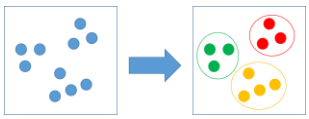
機械学習基盤を用いたクラスタリングの高速化に関する研究

情報基礎科学専攻　小林・佐藤研究室　村上 洸

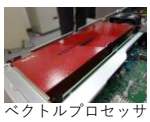
研究背景・アプローチ

- **機械学習の急速な普及**
  - 画像認識やデータ分類
  - 代表的な手法としてデータ分類に用いられる
    - クラスタリング等
- **クラスタリングの問題点:計算コスト**
  - 分析の対象になるデータ量の増大
    - データ同士の関係を計算
    - データ量の増大に伴う計算時間の増大



クラスタリング
の概念図

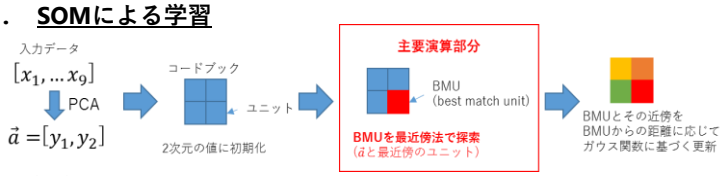
- **アプローチ:アクセラレータの活用**
  - 機械学習の高速化手法としてアクセラレータの活用
  - **アクセラレータとは**
    - 特定の演算を高速に実行するプロセッサ
    - 汎用のCPUをホストとして利用
    - GPUやベクトルプロセッサなど



ベクトルプロセッサ
SX-Aurora TSUBASA

SOMとk-means

1. **SOMによる学習**



- 全データを学習する期間をエポックと呼称し，学習時はエポックを繰り返す

2. **k-meansに基づく可視化**

主要演算部を機械学習基盤を用いてアクセラレータにオフロード

機械学習基盤とその課題

- **アクセラレータを用いた機械学習の高速化が主流**
  - 本研究ではGPUとベクトルプロセッサを活用

GPU	ベクトルプロセッサ
高演算性能	高メモリバンド幅
高並列性	長いベクトル長

アクセラレータごとに特徴が異なる

演算やデータごとに適切なアクセラレータが異なる可能性

- **機械学習基盤**
  - 機械学習を容易に実現できるライブラリ
    - 代表的なものとしてscikit-learn (汎用CPU用)
  - アクセラレータ用の機械学習基盤の拡充
    - アクセラレータへのオフロードが容易に

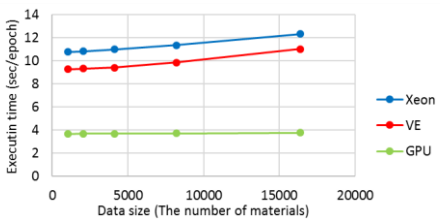
課題：適切なアクセラレータの選択方法が未確立

予備評価

評価に用いたプラットフォーム

プロセッサ	Xeon Gold 6126	Tesla V100	Vector Engine 10B
FLOPS (Double)	998.4 GFLOPS	7.8 TFLOPS	2.15 TFLOPS
メモリバンド幅	128 GB/s	900 GB/s	1.2 TB/s
コア数	12	2560	8
機械学習基盤	scikit-learn	RAPIDS	Frovedis

SOMの予備評価



- GPU, VE, Xeonの順番で実行時間が短い
  - VEは全コア使用できていなかったため，Xeon比で約10%の速度向上
    - VE1コアとXeonの演算性能効率ではVEの方が良い
- データサイズに比例して，実行時間が増大
  - データサイズに比例して，計算量が増加しているため

データサイズや使う関数によって実行時間が短いプロセッサが異なる

最適なプロセッサへの処理のオフロードが重要！

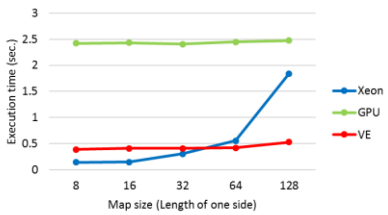
実験内容

- 液体物質クラスタリングコード[1]で評価
  - SOMで学習する部分とk-meansによる可視化部分で構成

1. **SOM部分の予備評価**
   - 9次元のダミーデータセットを作成し，物質数とSOMの実行時間の関係を評価(データサイズごとにエポック数が異なるため，1エポック当たりの実行時間で評価)
2. **k-means部分の予備評価**
   - 物質数を固定してマップサイズを変化させて，k-meansの実行時間を評価(2次元の値を持つユニットをクラスタリング)

[1] G. Kikugawa; et al. Data analysis of multi-dimensional thermophysical properties of liquid substances based on clustering approach of machine learning. Chemical Physics Letters, Vol. 728, pp. 109-114, August 2019.

k-meansの予備評価



- マップサイズが小さいときはXeon，大きいときはVEの実行時間が最も短い
  - プロファイルの結果，GPUは立ち上げ時間などの計算以外の時間がボトルネック
  - マップサイズが小さい時に，アクセラレータのベクトル長や並列性を生かし切れていない

今後の研究計画と結論

- **研究計画**
  - 最適なオフロード方法の確立
  - 熱物理特性が増えた場合に，ボトルネックとなり得るPCAについての調査を行い，アクセラレータを活用
  - 本手法以外のクラスタリング手法の検討を行い，実用的かつ高速に実行できる手法の提案
    - UMAP, t-sneなどの手法を検討
- **結論**
  - SOMではGPU，k-meansではマップサイズが大きい場合にVE，小さい場合にXeonの実行時間が最も短い
  - データサイズや使われる関数に応じて，適切なプロセッサを使い分けることが重要
  - 本手法以外の実用的かつ高速に実行できるクラスタリング・可視化方法の検討