通信ゼミナール
クラスタリング手法の評価

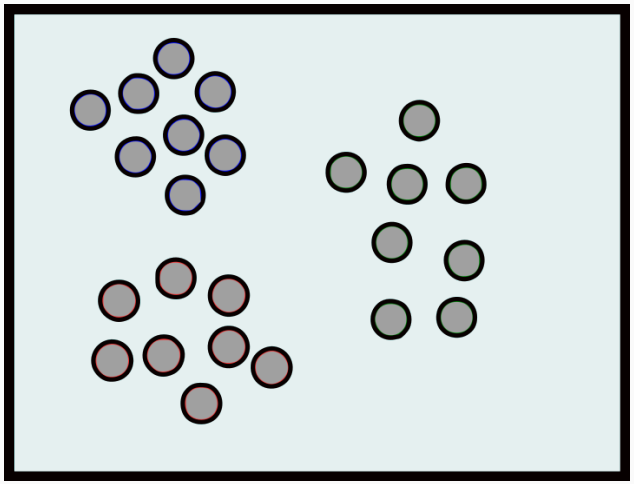
池辺 颯一

2018 年 12 月 12 日

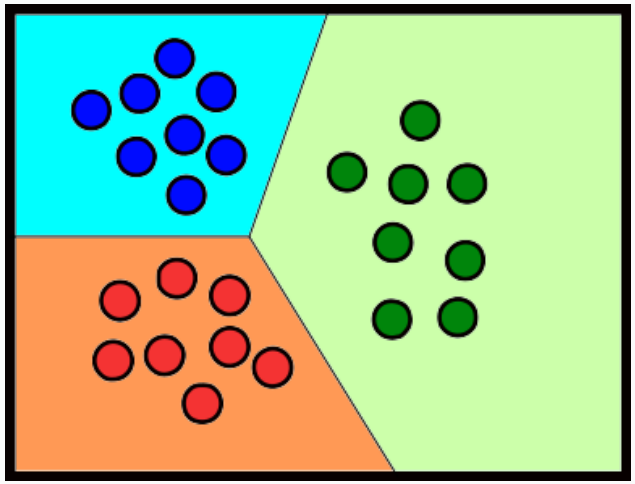
芝浦工業大学 工学部 通信工学科

概要・背景

- 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



クラスタリング前



クラスタリング後

目的・目標

目的

- クラスタリング手法の 1 つである Fussy c-means にクラスタサイズ調整変数を導入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を 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

実験対象

既存手法

- sFCM
- pFCM
- eFCM

比較対象手法

- **クラスタサイズ調整変数を導入**
- sFCMA
- pFCMA
- eFCMA

クラスタリングの最適化問題

eFCMA

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

N	データ数	x_k	データ数
C	クラスタ数	v_i	クラスタ中心
λ	ファジィ化パラメータ	$u_{i,k}$	帰属度
α_i	クラスタサイズ調整変数		

クラスタリングの最適化問題

qFCMA

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

N	データ数	x_k	データ数
C	クラスタ数	v_i	クラスタ中心
λ, m	ファジィ化パラメータ	$u_{i,k}$	帰属度
α_i	クラスタサイズ調整変数		

クラスタリングの最適化問題

sFCMA

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

$$\text{subject to} \sum_{i=1}^{C} u_{i,k} = 1 \ , \ \sum_{i=1}^{C} \alpha_i = 1 \text{ and } u_{i,k} \in [0,1] \quad m > 1$$

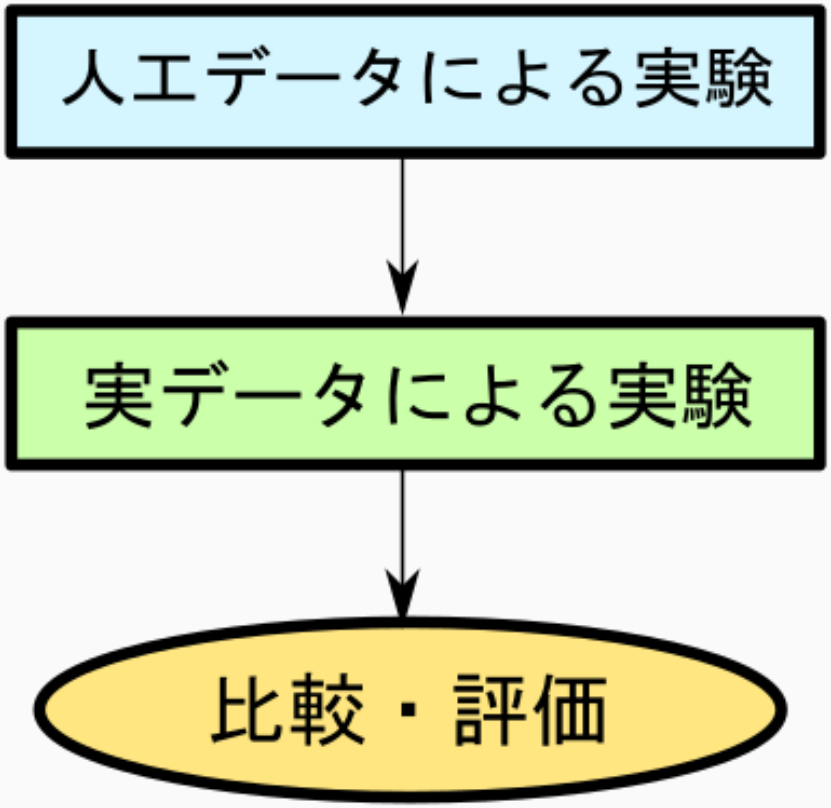
N	データ数	x_k	データ数
C	クラスタ数	v_i	クラスタ中心
m	ファジィ化パラメータ	$u_{i,k}$	帰属度
α	クラスタサイズ調整変数		

アルゴリズム

FCM(Fusssy 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。**

実験方法



評価方法

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
- ARI の値が高いほど高評価

使用する実データ

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

実験結果

考察

-

まとめ

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

実験結果

-

考察

-