

研究内容詳細

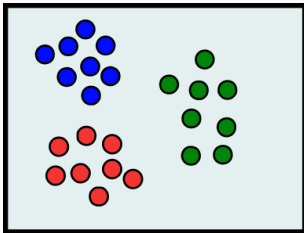
クラスタリング手法の評価に向けて

池辺 颯一

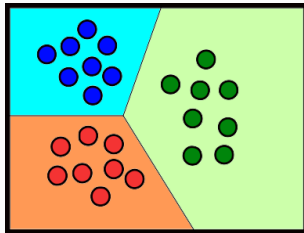
2018 年 12 月 15 日

芝浦工業大学

- 情報化社会の発展によりデータが複雑かつ膨大に
- ビッグデータを人の手で分類するのは難しい
- それらのデータを自動的に分類するクラスタリングに着目
- 機械学習における教師なし学習にあたる



クラスタリング前



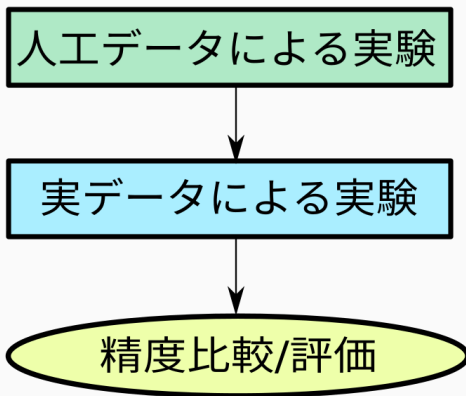
クラスタリング後

目的

- クラスタリング手法の1つである Fussy c-means を応用した最適化問題の中から最も精度が高いものを発見する

目標

- 各クラスタリング手法のプログラム C++を用いて開発
- プログラムの実行結果からクラスタリング精度を評価



クラスタリングの最適化問題

eFCMA

$$\underset{u,v,\pi}{\text{minimize}} \sum_{i=1}^C \sum_{k=1}^N u_{i,k} \|x_k - v_i\|_2^2 + \lambda^{-1} \sum_{i=1}^C \sum_{k=1}^N u_{i,k} \log\left(\frac{u_{i,k}}{\pi_i}\right)$$

qFCMA

$$\underset{u,v,\alpha}{\text{minimize}} \sum_{i=1}^C \sum_{k=1}^N (\alpha_i)^{1-m} (u_{i,k})^m \|x_k - v_i\|_2^2 \\ + \frac{\lambda^{-1}}{m-1} \sum_{i=1}^C \sum_{k=1}^N (\alpha_i)^{1-m} (u_{i,k})^m$$

sFCMA

$$\underset{u,v,\alpha}{\text{minimize}} \sum_{i=1}^C \sum_{k=1}^N (\alpha_i)^{1-m} (u_{i,k})^m \|x_k - v_i\|_2^2 \\ \text{subject to } \sum_{i=1}^C u_{i,k} = 1, \sum_{i=1}^C \alpha_i = 1 \text{ and } u_{i,k} \in [0, 1] \quad m > 1$$

- N : 個体数
- C : クラスタ数
- λ, m : ファジィ化パラメータ
- $u_{i,k}$: i 番目の個体におけるクラスタ k に対する帰属度
- v_i : i 番目のクラスタ中心
- x_k : k 番目の個体

FCM(Fussy c-means)

1. 初期クラスタ中心 V を与える
2. V から帰属度 U を更新する
3. V を更新する
4. 収束条件を満たせば終了。満たさなければ2へ。

ARI (Adjusted Rand Index)

- -1 から 1 までの範囲で精度評価を行う指標
- 1 の時に完全一致で 0 の時にランダム
- マイナスの値はランダムの期待値を下回る
- ARI の値が高いほど高評価

Yeast Data Set

- Yeast(酵母) の形など 9 属性を収録したデータ
- ソース : UCI Machine Learning Repository
- 個体数 : 1484
- クラス数 : 10

- sFCM を動作させるのに必要なプログラムが完成

- 処理の高速化
- 既存手法からの継承

目的

- クラスタリング手法の 1 つである Fussy c-means を応用した最適化問題の中から最も精度が高いものを発見する

目標

- 各クラスタリング手法のプログラム C++を用いて開発
- プログラムの実行結果からクラスタリング精度を評価

進捗

- sFCM を動作させるのに必要なプログラムが完成

課題

- 処理の高速化