クラスタサイズ調整変数を導入したクラスタリング手法の評価

池辺 颯一

情報数理工学研究室

af16009@shibaura-it.ac.jp

2019 **年** 1 **月** 16 **日**

研究背景

クラスタリング

- 情報通信社会の発展に伴いデータ量が増大
- データを類似度に基づきグループ化するクラスタリングに着目

クラスタリングの欠点

- 各クラスタのサイズに差がある場合に有意な結果が得られない場合がある
- クラスタのサイズを考慮して分類をする手法が提案されている

研究目的

目的

クラスタサイズ調整変数を導入した手法について

- 各手法の特性を把握
- 最も有用なものを発見

目標

- 各手法で 2 クラス分類を行い特性を把握する
- 各手法のクラスタリング精度の算出及び比較を行う

実験対象

提案手法

- sFCMA
- eFCMA
- qFCMA

クラスタリングの最適化問題

sFCMA

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

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

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_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

eFCMA

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

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

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

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

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

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		

研究方法

研究手順

1. 人工データ実験
2. 実データ実験
3. 実データ実験で算出した ARI により各手法を評価
4. 3 手法の中で ARI の最高値が算出された手法を高評価とする

研究方法

人工データ実験の評価指標

- 評価指標として分類関数を用いる

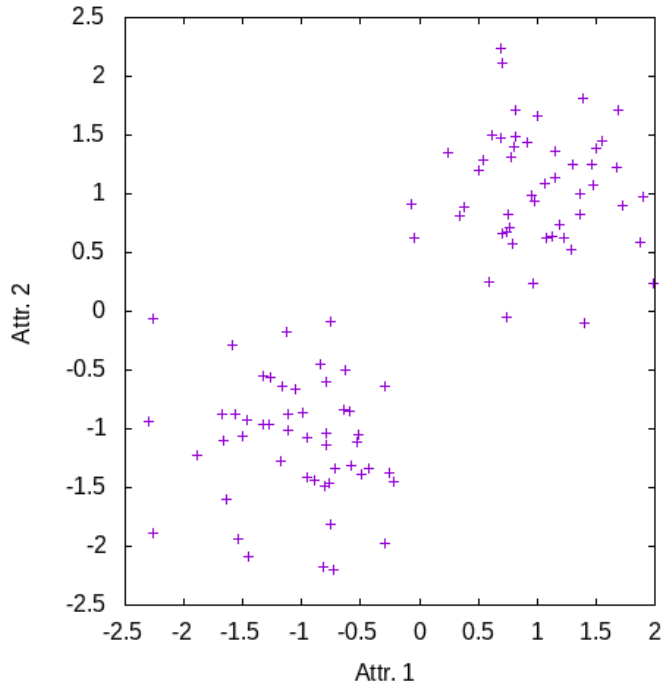
実データ実験の評価指標

- 評価指標として ARI(Adjusted Rand Index) を用いる
- -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

人工データ実験

使用する人工データ

- **平均値** (–1, –1), **標準偏差** (0.5, 0.5) **及び平均値** (1, 1), **標準偏差** (0.5, 0.5) **のガウスサンプリングで生成**
- **データ数**: 100
- **クラス数** : 2

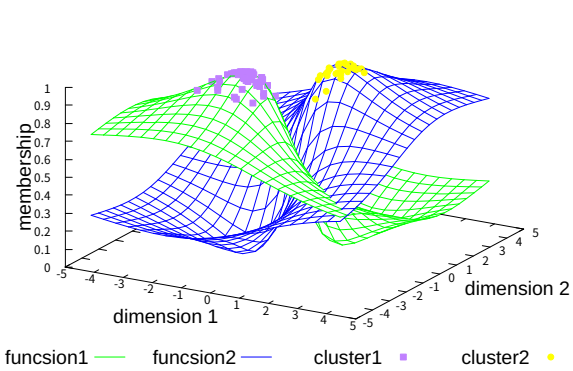


アルゴリズム

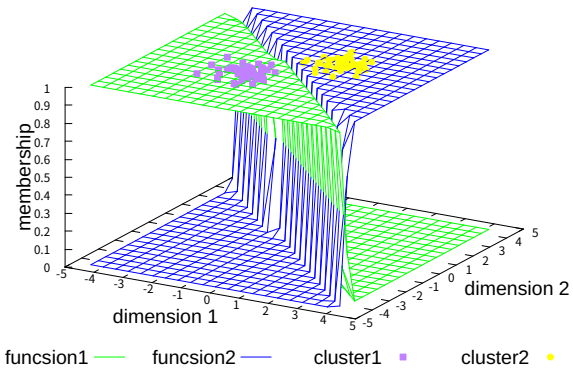
1. クラスタ中心をランダムに与える．
2. クラスタ中心を用いて帰属度を更新する．
3. 帰属度を用いてクラスタ中心及びクラスタサイズ調整変数を更新する．
4. 収束すれば終了し，そうでない場合は 2 に戻る．

実験結果

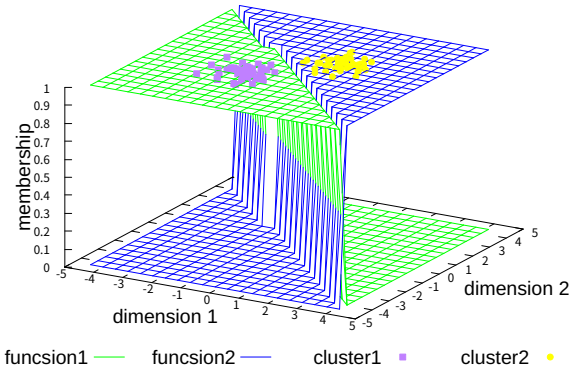
sFCMA



$m = 2.00$



$m = 1.01$



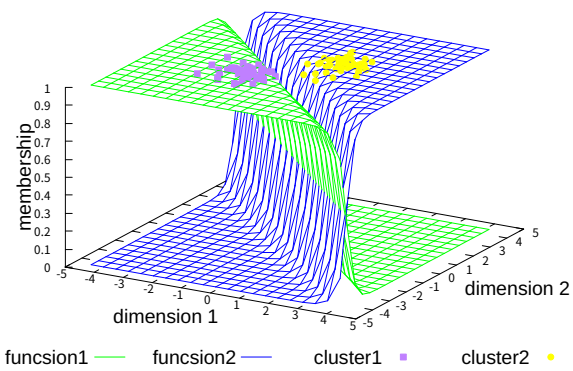
HCM

sFCMA の特徴

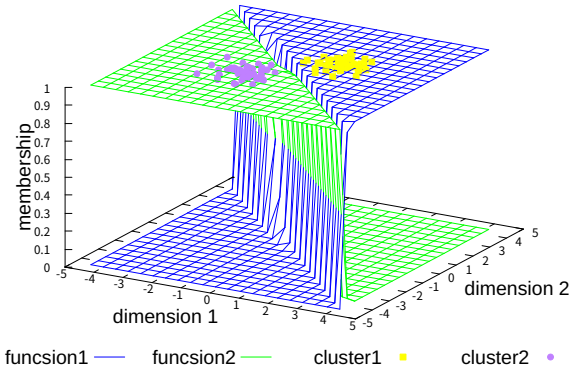
パラメータ m を 1 に近づけるほどクリスプになり，HCM に近づく

実験結果

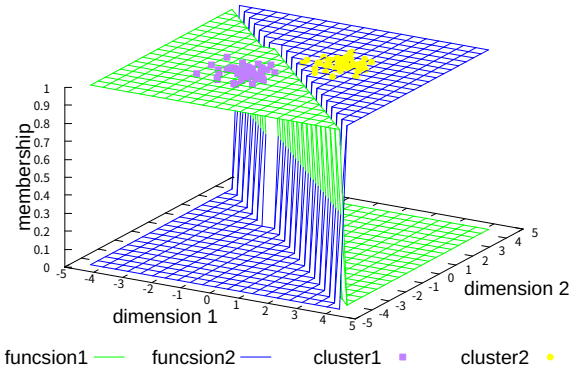
eFCMA



$\lambda = 1$



$\lambda = 10$



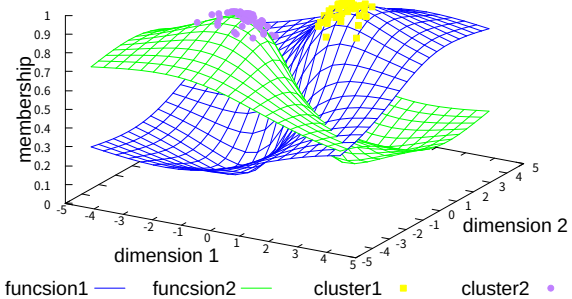
HCM

eFCMA の特徴

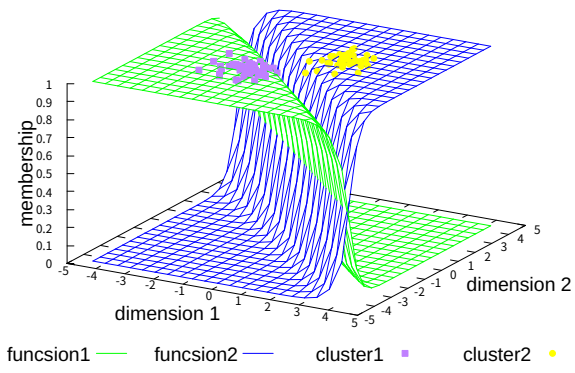
パラメータ λ を大きくするほどクリスプになり，HCM に近づく

実験結果

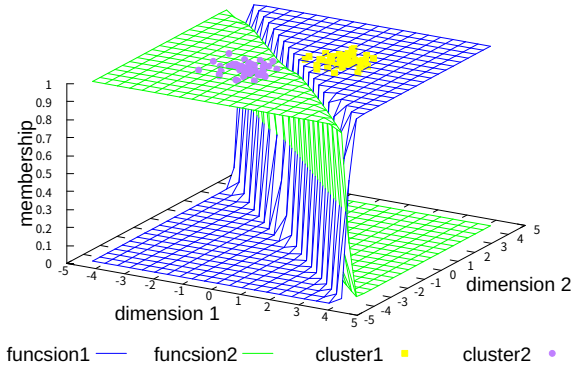
qFCMA



$m = 2.00\ ,\ \lambda = 1$



$m = 1.01\ ,\ \lambda = 1$

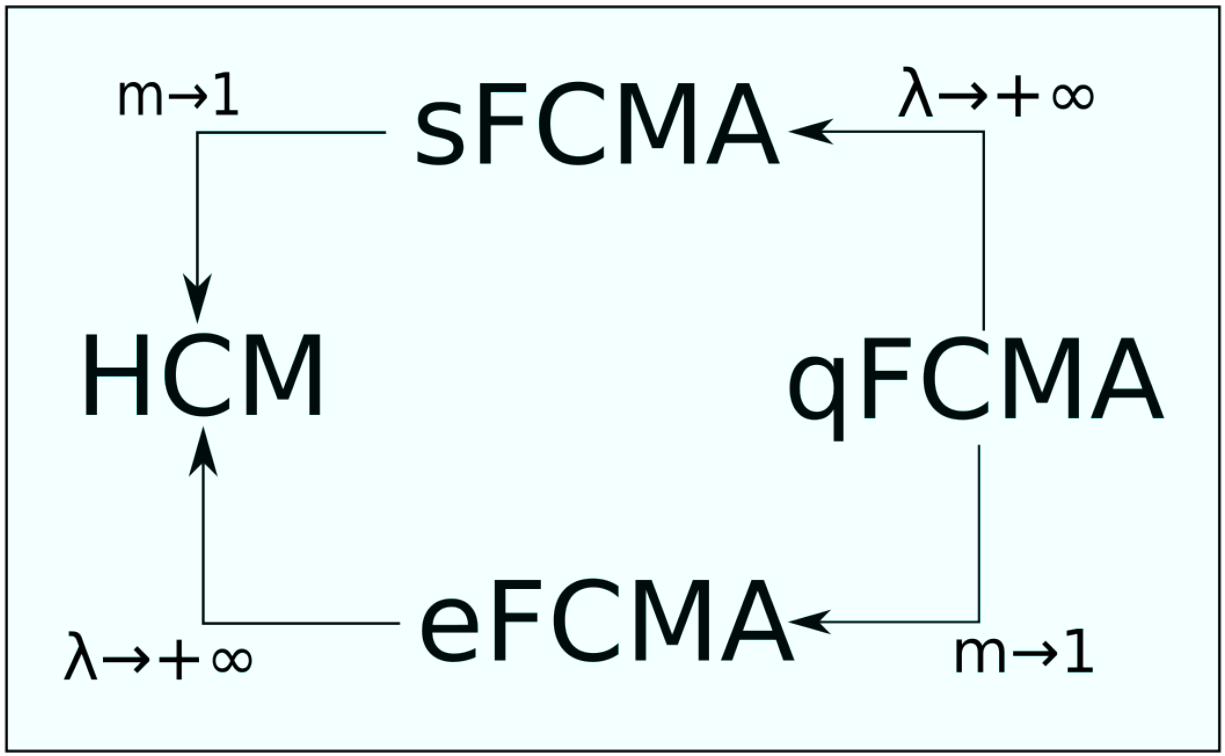


$m = 1.01\ ,\ \lambda = 10$

qFCMA の特徴

- パラメータ m を 1 に近づけるとクリスプになる
- パラメータ λ が大きいほどクリスプになる

各手法間の関係



実データ実験

使用する実データ

- User Knowledge Modeling Data Set
- **被験者の勉強時間，試験結果など 5 属性を収録したデータ**
- **ソース** : UCI Machine Learning Repository
- **個体数** : 403
- **クラス数** : 4(**非常に低い，低い，中央，高い**)

実験条件

sFCMA

パラメータ m を 1.1 から 0.1 刻みで 3.0 まで変化させる

eFCMA

パラメータ λ を 1 から 1 刻みで 100 まで変化させる

qFCMA

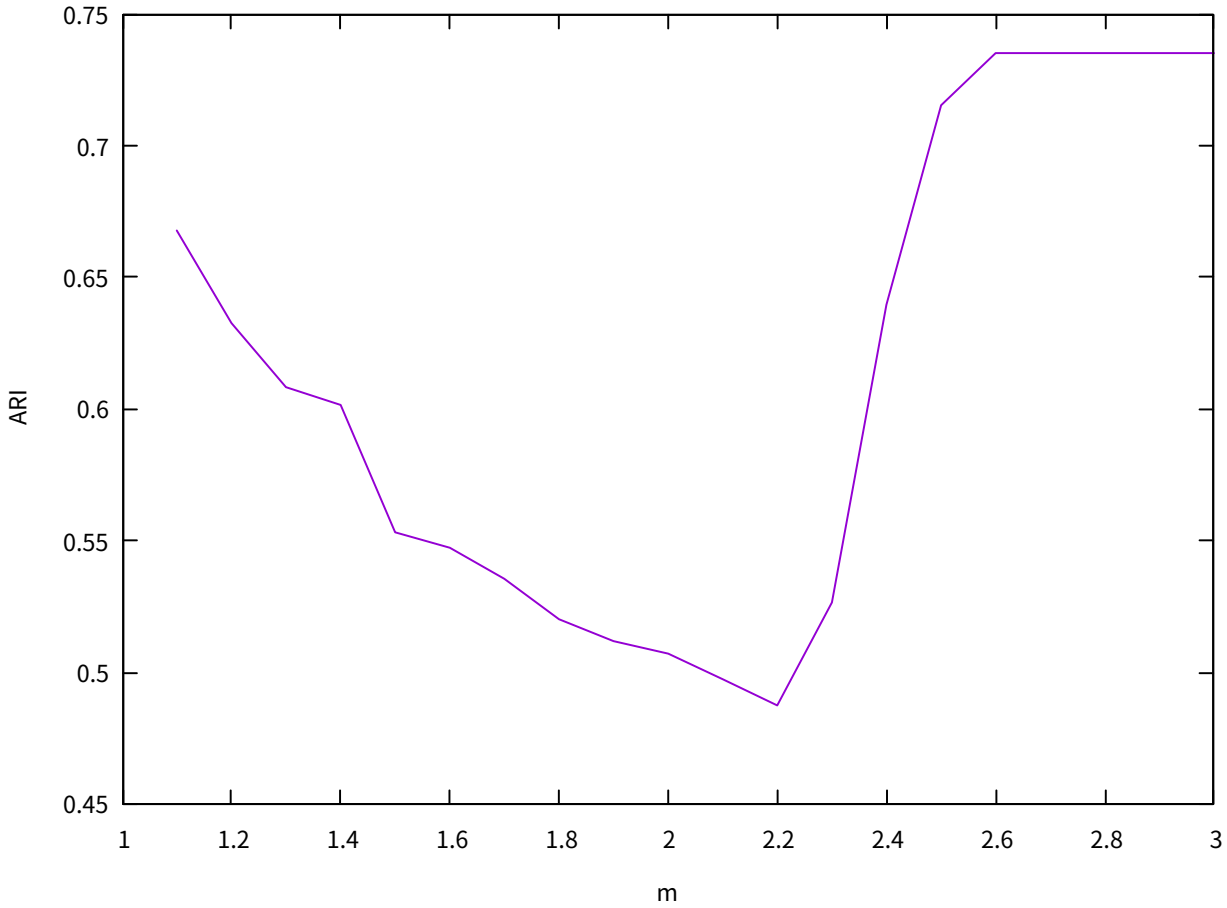
- **パラメータ λ を 1 から 1 刻みで 100 まで変化させる**
- **パラメータ m を 1.1 から 0.1 刻みで 3.0 まで変化させる**

アルゴリズム

1. 正解帰属度を用いて帰属度を初期化する．
2. 帰属度を用いてクラスタ中心及びクラスタサイズ調整変数を更新する．
3. 収束すれば終了し，そうでない場合は 2 に戻る．

実験結果

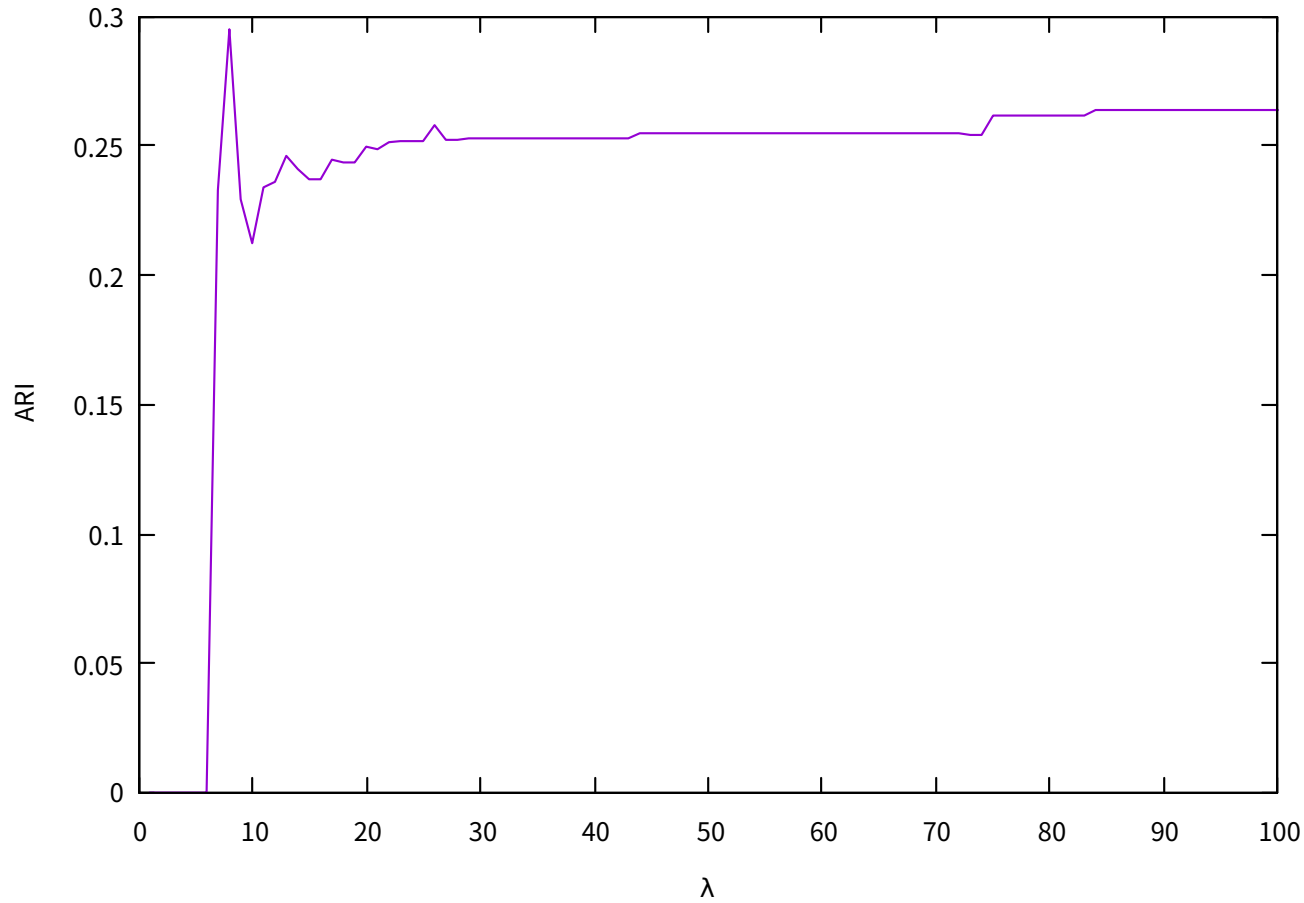
sFCMA



最高 ARI:0.73515 ($m = 3$)

実験結果

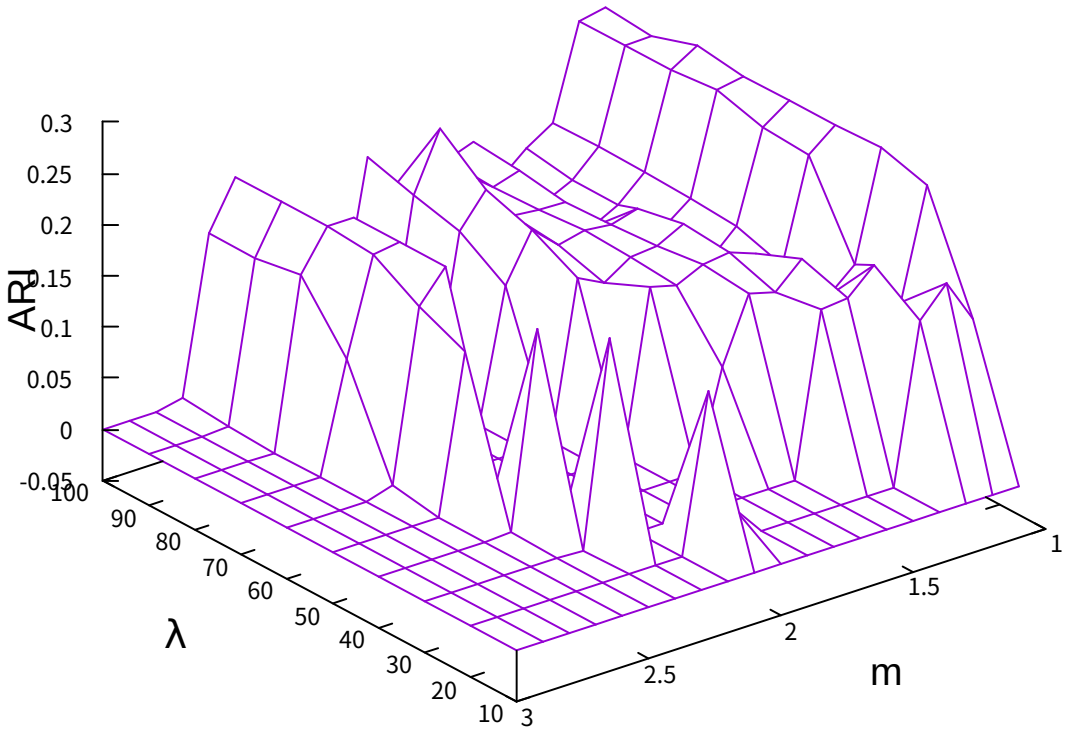
eFCMA



最高 ARI:0.29500 ($\lambda = 8$)

実験結果

qFCMA



最高 ARI:0.26286 ($\lambda = 80, m = 1.1$)

実験結果

各手法の最高 ARI

手法名	ARI の最高値	パラメータ値
sFCMA	0.73515	$m = 3$
eFCMA	0.29500	$\lambda = 8$
qFCMA	0.26286	$\lambda = 80, m = 1.1$

評価

sFCMA が最も高評価

考察

考察

- sFCMA と eFCMA 及び qFCMA との差は，エントロピー項の有無．
- エントロピー項を削除したことが計算結果に影響したと考えられる．

まとめ

背景

データ量の増大によりクラスタリングに注目が集まっている

目的

クラスタサイズ調整変数を導入した手法の中で有用なものを発見する

実験結果

sFCMA が最も高評価となった

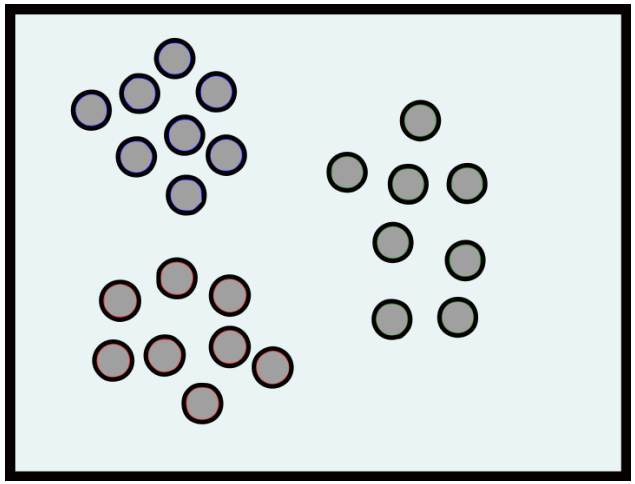
考察

エントロピー項を削除したことが計算結果に影響したと考えられる

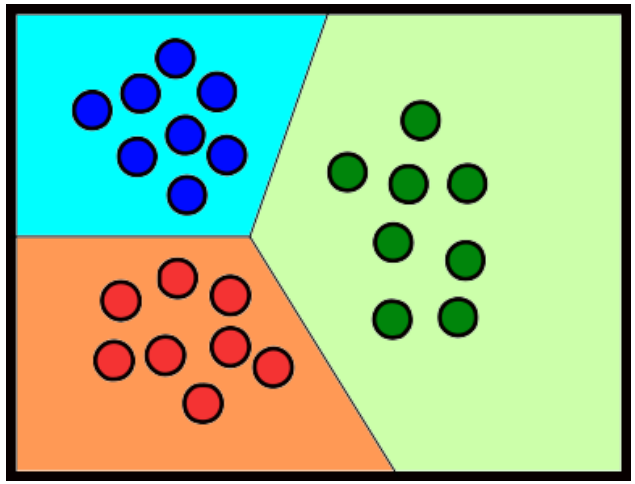
クラスタリングについて

クラスタリングとは

データを類似度に基づいてグループ化するデータ分析手法の 1 つ



クラスタリング前



クラスタリング後

HCM(Hard c-means)

最適化問題

$$\underset{u,v}{\text{minimize}} \sum_{i=1}^{C} \sum_{k=1}^{N} u_{i,k} \|x_k - v_i\|_2^2 \quad \text{subject to} \quad \sum_{i=1}^{C} u_{i,k}=1 \quad \text{and} \quad u_{i,k} \in \{0, 1\}$$

$$u_{i,k} = \begin{cases} 1 & (i = \arg\min_{1 \le j \le C} \{\|x_k - v_i\|_2^2\}) \\ 0 & (\text{otherwise}) \end{cases},$$

$$v_i = \frac{\sum_{k=1}^{N} u_{i,k} x_k}{\sum_{k=1}^{N} u_{i,k}}.$$

sFCMA の更新式

$$
\begin{aligned}
v_i &= \frac{\sum_{k=1}^{N} (u_{i,k})^m x_k}{\sum_{k=1}^{N} (u_{i,k})^m}, \\
u_{i,k} &= \frac{1}{\sum_{j=1}^{c} \frac{\alpha_j}{\alpha_i} \left(\frac{d_{j,k}}{d_{i,k}} \right)^{\frac{1}{1-m}}}, \\
\alpha_i &= \frac{1}{\sum_{j=1}^{C} \left(\sum_{k=1}^{N} \frac{(u_{j,k})^m d_{j,k}}{(u_{i,k})^m d_{i,k}} \right)^{\frac{1}{m}}}.
\end{aligned}
$$

eFCMA の更新式

$$
\begin{aligned}
v_i &= \frac{\sum_{k=1}^{N} u_{i,k} x_k}{\sum_{k=1}^{N} u_{i,k}}, \\
u_{i,k} &= \frac{\alpha_i \exp(-\lambda \|x_k - v_i\|_2^2)}{\sum_{j=1}^{C} \alpha_j \exp(-\lambda \|x_k - v_j\|_2^2)}, \\
\alpha_i &= \frac{\sum_{k=1}^{N} u_{i,k}}{N}.
\end{aligned}
$$

qFCMA の更新式

$$
\begin{aligned}
v_i &= \frac{\sum_{k=1}^{N} (u_{i,k})^m x_k}{\sum_{k=1}^{N} (u_{i,k})^m}, \\
u_{i,k} &= \frac{\alpha_i (1 + \lambda(1-m)\|x_i - v_k\|_2^2)^{\frac{1}{1-m}}}{\sum_{j=1}^{C} \alpha_j (1 + \lambda(1-m)\|x_j - v_k\|_2^2)^{\frac{1}{1-m}}}, \\
\alpha_i &= \frac{1}{\sum_{j=1}^{C} \left(\sum_{k=1}^{N} \frac{(u_{j,k})^m (1-\lambda(1-m)d_{j,k})}{(u_i,k)^m (1-\lambda(1-m)d_{i,k})} \right)^{\frac{1}{m}}}.
\end{aligned}
$$