3F3-04

混合距離定義に基づく一般化ピボット法の提案とその評価

Calculating Generalized Pivots based on Mixture Distance and its Evaluation

小林 えり[*1] 斉藤 和巳[*1] 池田 哲夫[*1]
Eri Kobayashi Kazumi Saito Tetsuo Ikeda

[*1]静岡県立大学経営情報イノベーション研究科
Graduate School of Management and Information of Innovation, University of Shizuoka

We address a problem of embedding complex objects by a contrective mapping. To this end, we propose a new method for calculating the generalized pivots based on a mixture distance of Manhattan and Euclid. In our experiments using a real data set, we confirmed that out proposed method works better than or comparable to representative conventional methods, in terms of improvements of objective function values, computation times, and the performance of presearching class cohisiness.

1. はじめに

近年，Web 上には多量のデータが蓄積されており，与えられたクエリから類似したオブジェクトを検索する類似検索研究の重要性はますます高まっている．類似検索とは，クエリと類似したオブジェクトをデータベースなどの中から検出する問題を指す．オブジェクト間の類似度は距離関数から求まり距離関数は，非負性，対称性，および三角不等式の性質を満たす．データの多くは高次元で表現されるが，高次元空間に存在するオブジェクト間の距離を求めるのに大量の計算が必要となる．そのため，類似検索ではこの計算量を削減し，検索を高速化させるために一部のオブジェクトをピボット集合として選定して利用する方法が提案されている．効果的なピボット集合を選択する手法として Bustos らはインクリメンタル法 [1](以下，BNC 法：Bustos-Navarro-Chavez 法と呼ぶ) を提案しており，このインクリメンタル法では，より良いピボット集合の指標として目的関数を定義し，目的関数を最大化するようにピボットを逐次追加することでピボット集合を得る．

これに対し，オブジェクト空間の任意の点をピボットとして求める一般化ピボット計算法も提案されている [2]．一般に，オブジェクト集合の中からピボット集合を選択する場合と比較し，オブジェクト空間の任意の点としてピボット集合を求めた結果のほうが良い結果を期待できる．これまでの研究により，一般化ピボット法はユークリッド距離，マンハッタン距離の両方の距離定義に対応できるようになり，Bustos らの提案したインクリメンタル法と比較しその有効性を確認してきた [2] [3]．しかしながら，データの中には距離定義を混合させて類似度を測るもの，また混合させたほうが良い分析結果が得られるようなデータセットが存在する．そのため，本論文ではユークリッド距離定義，マンハッタン距離定義，両者を混合させ距離を定義する，距離定義混合型一般化ピボット法を提案する．

2. 類似検索問題

類似検索には様々な手法が提案されているが，本稿ではクエリから一定のレンジ内にあるオブジェクトを検出するレンジクエリ問題を扱う．レンジクエリ問題は，オブジェクト集合 $Z = \{\mathbf{z}_1, \cdots, \mathbf{z}_N\}$ とクエリ $\mathbf{q}$ とレンジ r が与えられたとき，$\mathbf{q}$ と $\mathbf{z}_n$ の距離 $d(\mathbf{z}_n, \mathbf{q})$ が r 以下となるようなオブジェクト集合を求める問題である．本稿では，レンジクエリ計算時間を短縮させるためにピボット法を用いる．ピボット法は，オブジェクト間の距離計算回数を削減し検索を高速化させるために，一部のオブジェクトを選定してピボット集合を求める手法である．代表的な Bustos らの提案したインクリメンタル法では, クエリ $\mathbf{q}$ に対し，ピボット集合 P による最大の距離下界値 $\mathcal{D}(\mathbf{q}, \mathbf{z}_n; P)$ を次式で定義する．

$$D(\mathbf{q}, \mathbf{z}_n; P) = \max_{1 \le k \le K} |d(\mathbf{q}, \mathbf{p}_k) - d(\mathbf{z}_n, \mathbf{p}_k)| \qquad (1)$$

ここで，ピボット集合の k 番目の要素を $\mathbf{p}_k \in P$，集合の要素数を $K = |P|$ とし，$d(\cdot, \cdot)$ は上述した元空間での距離関数を表す．オブジェクト間の距離は距離公理の三角不等式，距離公理の対称性条件より次式が各ピボット $\mathbf{p}_k$ で成立し，クエリとピボットとの距離 $d(\mathbf{q}, \mathbf{p})$ からクエリとオブジェクトとの距離 $d(\mathbf{q}, \mathbf{z}_n)$ の下界値 $|d(\mathbf{q}, \mathbf{p}) - d(\mathbf{z}_n, \mathbf{p})|$ を算出することができる．これを K 個に拡張すれば式 (1) が導ける．すなわち，最大下界値 $\mathcal{D}(\mathbf{q}, \mathbf{z}_n; P)$ が r 以上のオブジェクト集合を $L = \{\mathbf{z}_n \mid D(\mathbf{q}, \mathbf{z}_n; P) > r\}$ と定義する．明らかに，L に属すオブジェクト集合に対しては距離計算が不要となるため，類似検索計算時間の短縮が期待できる．

3. 一般化ピボット計算法

Bustos ら [1] は，より良いピボット集合の指標として目的関数を定義し，目的関数を最大化させるピボット集合を求める問題として定式化した．Bustos らの手法では，与えられたオブジェクト集合の中から目的関数を最大化するようなピボットを逐次選択する ($\mathbf{p} \in Z$) [1]．

$$\mathcal{F}(P) \quad = \quad \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} D(\mathbf{z}_n, \mathbf{z}_m; P) \qquad (2)$$

これに対し，一般化ピボット計算法では，オブジェクト空間，ここでは H 次元のユークリッド空間 (R^H) の任意の点として，ピボット集合を構築する手法である ($\mathbf{p} \in R^H$)．最適化する目的関数は式 2 と同等の式で表される．

連絡先: 小林 えり，静岡県立大学院経営情報イノベーション学科，静岡県静岡市駿河区谷田 52-1，054-264-5436，rili0906@gmail.com

4. 混合型一般化ピボット計算法

単一の距離定義によるデータセットの場合，その距離定義による計算方法で距離を測り，目的関数を計算すればよい．しかしながら距離定義が混合する場合，各距離定義に合った計算方法でそれぞれの距離を計算しオブジェクト間の距離を決定する必要がある．ここで入力データとなるオブジェクトベクトル $\mathbf{z}$ はマンハッタン距離定義で計算する要素ベクトル $\mathbf{x}$，ユークリッド距離定義で計算する要素ベクトル $\mathbf{y}$ を含むオブジェクトベクトルであり，$\mathbf{z} = (\mathbf{x}^T, \mathbf{y}^T)^T$ と定義する．ここで混合距離に基づく距離関数 $d(\cdot,\cdot)$ は以下の式で計算する．

$$d(\mathbf{z}_n, \mathbf{z}_m) = \alpha d_1(\mathbf{x}_n, \mathbf{x}_m) + (1-\alpha) d_2(\mathbf{y}_n, \mathbf{y}_m) \tag{3}$$

ここで $\alpha(\alpha \in [0,1])$ は距離の混合比を表し，$d_1(\mathbf{x}_n, \mathbf{x}_m)$ はマンハッタン距離定義の要素を用いて計算する距離関数であり，$d_2(\mathbf{y}_n, \mathbf{y}_m)$ はユークリッド距離定義の要素を用いて計算する距離関数である．混合距離定義に基づくデータでの一般化ピボットは，それぞれの距離定義に対応するピボットを求め，それらを混合させることで決定される．$\mathbf{u}$ をマンハッタン距離定義に対応したピボットの要素，$\mathbf{v}$ をユークリッド距離定義に対応したピボットの要素としたとき，オブジェクトベクトル $\mathbf{z}$ に対応する一般化ピボット $\mathbf{w}$ は $\mathbf{w} = (\mathbf{u}^T, \mathbf{v}^T)^T$ で定義される．混合型一般化ピボット法では以下の目的関数を最適化するようなピボットを求める手法である．

$$\begin{aligned}
\mathcal{F}(W|\bar{W}) &= \sum_{n=1}^{N-1}\sum_{m=n+1}^{N} \max_{1\le k\le K} |d(\mathbf{z}_n, \mathbf{w}_k) - d(\mathbf{z}_m, \mathbf{w}_k)| \\
&= \sum_{k=1}^{K}\sum_{n=1}^{N} c_{n,k}(\bar{W}) d(\mathbf{z}_n, \mathbf{w}_k) \\
&= \alpha \sum_{k=1}^{K}\sum_{n=1}^{N} c_{n,k}(\bar{W}) d_1(\mathbf{x}_n, \mathbf{u}_k) \\
&\quad + (1-\alpha) \sum_{k=1}^{K}\sum_{n=1}^{N} c_{n,k}(\bar{W}) d_2(\mathbf{y}_n, \mathbf{v}_k)
\end{aligned} \tag{4}$$

ここで係数 $c_{n,k}(W)$ は，各オブジェクトペアごとに $\mathcal{D}(\mathbf{w}_k, \mathbf{z}_n; W)$ が最大となるピボット番号 k を求めたのち，絶対値を距離の大小で外すとき，距離 $d(\mathbf{z}_n, \mathbf{w}_k)$ がプラス符号で現れる回数と，マイナス符合での回数を相殺した係数を表す．また，一般化ピボット法はアルゴリズムを反復することで最適なピボット集合を構築する手法であり，$\bar{W}$ は更新前ピボット集合を表す．式 4 から第一項はマンハッタン距離定義に基づく一般化ピボット計算法 [3] を用いて求め，第二項はユークリッド距離定義に基づく一般化ピボット計算法 [2] を用いて求めることが出来る．

5. 実験評価

5.1 実験設定

実験データとして，静岡県が提供するオープンソースデータである，”ふじのくにエンゼルパワースポット”という観光データを用いた [5]．県民から広く募集した「恋愛・結婚・子宝」にまつわる噂のスポットに関するデータセットであり，本研究ではこのスポット情報をもとに観光コースを設定し，この観光コースをオブジェクトベクトルとする．観光コースは出発地点から自動車で $50km$ 圏内で回れるスポットによる集合から成り，全コース数，つまりはオブジェクト数は 4137 である．また本実験ではマンハッタン距離定義によるデータベクトルとして，スポットの概要から形成した単語頻度ベクトルを使用，次元数は概要内に出てきた単語数 1397 とする．ユークリッド距離定義によるデータベクトルとして出発地点の緯度経度からなる位置ベクトルを使用し，次元数は 2 とする．

本実験では提案法の基本性能を評価するため，従来法と比較し検証する．提案法である混合距離型一般化ピボット法 (以下 PVL 法)，従来法として Bustos らの提案したインクリメンタル法 (BNC 法) と，この他に，混合距離定義に対応でき，かつ代表的な埋め込み手法である多次元尺度構成法 (以下 MDS 法) [4] の 3 手法をデータ保存率，計算時間，クラス凝集率，可視化結果の 4 観点から比較していく．

また各実験結果でのグラフの色は赤が PVL 法，青が BNC 法，緑が MDS 法の結果を示す．

5.2 データ保存率での評価

本実験では各手法がどの程度元空間でのデータを保存できたか，で評価する．ピボット法での評価指標としてそれぞれの手法の理論上限値を分母に用いて，以下のような評価関数を定義する．

$$\mathcal{G}(W) = \frac{\mathcal{F}(W)}{\sum_{n=1}^{N-1}\sum_{m=n+1}^{N} d(\mathbf{z}_n, \mathbf{z}_m)}$$

すなわち，各手法による距離下限値の良さを比率で評価し，$\mathcal{G}(W)$ が高いほど元空間でのデータ量を保持した結果であると言える．また，縮小写像ではない MDS 法は一般的な寄与率で評価する．

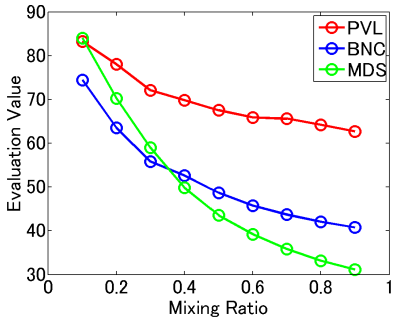
図 1(a) にデータ保存率の観点で 3 手法を比較した結果を示す．横軸には各距離定義情報の混合比を，縦軸には評価値 (データ保存率) を示す．横軸の混合比に関して，例えば 0.2 の場合，マンハッタン距離定義による情報が 20 %，ユークリッド距離定義による情報が 80 % の結果を示す．

結果を見てみると，PVL 法による結果が混合比に関わらず，常に高い評価値を算出している．また他 2 手法と比べて PVL 法は混合比を増加していった際の評価値の減少がゆるやかであり，常に 60 %以上のデータ保存が確認できる．故に PVL 法は本実験では最も元空間での情報を保存することのできる手法だと考えられる．混合比を増加していった際にデータ保存率が落ちてしまう結果に関して，混合比の増加はマンハッタン距離定義による情報の増加を意味しており，ユークリッド距離定義による情報は 2 次元で表現されるのに対し，マンハッタン距離定義による情報は 1397 次元と圧倒的に多い次元数である．故に混合比を増加させていった場合，理論上限値は広範囲のデータ量を表すようになり，そのため各手法がカバーできる情報量，つまりはデータ保存率が落ちていったのだと考えられる．

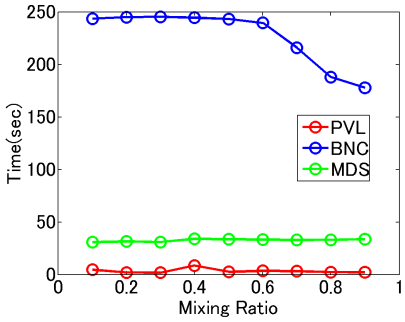
5.3 計算時間での評価

図 1(b) に計算時間の観点で 3 手法を比較した結果を示す．横軸には混合比を，縦軸には計算時間を秒単位で示す．

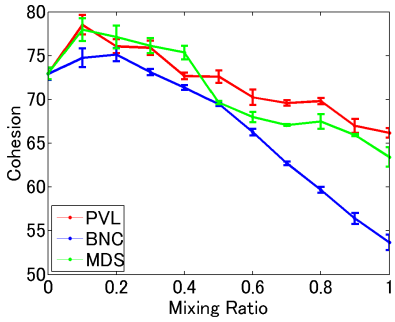
結果を見てみると PVL 法が他の手法に比べて最も短い計算時間で結果を算出していることが分かる．また PVL 法，MDS 法は混合比を変化させても一定の計算時間で結果を出力していることが確認できる．次元 H や反復回数 T に対してオブジェクト数 N の値が高い場合，BNC 法は遅延評価を用いてはいるものの，基本的には，$O(N^3)$ かかってしまうため，PVL 法の時間計算量 $\max(O(THN), O(TN^2))$ や，MDS 法の $\max(O(HN^2), O(TN^2))$ と比べて時間のかかる手法だと判断でき，故に本実験では最も時間がかかってしまったのだと考



(a) データ保存率

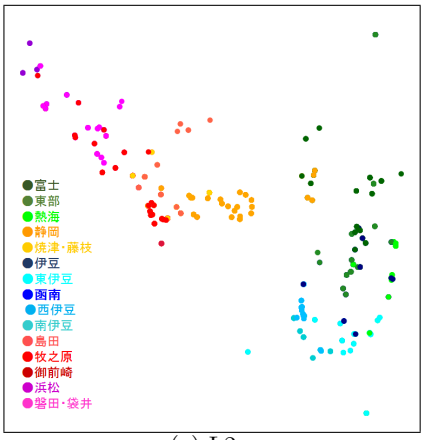


(b) 計算時間

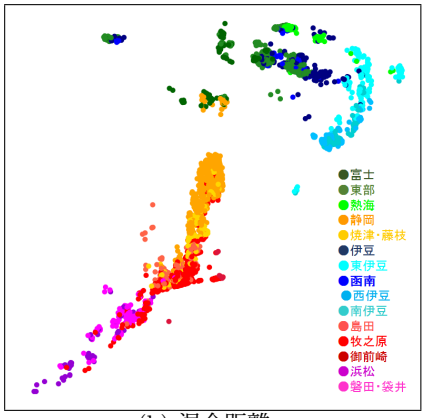


(c) クラス凝集率

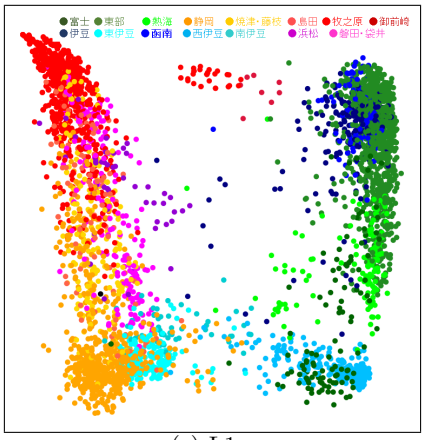
図 1: 評価実験



(a) L2

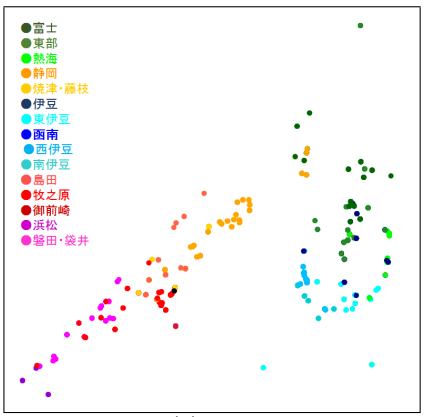


(b) 混合距離

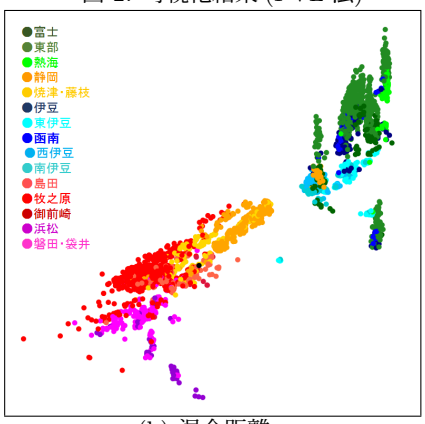


(c) L1

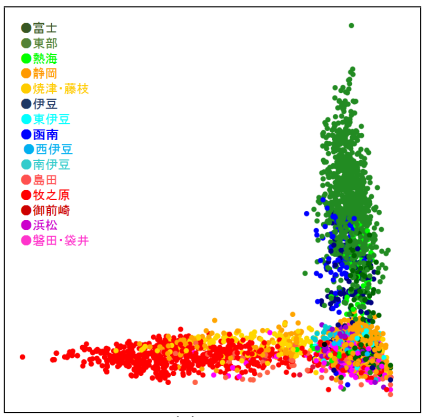
図 2: 可視化結果 (PVL 法)



(a) L2

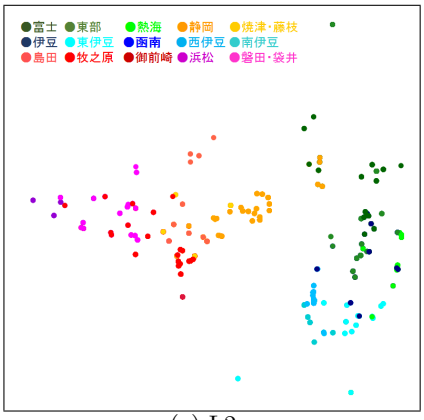


(b) 混合距離

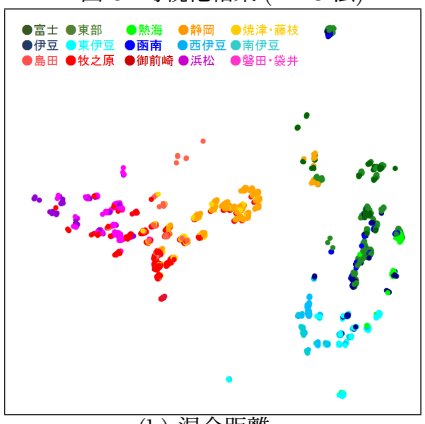


(c) L1

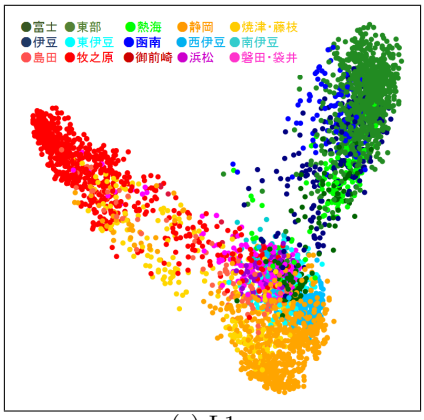
図 3: 可視化結果 (BNC 法)



(a) L2



(b) 混合距離



(c) L1

図 4: 可視化結果 (MDS 法)

えられる．また，MDS 法の計算時間 $O(TN^2)$ は最初に混合距離行列を求める計算量であり，このために多くの計算量を必要としているために MDS 法と PVL 法間にも差が生じたのだと考えられる．

5.4 クラス凝集率での評価

今回類似したコース同士が近傍に配置されているかを確認する評価指標として凝集率を用いた．ここで以下のようなデルタ関数を用いて凝集率を決定する．

$$\delta(A,B) = \begin{cases} 1 & (A = B), \\ 0 & (A \neq B). \end{cases} \tag{5}$$

$$\mathcal{CH}(J) = \frac{1}{NJ} \sum_{n=1}^{N} \sum_{j=1}^{J} \delta(cl(n), cl(r_j(n)))$$

$cl(n)$ は第 n オブジェクト $\mathbf{z}_n$ の属するクラスを表し，r_j は自身を除いた，$\mathbf{z}_n$ から j 番目に近いオブジェクト番号を表す．ここで言うオブジェクト間の近さの尺度は，2 次元空間に埋め込んだ座標でのチェビチェフ距離で求める．本実験ではクラスを観光コースの巡るスポットの属する地域から，最も多く巡る地域で設定する．例えば第 n オブジェクト $\mathbf{z}_n$ は静岡市にあるスポットを 3 つ，焼津市にあるスポットを 1 つめぐるようなコースだった場合，クラスを”静岡市”と設定する．つまり $\mathcal{CH}(J)$ は第 n オブジェクト $\mathbf{z}_n$ から近い順に J 個オブジェクトを取ってきた際，何割のオブジェクトが自身と同じクラスであるか，の平均値を表す．

図 1(c) にクラス凝集率の観点で 3 手法を比較した結果を示す．横軸には混合比を，縦軸には凝集率を示す．本実験では凝集率 $\mathcal{CH}(J)$ は J を 10 と設定した．

結果を見てみると，混合比 0.1 と設定したときの PVL 法の結果が最も高い凝集率を示し，類似したコースオブジェクトを近傍に配置し，非類似コースオブジェクトは遠くに配置していることが確認できた．また，混合比が 0.1〜0.5 間では MDS 法のが最もよい凝集率を示しているが，マンハッタン距離定義による情報量の割合が半分を超えると PVL 法よりも低い凝集率を示し，また 0.4〜0.5 間で凝集率が大きく減少することから次元数の多いデータに弱いということが推測される．またどの手法も単一の距離定義による情報のみの結果である混合比 0.0，1.0 の結果よりも距離定義を混合させた結果のほうが高い凝集率を示しており，距離定義を混合させることで，より精密なクラス凝集率が期待できることが分かった．混合比を増加させる，つまりはマンハッタン距離定義による情報量を増やしていくと，今度は単語頻度ベクトルによる影響が強くなるために位置情報よりも概要内容，つまりはスポットのジャンルを重視した結果になってしまうため，凝集率が低下していったのだと考えられる．

今回設定したクラスは位置情報だけでも事足りるように思えるが，複数のスポットを含むため位置情報も複数存在するため，例えば出発地点が静岡市の場合，そこから東部方面へ行くコースと西部方面に下るコースは類似していないコース同士であるが，出発地点の緯度経度情報だけでは判別することが出来ない．しかしながら，各スポットの概要情報を組み込むことにより，巡るスポットの情報が入るため，各観光コースが主にどの地域を巡るのかをより正確に表すことが出来るようになるのである．ゆえに異なる距離定義の情報を混合させることでより正確な同一クラスの抽出が期待できると本実験では判断できた．

5.5 可視化結果での評価

横軸にピボット 1 との距離を，縦軸にピボット 2 との距離でプロットし，各オブジェクトをクラス (地域) ごとに色分けした可視化結果でオブジェクト間の関係性を視覚的に確認してみる．図 2 に PVL 法での可視化結果を，図 3 に BNC 法での可視化結果を，図 4 に MDS 法での可視化結果を示す．また，各手法ともユークリッド距離定義情報のみの結果を図 2(a)，図 3(a)，図 4(a) で示し，最も凝縮率の良かった混合比での結果を図 2(b)，図 3(b)，図 4(b) で示し，マンハッタン距離定義情報のみの結果を図 2(a)，図 3(c)，図 4(c) に示す．

結果を見てみると，3 手法とも出発地点の位置情報だけだと多くのコースが重なってプロットされており，先ほど述べたとおり，出発地点が同じで進行方向の異なるコースの判別が出来ない．またコースの巡るスポットの概要情報だけの場合，比較的各地域ごとに色分けされてはいるが，伊豆地域クラスを示す青と東部地方クラスを示す緑が混在してしまっていたり，データが散らばったはいいが，正確に類似したオブジェクトコース同士が近傍にプロットされているわけではないことが分かる．しかしながら混合比を用いた結果の場合，出発地点と，スポット概要から得られる進行方向の情報が組み合わされたために，単一距離定義の結果よりも正確に各クラスごとにプロットされた結果であると判断でき，距離定義混合によるクラスごとの分類を視覚的に確認できた．

6. おわりに

本研究ではユークリッド距離，マンハッタン距離定義を混合させた混合型一般化ピボット法を提案し，データ保存率，計算時間，クラス凝集率，可視化結果，以上の 4 つの観点から提案法の特徴を明確化することに成功した．本実験では MDS 法は次元数が大きくなるにつれて凝集率が大きく減る，といった結果から，次元数が大きいデータに弱く，一方，PVL 法は次元数が大きくても対応できる，という仮説が考えられる．そのため，今後はマンハッタン距離定義要素，ユークリッド距離定義要素の両者が大きな次元数を取るようなデータをはじめ，多様なデータで検証を行い，提案法の有効性を検証していく．

謝辞 本研究は，科学研究費補助金基盤研究 (C)(No.23500312) の補助を受けた．

参考文献

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

[2] 木村 学，斉藤 和巳，上田 修功: ”効率的な類似検索のためのピボット学習法”，情報処理学会論文誌，Vol.50, No.8, 1883-1891（2009）

[3] 小林 えり，伏見 卓恭，斉藤 和巳，池田 哲夫: ”マンハッタン距離に基づく一般化ピボット計算法”，”第 6 回 Web とデータベースに関するフォーラム（WebDB2013），Nov.2013.（2013）

[4] J. A. Lee and M. Verleysen: ”Nonlinear Dimensionality Reduction”，Springer (2007).

[5] http://open-data.pref.shizuoka.jp/