ストリーミングデータを対象とした密度ベースクラスタリング

慶應義塾大学大学院　理工学研究科　齋藤勇太
慶應義塾大学　　理工学部　　鈴木秀男

1.はじめに

データ解析の重要性がより高まる昨今、特にリアルタイムでの分析対象となるストリーミングデータに対してのアプローチに多くの議論[1]がある。情報が更新され続けるストリーミングデータに対して、分析方法の動的調整や任意な構造を持つデータへの対応が強く求められる。本研究では、常に情報が更新され続けるストリーミングデータの教師なしという特性に対して、任意形状の抽出に優れた密度ベースクラスタリング手法である DBSCAN[2]を活用する。さらに、マイクロクラスタという概念を用いてストリーミングデータに対する処理を改善し、ストリーミングデータにおけるクラスタリング構造の把握を高い精度で実現することを目指す。

2.分析概要

本研究では、マイクロクラスタ[3]という概念を用いてデータを要約し、マイクロクラスタの仮想点を用いたクラスタリングを行うことで DBSCAN の計算量が多いという問題を解消する。また、スプリット＆マージによる逐次分類処理[4]を行うことでマイクロクラスタレベルでのエラーの発生を回避し、クラスタリング最終結果に及ぼすエラーを防ぐ。

ストリーミングデータ処理と DBSCAN への流れの概要は以下の通りである。

① 到着データのマイクロクラスタ(通常マイクロクラスタと外れ値マイクロクラスタ)振り分け
② 更新されたマイクロクラスタのスプリット&マージ判定
③ 外れ値マイクロクラスタの通常マイクロクラスタへの成長
④ 一定時間間隔におけるマイクロクラスタの存続チェック
⑤ マイクロクラスタの中心を仮想点とした DBSCAN の実行

参考文献

[1] R. M. M. Vallim, J. A. A. Filho, R. F. de Mello, A. C. P. L. F. de Carvalho, and J. Gama, “Unsupervised Density-based Behavior Change Detection in Data Streams,” *Intell. Data Anal.Unsupervised Density-based Behav. Chang. Detect. Data Streams*, vol. 18, no. 2, pp. 181–201, 2014.

[2] M. Daszykowski and B. Walczak, “Density-Based Clustering Methods,” *Compr. Chemom.*, vol. 2, pp. 635–654, 2010.

[3] F. Cao, M. Ester, W. Qian, and A. Zhou, “Density-based clustering over an evolving data stream with noise,” *Proc. Sixth SIAM Int. Conf. Data Min.*, vol. 2006, pp. 328–339, 2006.

[4] Suzuki, Enkawa. “A Data Clustering Based on MDL Criterion” *J-STAGE Free*. Vol22(1993) Issue3 Pages117-132.