1N-06

場所による密度の偏りを考慮した訪問履歴データのインタラクティブ・クラスタリング

渡邉慧汰 *藤田秀之 †大森匡 †新谷隆彦 †

1 背景

地理的な点データ集合として定義される，長期間にわたる個人の場所訪問履歴データを対象に，インタラクティブにクラスタリングを行うインタフェースを提案する．日常に訪れる領域と，そうでない領域では点の密度が異なるため，そこで，効率的に距離パラメータ更新を行う OPTICS クラスタリングを利用し，クラスタリング結果を示す地図と，距離パラメータとクラスタ構造の関係の概観を示す図を提示して，後者を，地図のズームやスクロール操作に応じて適切に更新することで，インタラクティブなパラメータ調整を支援するインタフェースを提案する．

2 インタラクティブ・クラスタリング

インタラクティブ・クラスタリングは，ユーザにとって適切なクラスタリング結果を得るために，ユーザが処理過程に介入するクラスタリング手法である．Bae らは，インタラクティブ・クラスタリングに関する 105 本の論文を対象に，その目的の一部として，クラスタリング過程の理解と興味深いデータの発見を挙げている [1]．クラスタリング過程にユーザを参加させることは，クラスタリング結果の理解や解釈に有用であり，興味深いデータの特定，外れ値や関心領域の発見を支援することができる．

また，同論文では，インタラクションを行う対象や段階として，パラメータへのインタラクションを挙げている．パラメータへのインタラクションでは，結果へのインタラクションと比較し，パラメータ自体を明示的に操作する．

本研究では，ユーザがクラスタリングの過程を理解し，興味深いデータを発見することを支援するため，主にパラメータ調整の過程でインタラクションを活用する手法を提案する．

* 電気通信大学情報理工学域 I 類情報数理工学プログラム 4 年
† 電気通信大学院情報理工学研究科

3 OPTICS

OPTICS(Ordering Points To Identify the Clustering Structure) は，密度ベースのクラスタリングである．OPTICS の入力は，点データセット D と，ふたつのパラメータ近傍距離 ϵ，最小近傍点数 $MinPts$ である．クラスタ数は不要であり，訪問履歴データのような，事前にクラスタ数を決定することが難しい場合に適正している．

距離 ϵ 以内の点の数が $MinPts$ 以上の点 $p \in D$ をコア点と呼ぶ．また，p からの距離 $MinPts$ 番目の点までの距離と，$p, q \in D$ 間の距離の大きい方を，p から q への到達性距離と呼ぶ．図 1 において，点 o がコア点，$r(p_1)$ が p_1 への，$r(p_2)$ が p_2 への到達性距離である．

OPTICS では，ϵ 最大値と $MinPts$ を入力とし，まず，各点のクラスタリング順序 $order$ と到達性距離 $reachability$ を出力する．これらを用いて，最大値以下の任意の ϵ に対し，効率良く $O(N)$ の計算でクラスタリング結果を出力する．ϵ 以下の到達性距離を持ち，クラスタリング順序の隣接する点集合がクラスタとなる．

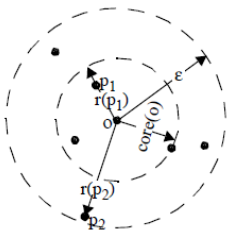

図 1　コア点と到達可能距離 [2]

横軸に $order$，縦軸に $reachability$ をとったグラフを到達性プロットと呼ぶ．例を図 2 に示す．階層的クラスタリングにおける樹形図に類する図であり，パラメータ ϵ と抽出されるクラスタの関係を可視化する．到達性プロットの ϵ 以下で連続する凹部と 1 つのクラスタが対応する．例えば，図 2 では，小さな ϵ で 3 つの小さなクラスタ，より大きな ε でそれらを囲むひとつのクラスタが生成されることが読み取れる．

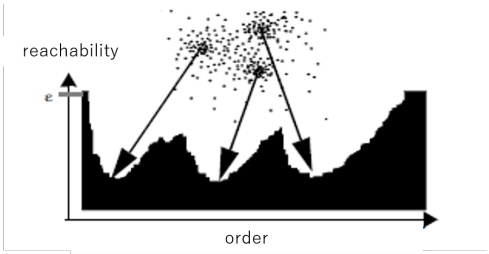

図 2　クラスタ構造と到達性プロット [2]

4　クラスタリングインタフェースの提案

4.1　OPTICS のインタフェースと課題

OPTICS によるインタラクティブ・インタフェースでは，到達性プロット上でクラスタ構造を確認しながら，ϵ の調整を行う．しかし，図 3 のような大規模データを対象とした OPTICS では，到達性プロットの視認性が低下し，クラスタ構造の読み取りが困難である．

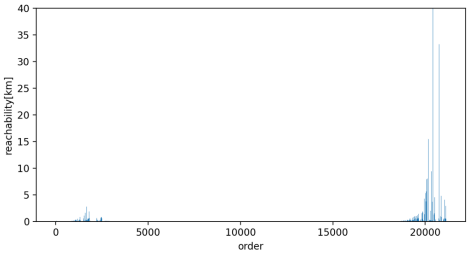

図 3　約 2 万点の到達性プロット ($reachability \geq 40$ の領域省略)

関心ある領域に絞り込む方法として，データをビューポートで絞り込んでからクラスタリングを行う方法が考えられる．データを絞り込んでからクラスタリングを行う例を図 4 に示す．四角形の枠がビューポートであり，クラスタの領域は凸包によって示している．全てのデータがビューポートに入るとき (図 4(1)) と，クラスタの一部のみビューポートに入るとき (図 4(2)) とで，得られるクラスタが異なり，地図操作に対して，結果の一貫性を保てない．

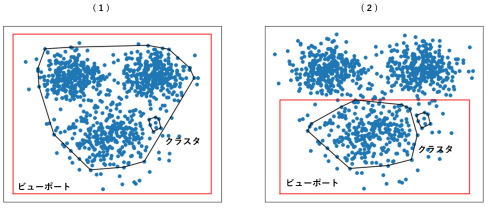

図 4　ビューポートによるクラスタリング結果の変化

4.2　ズームを考慮した到達性プロット

到達性プロットをビューポートに応じて集約化することで，関心ある領域に絞りこむ．ビューポート内の点群は，到達性プロット上で連続していると限らず，ビューポート外の点を間引くだけだと，クラスタ構造が表現できない．そこで，ビューポート外の連続する点列を，仮想的な 1 つの点に置き換えることで，ビューポート外のクラスタ構造を表現する．仮想的な点の到達距離の値は，置き換えられる点の到達距離の最大値とする．

4.3　インタフェース構成とインタラクション手順

インタフェースの構成を図 5 に示す．インタフェースは大きく，上部の地図と下部の集約化された到達性プロットに分けられる．地図では，スクロール・ズーム操作が可能であり，クラスタ領域を凸包で表示することで，現在の ϵ によるクラスタリング結果を可視化する．到達性プロットは，地図内のクラスタ構造を可視化する．地図を操作して関心のある領域を表示し，地図内の到達性プロットと，地図上の結果を確認しながら ϵ を調整し，クラスタリングを行う．

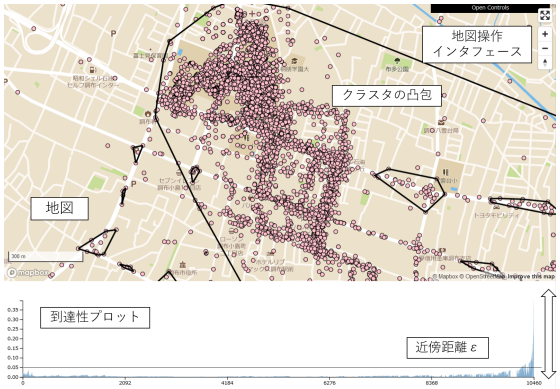

図 5　インタフェース構成例

5　適用結果

提案手法の有効性と限界を明らかにするため，現在，ズームを考慮した到達性プロットによる視認性の改善について，前節で示した GLH データを対象に定量評価を行っている．

参考文献

[1] Bae et al., Interactive Clustering: A Comprehensive Review, ACM Computing Surveys 53(1), pp.1-39, 2020

[2] Ankerst et al., OPTICS: Ordering points to identify the clustering structure. ACM Sigmod Record 28, pp.49-60, 1999