ISS-A-070

セル結合判定の順序付けに基づく CDBSCAN アルゴリズムの高速化

三木　直人† 　酒井　達弘†† 　田村　慶一‡

† 広島市立大学情報科学部　†† 島根大学学術研究院理工学系　‡広島市立大学大学院情報科学研究科

1. はじめに

IoT（Internet of Things）をはじめとして理工学，医学，生物学や社会学など様々な分野において分析のためのデータクラスタリング手法への関心が高まっている．クラスタリングを行う手法のひとつに密度に基づくクラスタリングがあり，代表的な手法として DBSCAN（Density-Based Spatial Clustering of Applications with Noise）アルゴリズムが提案されている．

近年，高速化手法としてセル分割に基づく DBSCAN アルゴリズム（以下，CDBSCAN アルゴリズムと呼ぶ），また，CDBSCANアルゴリズムのセルの結合判定を高速化した MBR-CDBSCAN アルゴリズム[1]が提案されている．しかしながら，大規模なデータセットにおいてはクラスタ構築時におけるセルの結合判定回数が増加して処理が遅いという課題があった．そこで本研究ではセルの結合判定の順序を工夫し，結合判定回数を削減することで CDBSCAN アルゴリズムの高速化を図る．

2. CDBSCAN アルゴリズム

DBSCAN アルゴリズムではデータ周辺のデータ数を当該データの密度と定義する．あるデータに対し，距離がϵ以内である自身を含めたデータを近傍データと呼び，近傍データの数が$MinPts$以上のデータをコアデータと呼ぶ．距離がϵ以内であるコアデータ同士を結合していくことでクラスタを形成していく．

CDBSCAN アルゴリズムではデータセット全体を対角線の長さがϵの矩形であるセルに分割する．セル内に含まれる全てのデータ間の距離はϵ以内であるため，セル単位でコアデータ判定や結合判定を行うことができる．コアデータが含まれるセルに所属するデータはセル単位でクラスタに割り当てられ，距離ϵ以内のコアデータが存在するような 2 つのセル同士を結合していくことでクラスタを構築していく．このセルのペア間に距離がϵ以内であるコアデータのペアが存在するか確かめる判定をセルの結合判定と呼ぶ．

3. 提案手法

提案手法のセルの結合判定は，最小外接矩形（MBR）を用いたセルの結合判定[1]をベースとする．セルの結合判定を行う際，2 つのセルについてそれぞれ自身に含まれるコアデータを囲む MBR を作成し，MBR 間のコアデータの最小距離の上限と下限を求めることで結合判定を行う．MBR を用いた判定が行えない場合のみコアデータ間の距離計算を行うため，距離計算の回数を大幅に削減可能である．

提案手法では，コアデータを含むセルをノード，2 つのセルが結合する関係をエッジとしてグラフを作成することでクラスタを構築する．グラフ作成の際，結合判定の順序を工夫し，距離が近いセルのペア間の結合判定を優先して行う．距離の近いセルのペアは比較的結合する可能性が高いため，早い段階で結合することで無駄な結合判定を削減することができる．

4. 評価実験

提案手法を評価するために，提案手法と提案手法からセルの結合判定の順序付けのみを取り除いたアルゴリズムについて，データ数 1000 万件の人工データを用いて比較実験を行った．実験ではセルの結合判定のみの処理時間を計測した．また，$\epsilon = 5000$，$MinPts = 100$とした．計測は 5 回行い，その平均値を表 1 に示す．表 1 より，提案手法は全ての次元について高速にクラスタリングを行えていることが分かる．

表 1 人工データを使用した実験結果[sec]

次元数	順序付けなし	提案手法
2	2.02	0.32
3	2.24	0.12
5	17.86	0.14
7	16.18	0.21

5. まとめ

本研究では，セルの結合判定の順序を工夫した新しいセル分割に基づくDBSCANアルゴリズムを提案した．評価実験の結果，セルの結合判定の順序を工夫することで高速にクラスタリングを行うことができると分かった．今後の研究として，データ件数の大きなデータセットを使用した実験を行うことで高速化の余地を見つけ出すことが挙げられる．

謝辞

本研究の一部はJSPS 科研費 JP18K11320の助成により行われた．

参考文献

[1] 酒井達弘，田村慶一，北上始，竹澤寿幸，最小外接矩形とセルの再帰分割を用いたセルベースの DBSCAN の高速化，電子情報通信学会論文誌，Vol. J101-D，No. 4，pp. 690-701，2018.