Vol.18, pp.727-737, 2005.