A-015

移動体データからの密集エリアと共通移動軌跡の同時クラスリング

Clustering of Density Areas and Trajectories simultaneously from Moving Object Data

史　旭† Xu Shi　豊田 真智子‡ Machiko Toyoda

1. はじめに

スマートフォンや GPS デバイスの普及に伴い，地理的，時間的な要素を含んだ人や車などの移動体の位置情報(以降，移動体データと呼ぶ)が高精度，低コストで収集可能になった．移動体データは，これまで把握が難しかった人々の詳細な活動を記録している．これらのデータから検出された人の行動習慣やパターンが小売・流通，観光，物流，交通，防犯・防災など様々な領域への活用が期待され，注目を集めている．

その中で，移動体データに対する捉え方の違いによって，分析の仕方や手法が主に 2 つに分けられる．密集エリアに関する分析と共通移動軌跡に関する分析である．密集エリアに関する分析は，人々が集まった場所の検出，及びそれら場所間の訪問順序の発見を目的として，詳細な軌跡を考慮することなく，その場所への流入・流出ルートを知ることができない．一方，共通移動軌跡に関する研究は， 人々が共通する移動ルートの発見を目的として，移動ルート上の集まる場所を知ることができない．このように密集エリアと共通移動軌跡を別々に検討する場合，人々はどの場所で集まったか，どのルートから集まったかといった移動傾向の全体把握や，なぜ密集しているのか，なぜその方向に移動するのかといった経路選択の意図や渋滞の原因推定が困難である．

以上の問題を解決するために，別々に行われる密集エリアと共通移動軌跡に関する分析を合わせることが望ましい．しかし，密集エリアと共通移動軌跡はそれぞれ異なる距離尺度を用いて検出していくため，単にそれぞれから得られた結果を組み合わるのみでは， 期待した結果が得られるとは限らない．例えば，東京駅を中心とした移動ルートの関係を検出したい場合を考える．様々な方面から人々が集まるため，東京駅内の密度は非常に高いが，流入ルートが複数存在するため，各ルート上の密度は分散される．密集エリアと共通移動軌跡を同じ粒度で組み合わせる場合，密度が高い東京駅の粒度に合わす時，周辺ルートの密度が低いため，共通移動軌跡として検出されない恐れがある．一方，周辺ルートの粒度に合わす時，密集エリアは東京駅だけではなく，周辺ルートも混ざって 1 つぼやけた意味のないものとして検出されてしまう可能性がある．

そこで，本研究ではデータ領域を複数の区域に分割し，区域毎のデータの密度を考慮してクラスタリングを行う手法を提案する．提案手法は，大きく 2 つのステップから構成される．(1)分割した区域に対して密集エリアまたは共通移動軌跡のどちらのクラスタを生成するのかを決定する．(2)どのように区域を分割すると最適なクラスタが得られるのかの区域割り当てを決定する．具体的に，(1)では密集エリアと共通移動の特徴を位置座標間の距離と角度の違いとして捉えた上で，同一指標で 2 種類のクラスタを生成するための距離関数を定義し，密集エリアと共通移動軌跡を密度ベースクラスタリング手法より同時に検出する．また，個々の区域におけるデータ点の特徴と密度度合に合わせ，区域単位で異なる密度閾値を用いて密集エリアクラスタ或いは共通移動軌跡クラスタのいずれを生成する．(2)では最小記述長(MDL: minimum description length)を導入し，MDL コスト最小化に基づいた適切な区域分割の数とそれぞれ区域の範囲を自動的に推定する．

本研究で提案する手法により検出される集まる場所とそこへの流入・流出を示すルートのような，お互いの関係性を示す移動パターンは移動傾向の全体把握及び渋滞や移動経路選択背後にある原因の推定を容易にすることができる．そして，従来取り扱わなかった移動パターンの発見により，新たなビジネスチャンスにつなげることが期待できる．

†‡日本電信電話株式会社

2. 関連研究

本章では，移動体データ分析に用いられた分析技術について述べる．

(1) 密集エリア分析に関する研究：ユーザの移動軌跡上の位置座標を分析対象とする技術.

文献[7][8][9]において密集エリアの検出に用いられる DBSCAN[1]は密度ベースクラスタリングの手法として，位置座標間のユークリッド距離を用いて近傍点を定義し，指定された範囲における近傍点の数がある閾値を越えている限り，それらの点を 1 つの密集エリアクラスタとして成長させ続ける．文献[4]では，密集エリアを指定される時間範囲以内に多くのオブジェクトが経由した長さlの正方形として定義し，各時刻のオブジェクトの軌跡を構成する位置座標から，自分と正方形距離範囲内，且つ時間範囲内における他のオブジェクトの軌跡の位置座標の数を順番に探索することで，その数が閾値を超えた正方形を検出する．

(2) 共通移動軌検出に関する研究：個々のユーザの移動軌跡を分析対象とする技術.

文献[5]で提案されている TCMRM は，EM アルゴリズムを用いて移動軌跡をクラスタリングする手法である．各移動軌跡を重回帰混合正規分布モデルに当てはめ，同じモデルに含まれる軌跡を 1 つの共通移動軌跡クラスタとして検出する．文献[6]で提案されている TRACLUS は，ユーザ毎の移動軌跡を短いラインセグメントに分割し，ラインセグメント間の方向性や角度を取り入れた線間の距離に基づいて，移動方向が一致するラインセグメントを共通移動軌跡クラスタとして形成していく．

(3)パターン検出に関する研究：密集エリア間の関係を分析対象とする技術.

T-Pattern[7]は，頻出シーケンシャルパターンマイニングアルゴリズムを用いて，与えられた閾値以上現れる一連のシーケンスを，密集エリア間のシーケンシャルパターンとして検出する．ここでの シーケンシャルパターンは，例えば，「密集エリア A → 密集エリア B → 密集エリア C」の

ような，密集エリア間の訪問順序を表すものである．T-Pattern のほか，Swarm[8]，Gathering[9] などの手法も提案されている．

これら既存研究は，密集エリア或いは共通移動軌跡のいずれに着目したマイニング技術であり，本研究で解決したい密集エリアと共通移動軌跡の関係性を示す移動パターンの検出課題に対応できない．

3. 密集エリア及び共通移動軌跡クラスタの生成

本章では，異なる距離尺度で検出される密集エリア及び共通移動軌跡を同時に扱うための距離関数とクラスタリング手法を提案する．また，異なる密度閾値で密集エリア及び共通移動軌跡を検出するための手法についても述べる．

本研究では，移動体データをユーザ毎に取得したデータ点 p_iの位置座標(経度，緯度)と移動方向$\overrightarrow{p_ip_{t+1}}$の系列$\{p_1, p_2 \dots p_n\}$として定義した上で，密集エリアと共通移動軌跡の特徴について次のような考察を行った．

- 密集エリア： 多くの人が集まる場所であり，位置座標間の距離が近い．また，駅やショッピングセンタのように，人々が様々な方向に移動するため，移動方向の交差が多い．
- 共通移動軌跡：多くの人が利用するルートであり，位置座標間の距離が近い．また，人々が同じ方向に移動をすることから，移動方向が一致する．

以上の考察から，密集エリアと共通移動軌跡は「位置座標の距離」と「移動方向の角度」の 2 つの指標で捉えられること，また，密集エリアと共通移動軌跡は「移動方向の角度」の違いによって判別できると考えられる．つまり，密集エリアを移動方向が交差し，近い距離の位置座標を持つデータ点の塊(クラスタ)として，共通移動軌跡を移動方向が一致し，近い距離の位置座標を持つデータ点の塊(クラスタ)としてみなすことができる．そこで，本研究では 2 種類のクラスタを検出するための 2 点間の距離関数を次のように定義する．

定義 1: 密集エリア距離. データ点p_i, p_jが与えられた時，p_i, p_j間の密集エリア距離は以下の式で計算される．

$$dist_D(p_i, p_j) = exp(-ED(p_i, p_j)/\delta) \cdot sin\theta \quad (1)$$

ここで，$ED(p_i, p_j)$ はp_i, p_j位置座標間のユークリッド距離，δ はスケール因子，θは$\overrightarrow{p_ip_{i+1}}$, $\overrightarrow{p_jp_{j+1}}$によって生成される角度を示す．θの範囲は方向の交差を表現するため $0 \le \theta \le \pi$とする．

定義 2: 共通移動軌跡距離．データ点p_i, p_jが与えられた時，p_i, p_j間の共通移動軌跡距離は以下の式で計算される．

$$dist_T(p_i, p_j) = exp(-ED(p_i, p_j)/\delta) \cdot cos\theta \quad (2)$$

θの範囲は同じ方向の移動を表現するため $0 \le \theta \le \pi/2$とする．

以上定義した 2 つの距離関数において，密集エリア距離関数の場合，2 つデータ点の位置座標間のユークリッド距離が 0 に近い，角度が$\pi/2$ に近いほど，近傍で直交する方向に移動することを表し，値が最大値の 1 に近くなる．一方，共通移動軌跡距離関数の場合，2 つデータ点の位置座標間のユークリッド距離が 0 に近い，角度の差が 0 に近いほど，近傍で同じ方向に移動することを表し，値が最大値の 1 になる．定義 1 と 2 より各データ点間の距離を定義することができた．次にこれらの距離を使って，どのようにクラスタを生成するのかについて述べる．

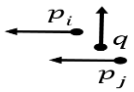

図 1　間接到達可能例

定義 3: 直接到達可能. データ点p_i, p_j，及び，距離の閾値Min_{dsit} が与えられた時，$dist_D(p_i, p_j) \ge Min_{dist}$ または$dist_T(p_i, p_j) \ge Min_{dist}$が成り立つなら，$p_i$と$p_j$ は直接到達可能であり，$p_i \ominus p_j$ として表す．

定義 4: 密度. データ点p_iが与えられた時，p_iと直接到達可能を満たすデータ点の数を密度と定義する．

定義 5: 間接到達可能. データ点$p_i, q_1, q_2 \dots q_n, p_j$について，$p_i \ominus q_1$，$q_i \ominus q_{i+1} (1 \le i \le n)$，$q_n \ominus p_j$ が成立する時，p_iとp_jは間接到達可能であり，$p_i \oslash p_j$として表す．

間接到達可能を満たすデータ点は，$p_i \oslash p_j \wedge p_j \oslash p_r \rightarrow p_i \oslash p_r$の性質を持つ．図 1 における例を考える．データ点 p_i と p_j が同じ移動方向であるため，$dist_D(p_i, p_j)$は最小となり，直接到達可能を満たさないが，データ点qとp_i, p_jの角度がそれぞれ$\pi/2$であるため，$p_i \oslash q \wedge q \oslash p_j$ が成立する時，間接到達可能の定義により$p_i \oslash p_j$が成立する．

定義 6: シード. データ点s，密度閾値$Min_{density}$が与えられた時，sが直接到達可能なデータ点の数は $Min_{density}$を満たす場合，sをシードと呼ぶ．

sから直接到達可能なデータ点を探索する場合，密集エリア距離関数及び共通移動軌跡距離関数における角度の差を評価する三角関数を最大である 1 に指定した時，sが直接到達可能なデータ点との最大ユークリッド距離Max_{ED}が$exp(-ED(s,q)/\delta) = Min_{density}$より計算され，$Max_{ED} = \delta \cdot \sqrt{-ln(Min_{density})}$ となる．つまり，s とのユークリッド距離がMax_{ED} より遠いデータ点は必ず s の直接到達可能の点ではない．従って，s と距離がMax_{ED}以上のデータ点の探索は必要がないことが分かる．

以上の定義から，本研究ではクラスタをシードとなるデータ点の集合，及びそれぞれのシードと直接到達可能なデータ点の集合として以下のように定義する．

定義 7: クラスタ. 最小クラスタサイズ|c| (クラスタに含まれるデータ点の数)の閾値Min_{pts}，Min_{dist}，$Min_{density}$が与えられた時，以下 3 つの条件を満たすデータ点の集合をクラスタcとする．

(A) シード$s \in c, s \oslash q \rightarrow q \in c$
(B) 任意の 2 つのデータ点 $p_i, p_j \in c \rightarrow p_i \oslash p_j$
(C) $|c| \ge Min_{pts}$

密集エリア距離(定義 1)を用いて生成されるクラスタを**密集エリアクラスタ**，共通移動軌跡距離(定義 2)を用いて生成さるクラスタを**共通移動軌跡クラスタ**とする．

クラスタ生成アルゴリズムは，文献[1]で提案された密度ベースのクラスタリングアルゴリズムをベースとしたもので，以下の 4 つのステップからなる．

・ステップ 1 データ点 p に対して，pがシードであるかどうかを判断する．pがシードである場合，p 及び p と直接到達可能な点をS に格納する．pがシードではな

い場合，他のデータ点を選び，シードを見つかるまで繰り返す．

・**ステップ 2** Sに含まれるp 以外のデータ点を対象に，シードであるかどうかを判断し，それらのデータ点と直接到達可能な点をS に追加する． S に含まれるすべてのデータ点に対して判断を行うまで繰り返す．

・**ステップ 3** $|S| \geq Min_{pts}$の時， Sをクラスタc_iとする．

・**ステップ 4** 現時点で生成されたクラスタ $\{c_1 \dots c_i\}$ のいずれに含まれないデータ点に対し，ステップ 1-3 を実行し，すべてのデータ点に対する処理が終了するまで繰り返す．最終的にクラスタに含まれないデータ点は***ノイズ***として扱われる．

上記に示したように，密集エリアと共通移動軌跡の 2 種類のクラスタを同一尺度で生成することができるようになった．次に検討すべきは，どうやって粒度の異なるクラスタを検出するのかである．この問題に対応するため，本研究ではデータ領域を複数の区域に分割し，区域単位でのクラスタリングを行う．すなわち，それぞれの区域内のデータ点の密度度合に合わせ，区域毎に異なる閾値を設けて，クラスタを生成する．

定義 8: セル. 与えられたデータ領域に対し，経度と緯度から構成されるメッシュによって表される最小の領域として定義される．

定義 9: 区域. 複数の隣接セルから構成される 1 つの領域として定義される．

区域が与えられた時，密集エリアまたは共通移動軌跡距離のどちらでクラスタを生成するのかを決定する必要がある．ここで，3 章で考察した密集エリア及び共通移動軌跡の性質について考える．密集エリアは移動方向の交差が多いことから，区域内の 2 つのデータ点p_i, p_jの角度θは

$$\theta_D\colon\ \pi/4 \leq \theta \leq 3\pi/4$$

を満たすものが多くなるべきである．一方，共通移動軌跡は移動方向が一致することから，区域内の 2 つのデータ点p_i, p_jの角度θは

$$\theta_T\colon\ 0 \leq \theta < \pi/4 \vee 3\pi/4 < \theta \leq \pi$$

を満たすものが多くなる．区間$0 \leq \theta < \pi/4$は移動方向が一致することを表し，区間$3\pi/4 < \theta \leq \pi$は移動方向が正反対であることを表す．

以上の考察から，区域内のすべてのデータ点について，他のデータ点との角度θをペアワイズに計算する時の総計算数をn_θ，θ_D 及び θ_T を満たすペアワイズの数をそれぞれn_{θ_D}，n_{θ_T}とする時，密集エリア区域及び共通移動軌跡区域を次のように定義する．

定義 10: 密集エリア区域.

$$n_{\theta_D}/n_\theta \geq n_{\theta_T}/n_\theta$$

が成立するなら，その区域を密集エリア区域として定義する．

定義 11: 共通移動軌跡区域.

$$n_{\theta_T}/n_\theta > n_{\theta_D}/n_\theta$$

が成立するなら，その区域を共通移動軌跡区域として定義する．

定義 10 と 11 により，データ領域を分割した区域が与えられた時，それらの区域内のデータ点を密集エリアと共通移動軌跡のどちらのクラスタとして生成するかを判断できるようになる．与えられた区域は密集エリア区域と判断されると，密集エリア距離関数と指定する密度閾値を用いて，密集エリアクラスタを生成し，共通移動軌跡区域と判断されると，共通移動軌跡距離関数と指定する密度閾値を用いて，共通移動軌跡クラスタを生成する．

区域毎の密度閾値は，各区域の大きいさを考慮したデータ点の密度度合に合わせるため，δを区域の長さと幅の和，Min_{dist}を区域内すべてのデータ点が自分との距離の上位$Min_{pts}/2$個の平均値の平均，$Min_{density}$を区域内すべてのデータ点の密度の平均として設定することが考えられる．

4. 区域割り当ての最適化

前章説明したように，提案手法は与えられた区域に対してクラスタを生成していく．ではどうやって最適な区域を決定するのか？この問題に対応するため，モデル選択基準の MDL を導入し，MDL コスト最小にするような区域割り当てを選択するアプローチをとる．次節以降で詳細に説明する．

4.1 データ圧縮

最小記述長((MDL: minimum description length)は情報理論に基づくデータの圧縮を目的とするモデル選択基準であり，パラメータの選択とその選択の下でのデータの適合度の両面を考慮した評価関数が最小となるモデルを最適なものとして選択する．MDL は 2 つのコスト関数の組み合わせにより，以下の式で表現される．

$$Cost(D,H) = Cost(H) + Cost(D|H) \qquad (3)$$

$Cost(H)$は圧縮モデルHを表現するコストであり，圧縮モデルが単純なほど$Cost(H)$ は小さい．一方，$Cost(D|H)$ は圧縮の質，すなわち，モデルH により，データDを表現するためのコストを表し，モデルとデータの適合度が高いほど，$Cost(D|H)$ は小さくなる．MDL はこの 2 つのコストの総計が一番小さなものを選び，少ない符号でデータを表現する最適なデータ圧縮を達成する．

本研究では，分割されたそれぞれの区域をその区域内のデータ点を圧縮するためのモデルとして捉える．各区域にはクラスタが生成され，クラスタに含まれないデータ点はノイズとなるが，ノイズが少ないほうが望ましい．そこで，モデルである区域が与えられた時，その区域に含まれるデータ点がノイズか否かを$Cost(D|H)$ で表現する．2 つのコストを最小化することで，適切な区域分割の数と区域の範囲を自動的に決定し，各区域内に生成された密集エリアクラスタ及び共通移動軌跡クラスタが得られる．以上より，与えられたデータ点を適切に区域に分割するために，2 つの MDL コスト関数$Cost(H)$，$Cost(D|H)$を設計する必要がある．

4.2 MDL コスト関数

本研究では，定義 8 のセルに基づいてデータ領域を分割するため，生成された個々の区域を表現するコスト関数

$Cost(H)$は，文献[7]の理論より，以下の要素から構成される．

- 生成された区域の個数 n：$log^*(n)$ ビット
- 各区域 n_i におけるセルの数 n_c：$\sum_{i=1}^{n} log^*(n_c)$ビット
- 各区域 n_i におけるデータ点の数n_p：$\sum_{i=1}^{n} log^*(n_p)$ ビット

ここで，$log^*(x)$は整数のユニバーサル符号長を表し，$log^*(x) = log_2(2.865) + log_2(x) + log_2 log_2(x) + \cdots$ として計算される[8].

次に，各区域のクラスタの集合を$C = \{c_1, c_2 \ldots c_m\}$，ノイズ集合を$K$とする時，データ点の表現コスト$Cost(D|H)$をエントロピーにより表す．

エントロピーとは，データ集合に含まれるすべての事象の生起確率を考慮したデータを表現するのに必要な最低限の符号長を示す指標であり，以下の式で表現される．

$$E = -\sum q_i log_2 q_i \qquad (4)$$

ここで，q_i は事象 i の生起確率を示す．データ集合に表した事象の数が多いほど，また，各事象の生起確率が均等であるほど，データを表現するための最低限の符号長が長くなり，エントロピーが大きくなる．一方，データ集合に含まれる事象の数が少ないほど，また，各事象の生起確率に偏りがあるほど，データを表現するための最低限の符号長が短くなり，エントロピーが小さくなる．

エントロピーの概念に基づいて，分割された各区域に含まれる個々のクラスタと個々のノイズを単独の事象とみなす時，ある区域内のデータ点p_iを表現するための最低限の符号長E_{p_i}は以下の式で表される．

$$E_{p_i} = -\sum_{i=1}^{m} \frac{|c_i|}{n_p} log_2 \frac{|c_i|}{n_p} - \sum_{i=1}^{k} \frac{1}{n_p} log_2 \frac{1}{n_p} \qquad (5)$$

ここで，$|c_i|$ はクラスタc_i に含まれるデータ点の数，m はクラスタの数，kはノイズの数とする．区域内に生成されるクラスタの数，また，ノイズとして判断されるデータ点の数が少ないほど，E_{p_i} の値は小さくなる．これは，各区域に含まれるデータ点をノイズの少ない単独のクラスタで表すことに寄与する．そして，ある区域 n_i 内のすべてのデータ点を表現する符号長$E_{n_i} = n_p \cdot E_{p_i}$で計算されるため，分割された区域の個数を$n$とする時，データ領域におけるすべてのデータ点を表現するための符号長を$Cost(D|H)$として，$Cost(D|H) = \sum_{i=1}^{n} E_{n_i}$ になる．

各区域内のすべてのデータ点を表現するコストE_{n_i} はその区域に対する再分割の必要性を知る目安になる．E_{n_i}の大きい区域は，その区域に属すデータ点を 1 つの区域に割り当てることは適切ではないことを意味する．区域に対する再分割を促すことによって，より適切なクラスタリング結果を求める．

以上まとめると，候補解$Z = \{\{C, k\} | \{n, \{n_c\}, \{n_p\}\}$ が与えられた時の MDL コスト関数は次のように表現される．

$$Cost(D,H) = log^*(n) + \sum_{i=1}^{n} log^*(n_c) + \sum_{i=1}^{n} log^*(n_p) + \sum_{i=1}^{n} n_p \cdot \left(-\sum_{i=1}^{m} \frac{|c_i|}{n_p} log_2 \frac{|c_i|}{n_p} - \sum_{i=1}^{k} \frac{1}{n_p} log_2 \frac{1}{n_p}\right) \qquad (6)$$

式(6)を用いて，ある区域分割の下で生成されるクラスタの良し悪しを評価することが可能となる．すなわち，式(6)を最小とする結果を求めることで，最適なクラスタを得る．

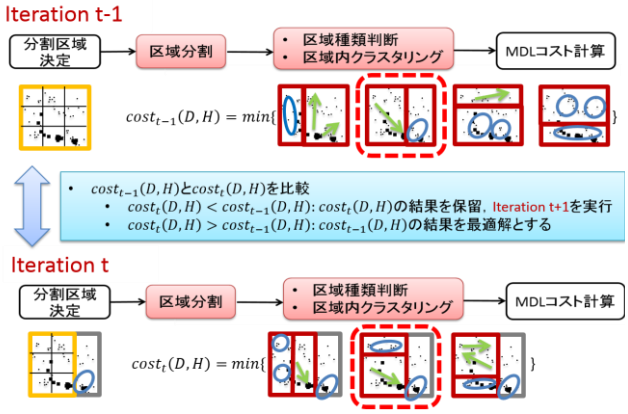

図 2　最適化アルゴリズムの流れ

4.3 最適化アルゴリズム

本節では，式(6)に基づき，最適な区域割り当て解を発見するためのアルゴリズムを提案する．

図 2 に提案するアルゴリズムの流れを示す．移動体データが与えられた時，データ領域に対する分割と区域内のクラスタリングを繰り返し実行することで MDL コストが最小となるクラスタリング結果を決定する．具体的には，データ領域をセルサイズに基づいて，メッシュ化した後，4 つの処理をイテレーション毎に実行する．現イテレーションと 1 つ前のイテレーションで生成されるローカル解を比較し，コストが増加した場合に最適解が得られたとして処理を終了する．

・ステップ 1(分割区域の決定) 分析対象となるデータ領域(図 2 におけるオレンジの四角)を決定する．最初のイテレーションは全データ領域を分割対象とする．

・ステップ 2(区域分割) データ領域に対して，経度と緯度をそれぞれ分割線として，区域割り当てのバリエーションを生成する．図 2 の Iteration t-1 には，2 つの縦線と 2 つの横線に従い，4 つの区域割り当てバリエーションが生成された例を表す．

・ステップ 3(区域種類判断＆区域内クラスタリング) ステップ 2 で生成された各区域割り当に対し，定義 10 と 11 を用いて個々の区域が密集エリア区域または共通移動軌跡区域のいずれに判断する．それぞれの区域に対して指定する密度閾値を用い，3 章で提案したクラスタリングアルゴリズムを実行する．図 2 では生成された密集エリアクラスタを青い丸，共通移動軌跡クラスタを緑の矢印として示している．

・ステップ 4(MDL コスト計算) 式(6)を用いて，ステップ 3 で生成されたクラスタの MDL コストを計算する．各イテレーションにおいて，最小 MDL コストを持つ結果を現イテレーションのローカル解(図 2 における赤の点線)として保存する．

・ステップ 5(ローカル解の比較) t 回目のイテレーションと t-1 回目のイテレーションで選ばれたローカル解の MDL コストを比較する．t 回目のイテレーションのコストが t-1 回目のイテレーションのコストより小さい

図 3　東大データセット 1 プロット

図 5　東大データセット 1 実験結果

図 4　東大データセット 2 プロット

図 6　東大データセット 2 実験結果

場合，t 回目のイテレーションの結果を保存し，t+1 回目のイテレーションを実行する(ステップ 6)．t 回目のイテレーションのコストが t-1 回目のイテレーションのコストより大きい場合，t-1 回目のイテレーションの結果を最適解として処理を終了する．

・ステップ 6(分割区域の決定)　ステップ 5 で t+1 回目のイテレーションの実行が決まった時，t 回目イテレーションのローカル解について，コストE_{n_i}が大きい区域を次の分割対象区域として選び，ステップ 1-5 の手順に従い，t + 1 回目のイテレーションを実行する．図 2 の例では，Iteration t-1 のローカル解の結果について，左側区域のコスト E_{n_i} は右側区域より大きいため，Iteration t で再分割対象として選択される．

5. 評価実験

本研究では，提案手法の有効性を検証するため，東大人の流れプロジェクトの人流データ[11] を用いた実験を行った．(1) 移動方向の違いを考慮した密集エリア及び共通移動軌跡クラスタを検出できたか，(2)地図とマッピングし，ランドマークや道路などを示す意味のあるクラスタを検出できたか，という 2 つの観点から提案手法により検出されたクラスタの精度を考察する．

5.1 実験データ

人の流れプロジェクトの人流データは国や地方自治体等で実施するパーソントリップ調査データから推定した各都市圏周辺の人の流れを表すデータである．データには日別にユーザ毎の経度，緯度，時刻などの属性が集計される．実験用データセットは 2013 年 7 月 1 日のデータから以下 2 つのデータセットをピックアップし，定義した移動体データの形に変換を行った．

・データセット 1

経度 139.770 から 139.776 まで，緯度 35.695 から 35.700 までの範囲における 118 名ユーザの 815 個のデータ点．図 3 にすべてのデータ点の位置座標と移動方向をプロットしたものを示す．

・データセット 2

経度 139.420 から 139.600 まで，緯度 35.700 から 35.710 までの範囲における 234 名ユーザの 3296 個のデータ点．図 4 にすべてのデータ点の位置座標と移動方向をプロットしたものを示す．

5.2 実験結果

2 つデータセットのスケールをわせて，データセット 1 を用いた実験においてはセルサイズを 100m×100m，Min_{pts}を 10 として，データセット 2 を用いた実験においてはセルサイズを 200m×200m，Min_{pts} を 50 として設定した時の実験結果を図 5 と図 6 に示す．ここで，青い四角は密集エリア区域，緑の四角は共通移動軌跡区域と判断された区域であることを表し，青い丸と緑の矢印は密集エリアクラスタと共通移動軌跡クラスタを表す．丸と矢印の線の太さはそのクラスタに含まれるデータ点の数を示し，線が太いほど，そのクラスタに含まれるデータ点の数が多いことを意味する．そして，灰色のノイズを除いて，それぞれのクラスタに含まれるデータ点を色分けで表示している．

図 5，6 から，移動方向の違いにより 2 種類のクラスタが生成されていることが分かる．また，区域内の密度が異なるクラスタを検出できることも確認される．これらのクラスタから，集まる場所や移動の流れを把握することができる．例えば，密集エリアクラスタは駅やランドマーク，共通移動軌跡クラスタは主要道路に該当していた．

一方で，図 5 において，秋葉原駅を含む密集エリア区域には，駅の範囲を大幅に超えた密集エリアクラスタが生成されおり，中には共通移動軌跡クラスタとして抽出するのが望ましいと思われるデータ点が含まれている．また，図 6 においても，飯田橋駅を含む密集エリア区域には，駅の

範囲を大幅に超えた密集エリアクラスタが生成されおり，その中には主要道路を含まれていた．

以上の結果から，区域内の n_{θ_D} と n_{θ_T} がほぼ同数存在する区域については，密集エリアクラスタと共通移動軌跡 クラスタが混在している状態であると推測される．このような区域に対し，再分割が行われること，そして，再分割によって生成される新たな区域において，n_{θ_D} と n_{θ_T} の差が大きくなることが望ましい．各イテレーションにおいて，どの区域を分割するのかはコスト E_{n_i} で判断されるため，図 5 の秋葉原駅を含む区域，および，図 6 の飯田橋駅を含む区域のコスト E_{n_i} の値を調査した．すると，どちらの区域においても E_{n_i} の値が大きくならず，再分割対象区域に選ばれていなかったことがわかった．つまり，ノイズとして生成すべきデータ点はノイズとして検出できなかったことになる．これは，実験時 2 つのパラメータ Min_{dist}，$Min_{density}$ を区域におけるデータ点の平均として設定し，ノイズであるべきデータ点により閾値の値を引きずられた影響であることが考えられる．より精度の高いクラスタリング結果を得るため，それぞれ区域内のデータ点の特性に基づいたクラスタリングパラメータのチューニングへの工夫が必要であることが見えてきた．

6. おわりに

本研究では，密集エリアと共通移動軌跡を 2 種類のクラスタとして同時に扱い，区域毎の密度に合わせて 2 種類のクラスタを検出する手法を提案した．そして，2 つの実データを用いて提案したアルゴリズムの評価を行った．実験結果を地図とマッピングした結果，検出された個々の密集エリアクラスタがランドマーク，共通移動軌跡クラスタが道路とマッチし，集まる場所と移動の流れを同時に知ることができた．

今後の課題としては，精度の向上にあたり，より合理的なクラスタリングパラメータの自動決定方法の検討，そして，より規模の大きいデータセットを用いた性能評価を行う予定である．

参考文献

[1] Martin Ester, Hans-Peter Kriegel, Jorg Sander, XiaoweiXu: “A density-based algorithm for discovering clusters in large spatial databases with noise”, In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp.226-231, 1996.

[2] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, Loren Terveen: “Discovering Personal Gazetteers: An Interactive Clustering Approach,” In Proceedings of ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS), pp. 266-273, 2004.

[3] Daoying Ma and Aidong Zhang: “An Adaptive Density Based Clustering Algorithm for Spatial Database with Noise,” In Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM), pp. 467-470, 2004.

[4] M. Benkert, B. Djordjevic, J. Gudmundsson, and T. Wolle.: “Finding popular places,” In Proceedings of J. Comput. Geometry Appl, 20(1):19-42, 2010.

[5] Scott Gaffney, and Padhraic Smyth: “Trajectory Clustering with Mixtures of Regression Models,” In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 63-72, 1999.

[6] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang: “Trajectory clustering: a partition-and-group framework,” In Proceedings of ACM SIGSPATIAL International Conference on Management of Data (SIGMOD), pp. 593-604, 2007.

[7] Spiros Papadimitriou, Aristides Gionis, Panayiotis Tsaparas: “Parameter-Free Spatial Data Mining Using MDL,” In Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM), 2005.

[8] Jae-Gil Lee, Jiawei Han, Xiaolei Li, Hector Gonzalez: “TraClass: Trajectory Classification Using Hierarchical Region-Based and Trajectory-Based Clustering,” In Proceedings of the Very Large Data Base Endowment (VLDB), pp. 1081-1094, 2008.

[9] Xiao He, Jing Feng, Bettina Konte, Son T.Mai, Claudia Plant: “Relevant Overlapping Subspace Clusters on Categorical Data,” In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 213-222, 2014.

[10] Zaiben Chen, Heng Tao Shen, Xiaofang Zhou: “Discovering Popular Routes from Trajectories,” In Proceedings of the 27th IEEE International Conference on Data Engineering (ICDE), pp. 900-911, 2011.

[11] “東京大学人の流れプロジェクト”：http://bdm.change-jp.com/?p=1705/