Vol.2018-MPS-120 No.17
2018/9/26

グラフ連結性に基づく多様体上での頑健なクラスタリング手法

伏見 卓恭[1,a)]　斉藤 和巳[2,b)]　池田 哲夫[3,c)]　風間 一洋[4,d)]

概要：本研究では，多様体上に分布するオブジェクト群に対して，多様体学習において代表的な ISOMAP と同様に，局所的に近傍するオブジェクト間にリンクを付与することでグラフを構築し，連結性に基づきグラフをクラスタリングすることで頑健で高速なクラスタリングを実現する．この際，距離に比例する確率にしたがってリンクを付与するシミュレーションを多数回実行し，可到達となる期待値が高いオブジェクトノードをクラスタとし抽出する．実データを用いた評価実験では，提案手法の有用性について，クラスタリング精度，データ欠損に対する頑健性，計算速度の観点から評価する．近接中心性を拡張した手法と比較して，提案手法は精度良く高速に頑健なクラスタリング結果を出力することを確認した．

Robust Clustering over Manifold Based on Graph Connectivity

Fushimi Takayasu[1,a)]　Kazumi Saito[2,b)]　Tetsuo Ikeda[3,c)]　Kazuhiro Kazama[4,d)]

1. はじめに

古来より，大量のデータ間の関係や特性を把握することは主要な研究課題の 1 つである．近年でも，膨大な量のデータが Web 上に蓄積されており，情報検索やデータマイニングの分野では，データに潜む有用な情報の抽出や大量データの高速な処理のために多くの手法提案されている．しかし，これらのデータの多くはメトリック空間や高次元ベクトル空間に分布しているため，その実態を把握することは困難である．この問題を緩和するために，多くの次元削減手法 [1], [2], [3]，クラスタリング手法 [4], [5] が提案されている．次元削減やクラスタリングにおいては，オブジェクト間の距離が結果を左右する重要な要因の 1 つである．オブジェクト群が多様体上に分布する場合では，大域的な距離は有用でないことが知られており，局所的な距離に基づく次元削減手法が提案されている [6], [7]．また，文字列や木構造，確率分布などは，オブジェクト間の距離や類似度は定義できるが，ベクトル表現できないため，広く用いられている k-means クラスタリングは適用できない．文献 [8] では，メトリックデータから凝集性と混合性の観点から重要なオブジェクトを抽出する手法を提案している．文献 [6] では，多様体上に分布するオブジェクト群に対し k-NN グラフを構築し，グラフ上での測地距離に基づき MDS により低次元表現を求める ISOMAP を提案している．

本稿では，上述したようなメトリック空間のオブジェクト群，あるいは，多様体上に分布しているようなオブジェクト群に対して，データの欠損などに頑健なクラスタリングを実現する手法を提案する．具体的には，ISOMAP と同様に，オブジェクト間の距離に基づき k-NN グラフを構築する．k-NN グラフ対して集合連結中心性によりノード群をクラスタリングする．文献 [9] では空間ネットワークを対象に，確率的に発生するリンク切断の状況下で到達可

[1] 東京工科大学　コンピュータサイエンス学部
School of Computer Science, Tokyo University of Technology
[2] 神奈川大学　理学部
Faculty of Science, Kanagawa University
[3] 静岡県立大学　経営情報学部
School of Management and Information, University of Shizuoka
[4] 和歌山大学　システム工学部
Faculty of Systems Engineering, Wakayama University
[a)] fushimity@stf.teu.ac.jp
[b)] k-saito@kanagawa-u.ac.jp
[c)] t-ikeda@u-shizuoka-ken.ac.jp
[d)] kazama@sys.wakayama-u.ac.jp

能なノード数の期待値が多いノードを抽出する手法である連結中心性を提案した．さらに，この期待値を厳密に計算するには 2 のリンク数乗の計算量が必要となるが，シミュレーションベースの効率的なアルゴリズムを提案した．

連結中心性は，ノードごとに独立した性質を定量化しているだけであり，ノード間の影響の重複を排除していないため，複数のノードを抽出する際には，一部のノード群に偏った結果が得られる．本稿では，文献 [10] と同様に中心性指標を拡張し，ノード集合に対する連結度を定義し，この値に基づき代表ノード集合を抽出する．集合中心性スコアが高いノード集合を抽出する問題はある種の組合せ最適化問題であるが，目的関数のサブモジュラ性により貪欲解法における最悪ケースの解品質が保証されいるため貪欲法を採用する．本稿では，拡張した連結中心性に基づき k-NN グラフを分割することで，高速かつ頑健なクラスタリングを実現する．

2. 連結中心性

提案指標の前に，ベースとなる連結中心性 [9] について説明する．与えられた無向グラフ構造を $G = (\mathcal{V}, \mathcal{E})$ とし，リンク $e \in \mathcal{E}$ の非切断確率を $p(e; s)$ とする．ここで，s は災害の規模など非切断確率 $p(e; s)$ を制御するパラメータを表しており，便宜上 $0 \leq s \leq 1$ とする．いま，リンク e が切断されれば $x(e) = 0$，さもなければ $x(e) = 1$ となる確率変数を導入する．ここで，リンク非切断確率は $p(x(e) = 1; s) = p(e; s)$ である．これら確率変数の実現値を並べて構成するベクトルを $\mathbf{x} = (\cdots, x(e), \cdots) \in \{0, 1\}^{|\mathcal{E}|}$ とする．この場合，2 値の値を取る確率変数を並べた $\mathbf{x}(s)$ の組合せ数は $|\{0, 1\}|^{|\mathcal{E}|} = 2^{|\mathcal{E}|}$ である．ある組合せ $\mathbf{x}(s)$ の実現確率は同時確率 $q(\mathbf{x}; s) = q(G_{\mathbf{x}}; s) = \prod_{e \in \mathcal{E}} p(e; s)^{x(e)}(1 - p(e; s))^{1-x(e)}$ となる．実現値ベクトル $\mathbf{x}$ に対し，切断されなかったリンク集合 $\mathcal{E}_{\mathbf{x}} = \{e \mid e \in \mathcal{E}, x(e) = 1\}$ とし，そのグラフ構造を $G_{\mathbf{x}} = (\mathcal{V}, \mathcal{E}_{\mathbf{x}})$ とする．$G_{\mathbf{x}}$ においてノード $v \in \mathcal{V}$ と同じ連結成分に属するノード集合を $c(v; G_{\mathbf{x}})$ とする．当然，同じ連結成分に属するノード u と v に対して，$c(u; G_{\mathbf{x}}) = c(v; G_{\mathbf{x}})$ が成り立つ．

リンク切断が発生した状況下での可到達ノード数，すなわち，同一の連結成分に属するノード数の期待値により各ノードの連結度を定義する：

$$cnc_1(v) = \int_0^1 \sum_{\mathbf{x} \in \{0,1\}^{|\mathcal{E}|}} |c(v; G_{\mathbf{x}})| q(\mathbf{x}; s) r_1(s) ds. \quad (1)$$

ここで，$r_1(s)$ はパラメータ s の分布であり，災害の規模に対する想定発生確率から決まるとする．例えば，規模の小さい地震は高確率で発生するが，大震災は低い確率で起きる．式 (1) は，確率 $r_1(s)$ で発生する規模 s の災害によってリンク切断パターン $\mathbf{x}$ が確率 $q(\mathbf{x}; s)$ で発生し，その時のグラフ構造におけるノード v の可到達ノード数 $|c(v; G_{\mathbf{x}})|$ を全切断パターンに関して和を取り，すべての災害規模 $0 \leq s \leq 1$ に関して確率 $r_1(s)$ を掛けながら積分した期待値である．

最も基本的な問題設定として，非切断確率を $p(e; s) = p(s) = s$，$r(s)$ を一様分布に設定し，s による積分を $H+1$ 等分した分割和で求める．h $(0 \leq h \leq H)$ 番目の分割区間に対して，リンク非切断確率は $p(h) = h/H$ と設定する．さらに，$2^{|\mathcal{E}|}$ の和を厳密に求めることは困難なため，J 回のモンテカルロ・シミュレーションにより求める．

j 回目のシミュレーションの h 番目の分割区間において得られるグラフ構造を $G_{(h,j)} = (\mathcal{V}, \mathcal{E}_{(h,j)})$ とすれば，連結中心性の式 (1) に対し，以下の推定式 $cnc_2(v)$ を考えることができる：

$$cnc_2(v) = \frac{1}{HJ} \sum_{h=1}^{H} \sum_{j=1}^{J} |c(v; G_{(h,j)})| r_2(h). \quad (2)$$

ここで，$r_2(h) = r_1(h/H) / \sum_{h'=1}^{H} r_1(h'/H)$ である．明らかに，H や J を十分に大きく設定すれば，式 (2) の推定値 $cnc_2(v)$ は式 (1) の十分精度の高い近似となる．しかしながら，与えられたネットワークのノードとリンクの総数をそれぞれ $N = |\mathcal{V}|$ と $L = |\mathcal{E}|$ とすれば，全ての $v \in \mathcal{V}$ に対して式 (2) を求める計算量は $O(HJ(N + L))$ となり大規模ネットワークへの適用は困難になる．ここで，ノード数 N とリンク数 L のグラフを連結成分分解する計算量は $O(N + L)$ であり，この計算過程で各連結成分のサイズ（可到達ノード数）を計算できる．

以下では，計算量 $O(J(L + N \log N))$ で 式 (2) と同等な精度の推定値を $cnc_3(v)$ として求めるアルゴリズムを説明する．基本アイデアは，全ノードが孤立ノード，すなわち，全リンクが切断された状態 $p(0) = 0$ の初期状態から，もとのネットワークの全リンク $\mathcal{E}$ が追加された状態 $p(H) = 1$ に至るまで，ランダムに選んだリンクを 1 本づつ追加していくことを繰り返す．このとき，リンクを 1 本追加するごとに各ノード $v \in \mathcal{V}$ の期待連結ノード数（可到達ノード数）の差分値を利用して，効率的に連結中心性スコアを計算する．ここで，$H = L$ とする，すなわち，積分の分割数 H をリンク数 L に設定し，J 回繰り返すシミュレーションの第 j 番目では，全てのリンクをランダムにシャッフルし，1 から $H(= L)$ までの ID （$1 \leq h \leq H$）を付与する．ID h を付した各リンクを $e^{(h,j)}$ として，$h = 1$ から順にネットワークに追加する．いま，第 h 番目までのリンクが追加されたリンク集合を $\mathcal{E}^{(h,j)} = \{e^{(h',j)} \in \mathcal{E} \mid h' \leq h\}$ とし，そのときのグラフ構造を $G^{(h,j)} = (\mathcal{V}, \mathcal{E}^{(h,j)})$ とすれば，連結中心性の基本式 (1) に対し，以下の推定式 $cnc_3(v)$ を考えることができる：

$$cnc_3(v) = \frac{1}{JH} \sum_{j=1}^{J} \sum_{h=1}^{H} |c(v; G^{(h,j)})| r_2(h). \quad (3)$$

ここで，グラフ構造 $G^{(h,j)}$ から求まる非切断確率の最尤推定値は $\hat{p} = h/H$ であり，j ごとに独立かつランダムに h 本のリンクが選定（$H-h$ 本のリンクが切断）されていることより，十分大きな J に対し，$\frac{1}{J}\sum_{j=1}^{J}|c(v;G_{(h,j)})|$ と $\frac{1}{J}\sum_{j=1}^{J}|c(v;G^{(h,j)})|$ は同等な精度の推定値を与えることが分かる．したがって，与えられた J に対し，これらの総和として求まる $cnc_2(v)$ と $cnc_3(v)$ も同等な精度の推定値となることが分かる．

以下に，第 j シミュレーションでのリンク追加アルゴリズムの計算量を示す．まず，リンク数 0 の初期状態では，各ノードはそれぞれ異なる連結成分に属するとする．ここで，各ノード $v \in \mathcal{V}$ に一意な連結成分番号 $n(v) \in \{1, \cdots, N\}$ を割り当てる．リンク $e^{(h,j)} = (x,y)^{(h,j)}$ が追加されるとき，ノード x と y が同じ連結成分に属すなら，何もせず次のリンク追加に進む．さもなければ，片方の連結成分に属する全ノードに対して，連結成分番号を更新する．たとえば，$|c(x;G^{(h,j)})| \leq |c(y;G^{(h,j)})|$ とすると，小さい連結成分に属するノードの連結成分番号を大きい連結成分の番号に書き換える：$n(z) \leftarrow n(x)$ for each $z \in c(y;G^{(h,j)})$．したがって，1 本のリンク追加において，連結成分番号が更新されるノード数の最大値は高々 $N/2$ である．よって，すべてのリンク追加では，連結成分番号の更新にかかる計算量は $O(N\log N)$ となる．

いま，$cnc_3^{(j,h)}(v)$ を以下のように定義される第 j シミュレーションにおける $h' = h$ までの $|c(v;G^{(h',j)})|$ の部分和とする：

$$cnc_3^{(j,h)}(v) = \sum_{h'=1}^{h} |c(v;G^{(h',j)})| r_2(h'). \tag{4}$$

新たなリンク $e^{(h,j)} = (x,y)^{(h,j)}$ が第 h ステップで追加されノード x と y が初めて同じ連結成分に属するようになったとき，任意の $h' \geq h$ に対して $c(x;G^{(h',j)}) = c(y;G^{(h',j)})$ であるため，以下の関係が成り立つ：

$$cnc_3^{(j,h')}(x) - cnc_3^{(j,h')}(y) = cnc_3^{(j,h-1)}(x) - cnc_3^{(j,h-1)}(y). \tag{5}$$

よって，各連結成分において，1 つの代表ノード x のみ部分和 $cnc_3^{(j,h')}(x)$ を保持しておき，同じ連結成分内の他のノード y との差分値 $cnc_3^{(j,h-1)}(x) - cnc_3^{(j,h-1)}(y)$ を保持しておくことで，最終ステップ H での和 $cnc_3^{(j,H)}(y)$ は式 (5) を用いて計算できる．ここで，すべてのノード $v \in \mathcal{V}$ に対して，$cnc_3^{(j,H)}(v)$ を計算するのにかかる計算量は $O(N)$ である．また，連結成分番号の更新と同時に差分値も更新するため，差分値更新の計算量は $O(N\log N)$ である．第 j シミュレーションにおいて全リンクをシャッフルして 1 本ずつ追加するため，アルゴリズム全体の計算量は $O(J(L + N\log N))$ である．

3. 提案手法

本研究では，多様体上に分布するオブジェクト群に対して頑健なクラスタリングを実現するための手法を提案する．提案手法は，全 N 個のオブジェクトの集合を $\mathcal{U}$ を入力とし，以下の手順で全オブジェクトを κ 個のクラスタに分割する：

(1) オブジェクト間の距離に基づき，k-NN グラフを構築する；
(2) k-NN グラフに対して，集合連結中心性により代表オブジェクトを抽出する；
(3) オブジェクト群を代表オブジェクトとの連結度に従い κ 個のクラスタに分割する；

第 1 ステップで，ISOMAP と同様に k-NN グラフを構築し，無向化する．これにより，すべてのオブジェクト $\mathcal{U}$ はグラフのノード $\mathcal{V}$ として扱われ，いくつかのオブジェクトノード間には距離重みの付されたリンクが生成される．k-NN グラフにおける k の値は，$k = 1$ から増やしていき，全オブジェクトが単連結となる最小の k を採用する．第 2 ステップで，構築した k-NN グラフに対して，集合連結中心性を適用し，連結度の高い代表ノード集合を抽出する．第 3 ステップで，その他ノードそれぞれを，自身と最も連結度の高いの代表ノードのクラスタに割り当てる．次節以降で，連結中心性を拡張した集合連結中心性，集合連結中心性により抽出した代表ノードにより，他のノードを分割する方法について説明する．

3.1 集合連結中心性

集合連結中心性では，ノード集合 $\mathcal{R}$ に対して連結度を定義し，連結度が最大となるノード集合を代表ノード集合 $\mathcal{R} \subset \mathcal{V}$ として抽出する．ここで，$\mathcal{V}$ はオブジェクトノード集合である．与えられた k-NN グラフの構造を $G = (\mathcal{V}, \mathcal{E})$ とする．任意のノード集合 $\mathcal{R}$ に対する連結度を以下のように定義する：

$$cnc_1(\mathcal{R}) = \int_0^1 \sum_{\mathbf{x}\in\{0,1\}^{|\mathcal{E}|}} |c(\mathcal{R};G_{\mathbf{x}})| q(\mathbf{x};s) r_1(s) ds. \tag{6}$$

ここで，$c(\mathcal{R};G_{\mathbf{x}}) = \bigcup_{r\in\mathcal{R}} c(r;G_{\mathbf{x}})$ であり，集合 $\mathcal{R}$ の各ノード r が属する連結成分の合併集合である．すなわち，$cnc_1(\mathcal{R})$ は，リンク切断が確率的に発生する状況下で $r \in \mathcal{R}$ のいずれかに到達可能であるノード数の期待値を表している．連結中心性と同様に，$r(s)$ を一様分布に設定し，s による積分を $H+1$ 等分した分割和で求める．

$$cnc_3(\mathcal{R}) = \frac{1}{JH}\sum_{j=1}^{J}\sum_{h=1}^{H} |c(\mathcal{R};G^{(h,j)})| r_2(h). \tag{7}$$

本稿では，貪欲法により，以下の Marginal Gain が最大になるように代表ノードを 1 つずつ求めていく．

$$MG(v;\mathcal{R}) = cnc_3(\mathcal{R}\cup\{v\}) - cnc_3(\mathcal{R}) = \frac{1}{JH}\sum_{j=1}^{J}\sum_{h=1}^{H} mg(v;\mathcal{R})^{(h,j)} r_2(h).$$

ここで，$mg(v;\mathcal{R})^{(h,j)} = |c(\mathcal{R}\cup\{v\}; G^{(h,j)}) \setminus c(\mathcal{R}; G^{(h,j)})|$ は，代表ノードの候補であるノード v を $\mathcal{R}$ に追加したとき，代表ノード群から可到達なノード数の増分を表している．この増分 $MG(v)$ が最も高い候補ノード v を代表ノードとして抽出する．κ 個の代表ノードを抽出する過程の中で，k 番目の代表ノードは，$\hat{r}_k \leftarrow \arg\max_{v\in\mathcal{V}\setminus\mathcal{R}_{k-1}} MG(v;\mathcal{R}_{k-1})$ により求め，$\mathcal{R}_k \leftarrow \mathcal{R}_{k-1}\cup\{\hat{r}_k\}$ とする．

提案手法では，リンク $e=(x,y)$ の非切断確率 $p(e;s)$ は，k-NN グラフの各リンクに付された距離重みに比例するように定める：

$$p((x,y)) = \frac{\exp(-d(x,y))}{\sum_{e\in\mathcal{E}}\exp(-d(e))}. \quad (8)$$

3.2 連結度分割

リンク追加シミュレーションにおいて，ノード v が代表ノード r と同じ連結成分に属するようになる，すなわち，可到達な関係になるステップ数を h とすると，第 j 回目のシミュレーションにおけるノード v と代表ノード r の連結度を $f(v,r)^j = 1-h/H$ と定義する．シミュレーションにおいて，より早いステップ h で連結関係になる代表ノードとの連結度は高くなる．したがって，全 J 回のシミュレーションにおける連結度は $F(v,r) = J^{-1}\sum_{j=1}^{J} f(v,r)^j$ となる．そして，すべてのノードについて，連結度の値が最も高い代表ノードのクラスタに分割する：

$$\mathcal{V}^{(k)} = \{v\in\mathcal{V}; r_k = \arg\max_{r\in\mathcal{R}} F(v,r)\}.$$

4. 評価実験

評価実験では，次節で説明する実データを用いて，計算精度，計算速度，頑健性の観点から提案手法を評価する．

4.1 データセット

1 つ目は，日本国内にある 1,909 の市区町村役場の座標（緯度経度）データである．市役所と区役所が同一の建物にある場合は一方を取り除いてある．座標間の距離としてユークリッド距離を採用する．本稿では，役場データと呼ぶ．$k=12$ で k-NN グラフが単連結となったため，12-NN グラフを対象とする．

2 つ目は，手書き文字認識用データベースに含まれるデータを 5,000 個抽出したものである．10 クラス（0, 1, 2, 3, 4, 5, 6, 7, 8, 9）の数字の手書き文字データであり，各文字は 28 × 28=784 画素で，各画素値は 0〜255 の 256 階調グレースケールで表されている．各手書き文字を，画素値を要素とする 784 次元ベクトルで表現し，文字間の類似度として，頻繁に用いられる相関係数 [11] を採用する．文字 i,j 間の相関係数 $r(i,j)$ を以下のように距離に変換し，オブジェクト間のメトリックとした：

$$d(i,j) = \sqrt{2(1-r(i,j))}.$$

距離の公理（三角不等式）を満たすように，一般的に用いられる平方根をとった距離を採用する．本稿では，文字データと呼ぶ．$k=3$ で k-NN グラフが単連結となったため，3-NN グラフを対象とする．文字データの 3-NN グラフの無向リンク数は 11,531 であった．

3 つ目は，全国の鉄道駅名をローマ字表記した単語データであり，5,000 駅を抽出したものである．駅名間の距離として，代表的な編集距離であるレーベンシュタイン距離 [12] を用いる．これは，二つの文字列が文字（character）の挿入・削除・置換で同一になる場合の最小の操作回数をコストとする．ただし，そのままでは長い駅名は必然的に距離が大きくなるため，距離を測定する長い方の駅名の長さで除して正規化する．駅名 i,j 間のレーベンシュタイン距離を $L(i,j)$，駅名 i の単語長を $\mathrm{length}(i)$ とすると，

$$d(i,j) = \frac{L(i,j)}{\max\{\mathrm{length}(i), \mathrm{length}(j)\}}$$

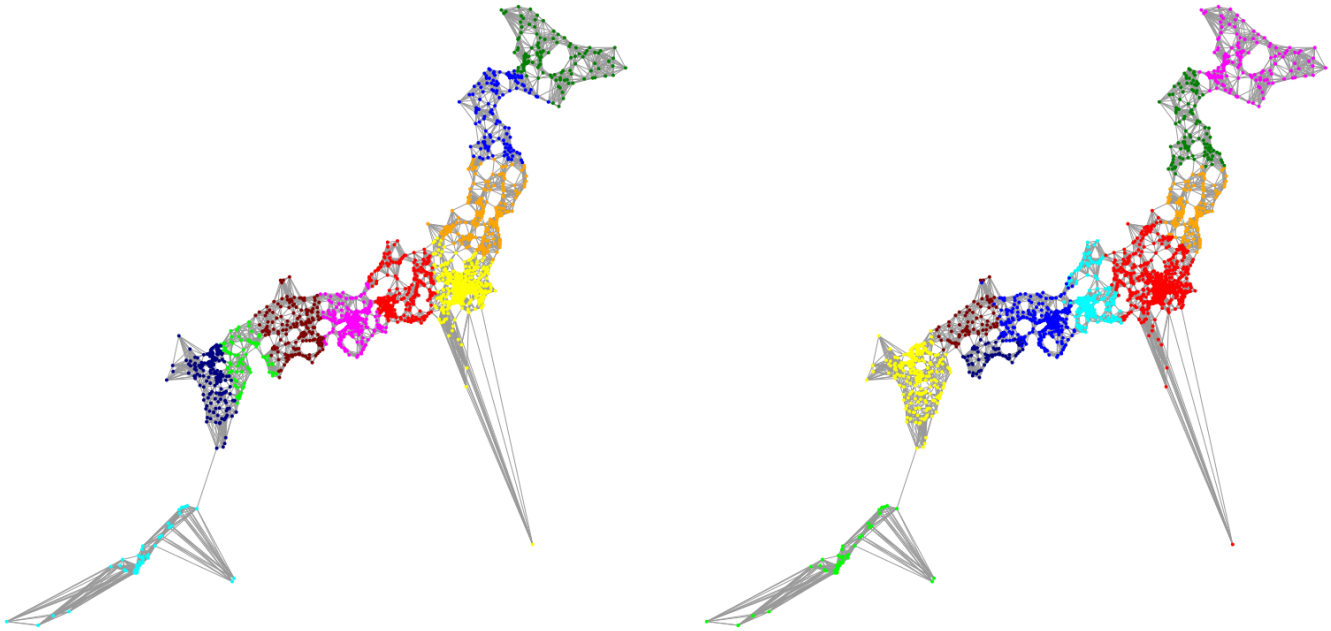
を正規化編集距離と呼び，オブジェクト間のメトリックとした．本稿では，駅名データと呼ぶ．$k=3$ で k-NN グラフが単連結となったため，3-NN グラフを対象とする．駅名データの 3-NN グラフの無向リンク数は 10,899 であった．

4 つ目は，写真共有サイト Flickr に投稿された写真データであり，東京都で撮影された写真のうち 5,000 枚を抽出したものである．抽出した写真データを MPEG-7 の Color Structure ディスクリプタと呼ばれる特徴量で数値化した．写真間の距離として，MPEG-7 で推奨されるマンハッタン距離を採用する．本稿では，写真データと呼ぶ．$k=4$ で k-NN グラフが単連結となったため，4-NN グラフを対象とする．写真データの 4-NN グラフの無向リンク数は 15,320 であった．

5 つ目は，毎日新聞国際面の新聞記事データであり，データベースから古い順に 5,000 記事を抽出した．含まれる単語数は 23,135 であり，記事間の類似度としてベクトル空間モデル [13] で頻繁に用いられる単語頻度ベクトル（Bag Of Words）間のコサイン類似度を採用する．記事 i,j 間のコサイン類似度 $s(i,j)$ を以下のように距離に変換し，オブジェクト間のメトリックとした：

$$d(i,j) = \sqrt{2(1-s(i,j))}.$$

距離の公理（三角不等式）を満たすように，一般的に用いられる平方根をとった距離を採用する．本稿では，新聞データと呼ぶ．$k=4$ で k-NN グラフが単連結となったため，4-NN グラフを対象とする．新聞データの 4-NN グラフの無向リンク数は 15,446 であった．



(a) 比較手法：集合近接中心性　　(b) 提案手法：集合連結中心性

図 1　役場データのクラスタリング結果 ($\kappa = 10$)

4.2　比較手法

上述した種々の距離重みが付されたリンクに対して，ダイクストラ法を用いて全オブジェクト (u, v) 間の測地距離 $g(u, v)$ を求める．オブジェクト間の測地距離 $g(u, v)$ に対して，集合近接中心性 [10] により代表ノードを抽出し，他のノードを最も近い代表ノードのクラスタに分割する．すなわち，測地距離に基づき k-medoids クラスタリングしているのと等価である．

5.　実験結果

提案手法と比較手法によるクラスタリング結果について，複数の観点から評価する．今回の実験では，集合連結中心性のシミュレーション回数は $J = 100$ とした．

5.1　定性評価

図 3.2 に，クラスタ数 $\kappa = 10$ とした際の役場データのクラスタリング結果を示す．図中の色はクラスタを表している．図 3.2(a) と (b) の結果を比較すると，福岡と山口や中国地方と四国地方が距離的に近いため同一クラスタとなっているのに対し，提案手法の集合連結中心性によるクラスタリング結果では概ね地方ごとに分かれている．提案手法では，距離ではなく連結性に着目しているため，九州と本州，中国地方と四国地方のように連結度の低い部分は別クラスタとなる．

5.2　計算精度

文字データ（MNIST）は，各手書き文字がどの数字を表しているのかラベルが付されている．これを正解ラベルとし，クラスタリング結果が正解ラベルとどの程度一致しているのかを正規化相互情報量（NMI：Normalized Mutual Information）により評価する．集合近接中心性によるクラスタリング結果と正解ラベルの NMI は 0.56，集合連結中心性の結果と正解ラベルの NMI は 0.72 であり，提案手法の方が高いクラスタリング精度が得られた．定性評価と NMI の評価結果から，提案手法は精度よくクラスタリングできると言える．

5.3　頑健性

データの欠損に対する頑健性について評価する．具体的には，全 N オブジェクトから一定の割合 $rate$ のオブジェクトをランダムに選び，k-NN グラフから削除する．この k-NN グラフは，場合によっては非連結なグラフとなる．この k-NN グラフから集合連結中心性により代表ノードを抽出，全ノードを連結度に基づき κ 分割する．そして，残った $M = N \times rate$ オブジェクトに関して，NMI により計算するクラスタリング結果の類似性で評価する．NMI が高いほど，削除前のクラスタリング結果に近い結果が得られたことになり，データ欠損に対する頑健性があるといえる．

図 2 に，横軸に削除比率 $rate$，縦軸に NMI の値をプロットした．削除オブジェクトのランダム抽出は 5 回行い，NMI 値の平均値をプロットした．図 2 を見ると，いずれのデータセットでも高い NMI 値が得られることが確認できる．1 割くらいのデータが欠損しても（$rate = 0.1$），

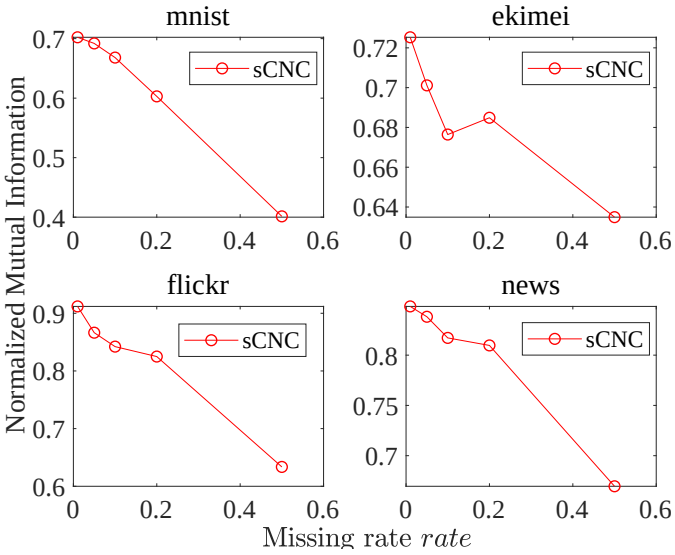


図 2　頑健性

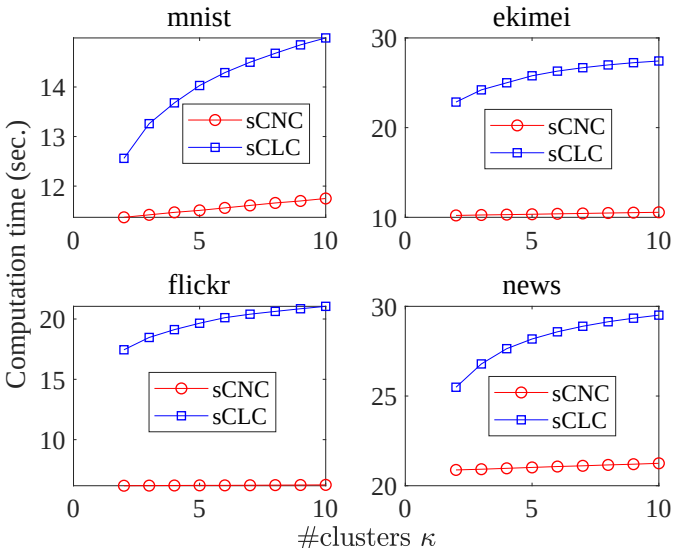


図 3　計算時間

7,8 割の精度が得られることから，サンプリングによるクラスタリングの高速化への応用も期待できる．

5.4　計算速度

図 3 に，集合近接中心性（sCLC）と集合連結中心性（sCNC）によるクラスタリングの計算時間をプロットした．横軸はクラスタ数 κ，縦軸は計算に要した実時間を秒単位でプロットした．ただし，k-NN グラフ構築にかかる時間は含まれていない．図 3 を見ると，いずれのデータセットにおいても，赤線の提案手法の方が青線の比較手法より 2 倍以上早いことがわかる．特に，提案手法はクラスタ数 κ が増えても計算時間に大きな増加は見られないため，両手法の差は大きくなる傾向にある．このことから，提案手法は頑健で高速なクラスタリング手法であると言える．

6.　おわりに

本研究では，多様体上に分布するオブジェクト群のクラスタリングを目的に，計算速度と頑健性の観点から優れている手法を提案した．実データを用いた評価実験により，提案手法の有用性について，精度，頑健性，計算速度の観点で定性的，定量的に確認した．今後は，相対近傍グラフ（RNG）や非連結な状態の k-NN グラフでの集合連結中心性によるクラスタ分割の有効性について評価していく．

謝辞　本研究は，JSPS 科研費（No.17H01826, No.16K16154）の助成を受けたものである．

参考文献

[1] Torgerson, W.: Multidimensional scaling: I. Theory and method, *Psychometrika*, Vol. 17, pp. 401–419 (1952).
[2] Sammon, J. W.: A Nonlinear Mapping for Data Structure Analysis, *IEEE Trans. Comput.*, Vol. 18, No. 5, pp. 401–409 (1969).
[3] Lee, J. A. and Verleysen, M.: *Nonlinear dimensionality reduction*, Springer, New York; London (2007).
[4] von Luxburg, U.: A tutorial on spectral clustering, *Statistics and Computing*, Vol. 17, No. 4, pp. 395–416 (2007).
[5] Park, H.-S. and Jun, C.-H.: A simple and fast algorithm for K-medoids clustering, *Expert Systems with Applications*, Vol. 36, No. 2, Part 2, pp. 3336 – 3341 (2009).
[6] Tenenbaum, J. B., Silva, V. and Langford, J. C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science*, Vol. 290, No. 5500, pp. 2319–2323 (2000).
[7] Belkin, M. and Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, *Neural Computation*, Vol. 15, No. 6, pp. 1373–1396 (2003).
[8] 伏見卓恭，斉藤和巳，風間一洋：メトリック空間における複数カテゴリに属するハイブリッドオブジェクト抽出法の提案，情報処理学会論文誌，Vol. 58, No. 6, pp. 1258–1267 (2017).
[9] 伏見卓恭，斉藤和巳，池田哲夫，風間一洋：リンク切断に頑健な連結中心性とその高速計算法，情報処理学会論文誌数理モデル化と応用（TOM），Vol. 11, No. 2, pp. 1–11 (2018).
[10] 伏見卓恭，斉藤和巳，池田哲夫，武藤伸明：ノード群の協調的振舞いに着目した集合媒介中心性の提案と応用，電子情報通信学会和文論文誌 D，Vol. J96-D, No. 5, pp. 1158–1165 (2013-05).
[11] Seewald, A. K.: On the brittleness of handwritten digit recognition models, *ISRN Machine Vision*, Vol. 2012 (2011).
[12] Levenshtein, V.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals, *Soviet Physics Doklady*, Vol. 10, p. 707 (1966).
[13] Salton, G., Wong, A. and Yang, C. S.: A Vector Space Model for Automatic Indexing, *Commun. ACM*, Vol. 18, No. 11, pp. 613–620 (1975).