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

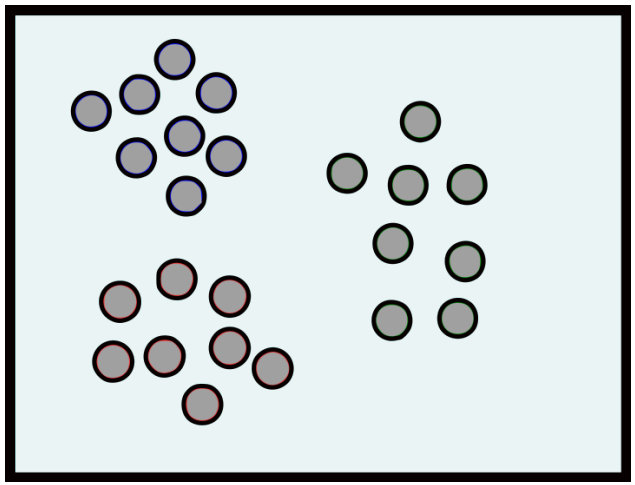
池辺 颯一

芝浦工業大学 工学部 通信工学科

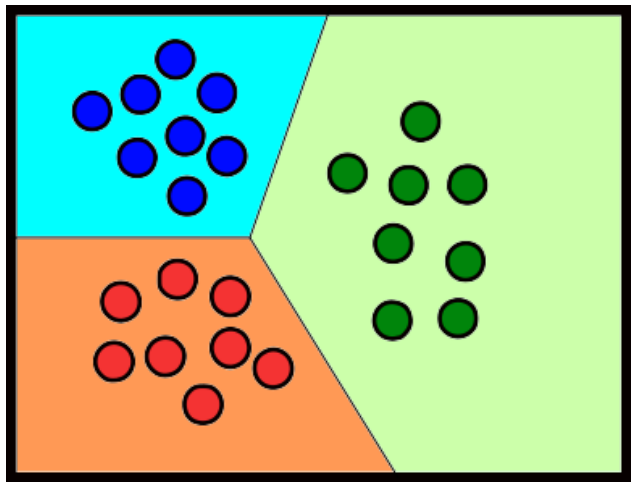
2018 年 1 月 16 日

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 の中からよりよい手法を発見する

目標

- 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
- C++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

- eFCM
- qFCM
- sFCM

比較対象手法 (上記手法にクラスタサイズ調整変数を導入)

- eFCMA
- qFCMA
- sFCMA

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

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

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

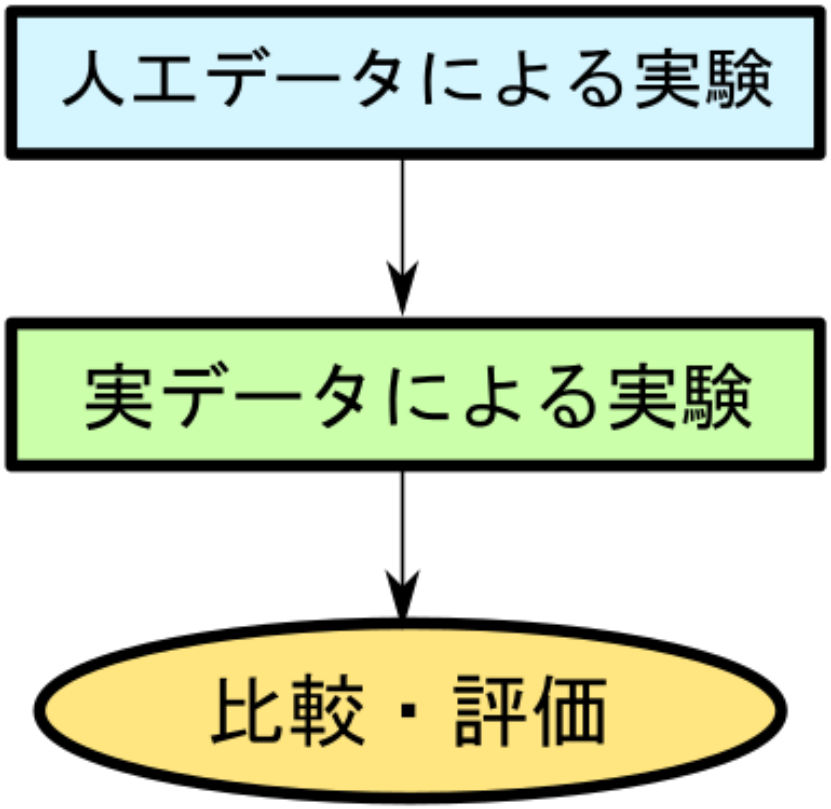
N	データ数	x_k	データ数
C	クラスタ数	v_i	クラスタ中心
m	ファジィ化パラメータ	$u_{i,k}$	帰属度
α	クラスタサイズ調整変数		

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

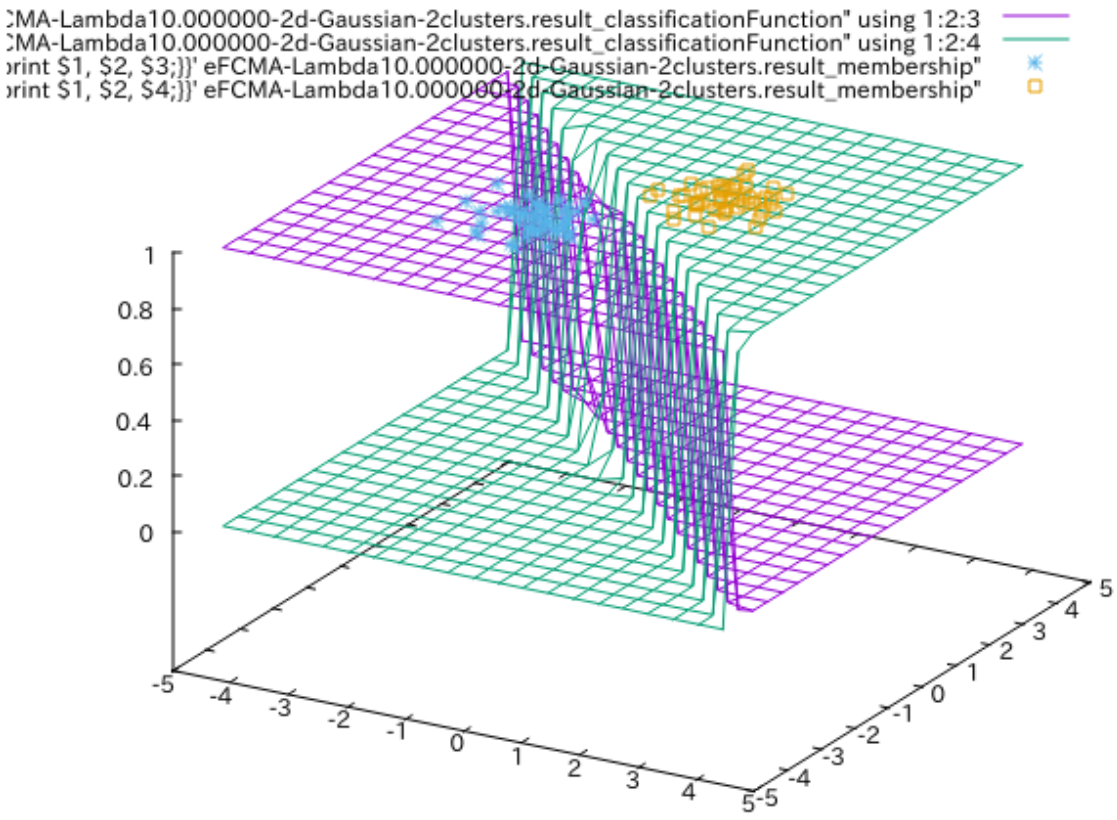
人工データ実験

使用する人工データ

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

実験結果:人工データ

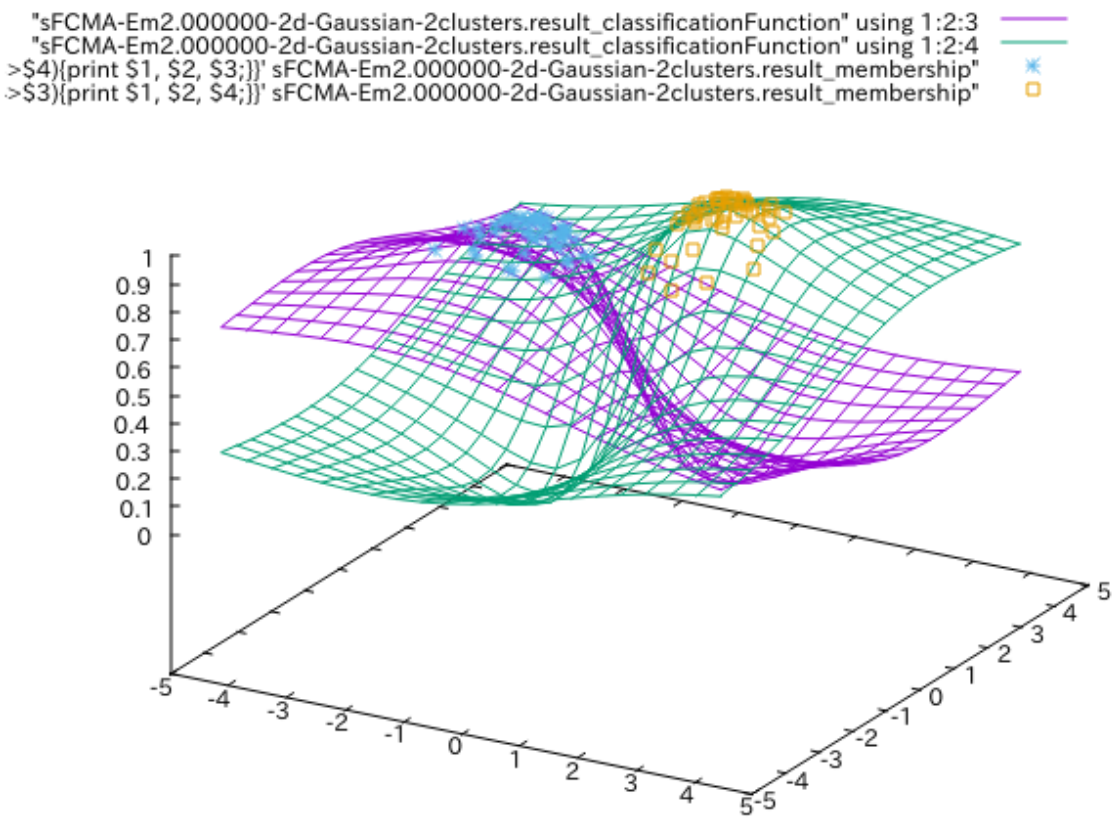
eFCMA



$$\lambda = 10.0$$

実験結果:人工データ

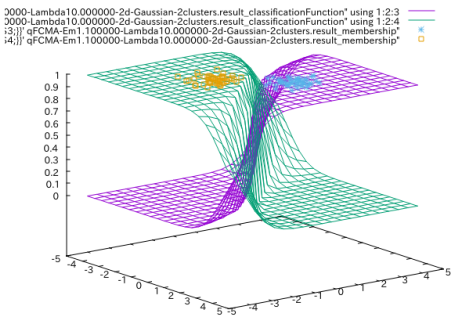
sFCMA



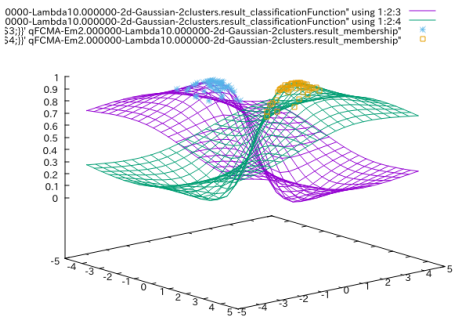
$$m = 2.0$$

実験結果:人工データ

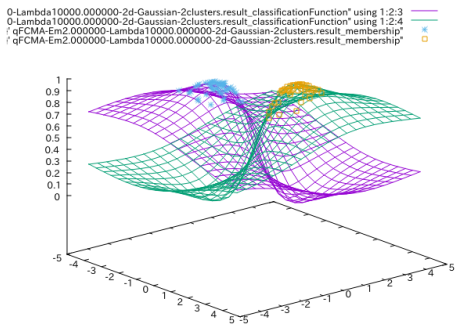
qFCMA



$m = 1.1, \lambda = 10.0$



$m = 2.0, \lambda = 10.0$



$m = 2.0, \lambda = 10000$

実データ実験

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

eFCMA

パラメータ λ を 6.0 から 0.1 刻みで 10.0 まで変化させた際の ARI を導出

qFCMA

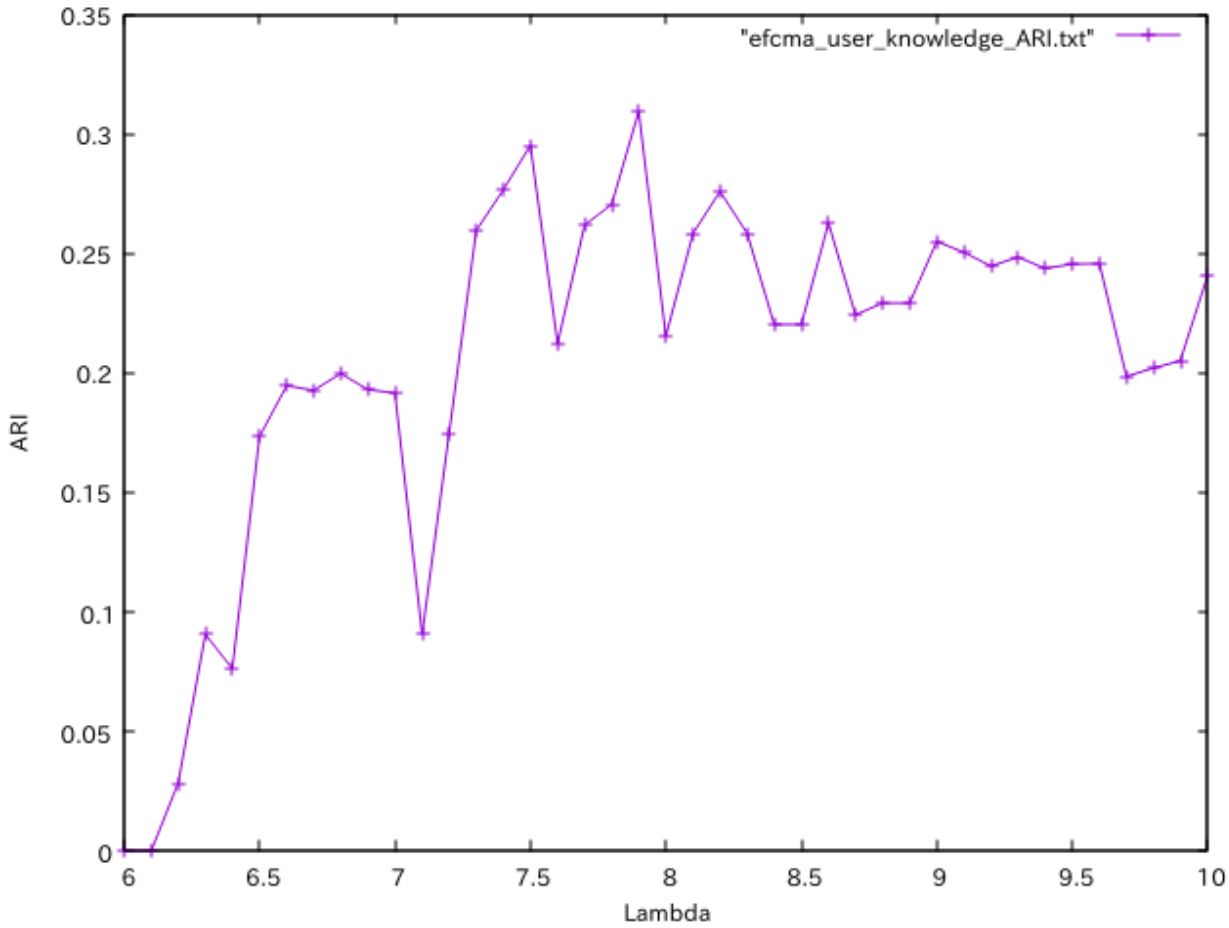
- **パラメータ λ を 10.0 で固定し、パラメータ m を 2.0 から 0.1 刻みで 1.1 まで変化させた際の ARI を導出**
- **パラメータ m を 2.0 で固定し、パラメータ λ50.0 から 10 刻みで 200 まで変化させた際の ARI を導出**

sFCMA

パラメータ m を 3.0 から 0.1 刻みで 1.1 まで変化させた際の ARI を導出

実験結果:実データ

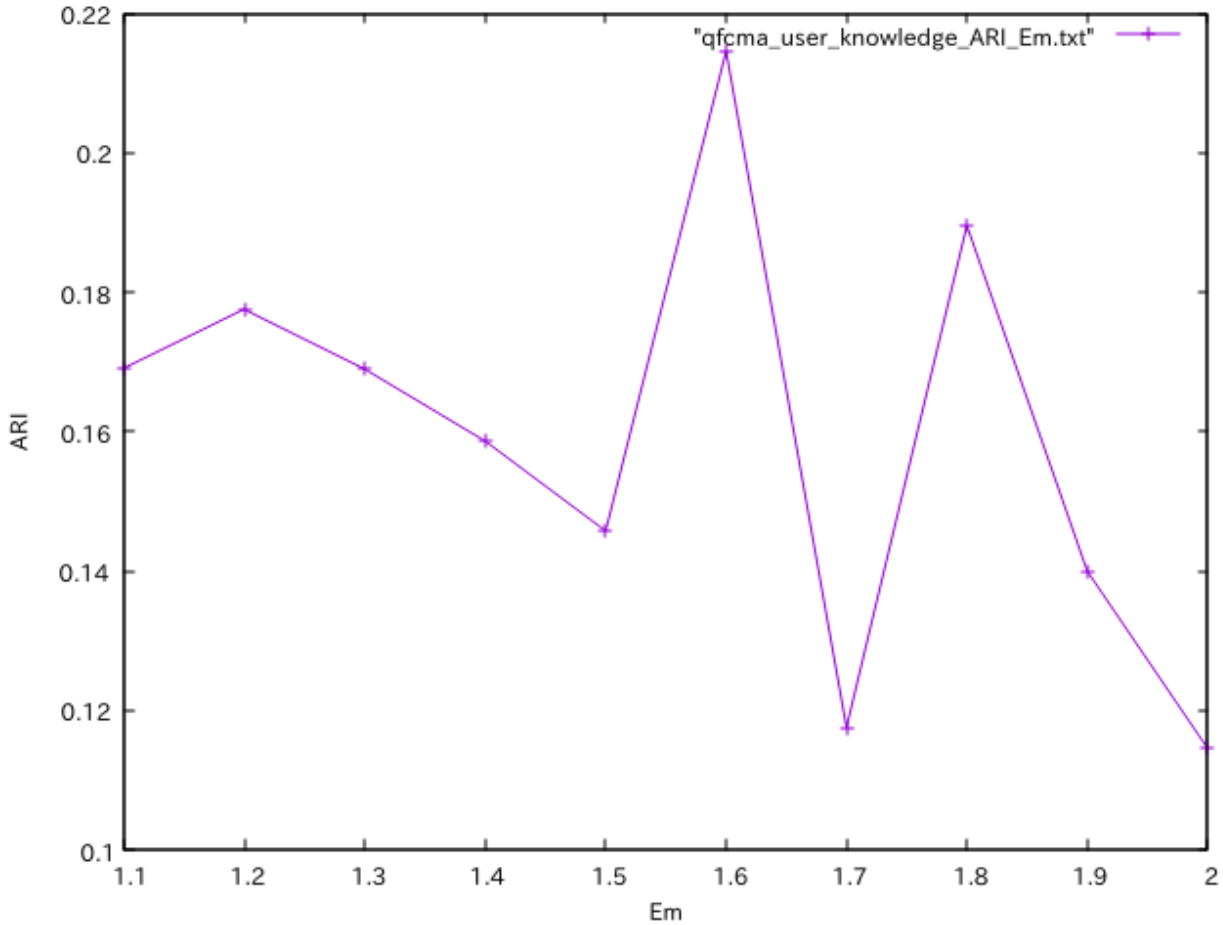
eFCMA



最高 ARI:0.309493

実験結果:実データ

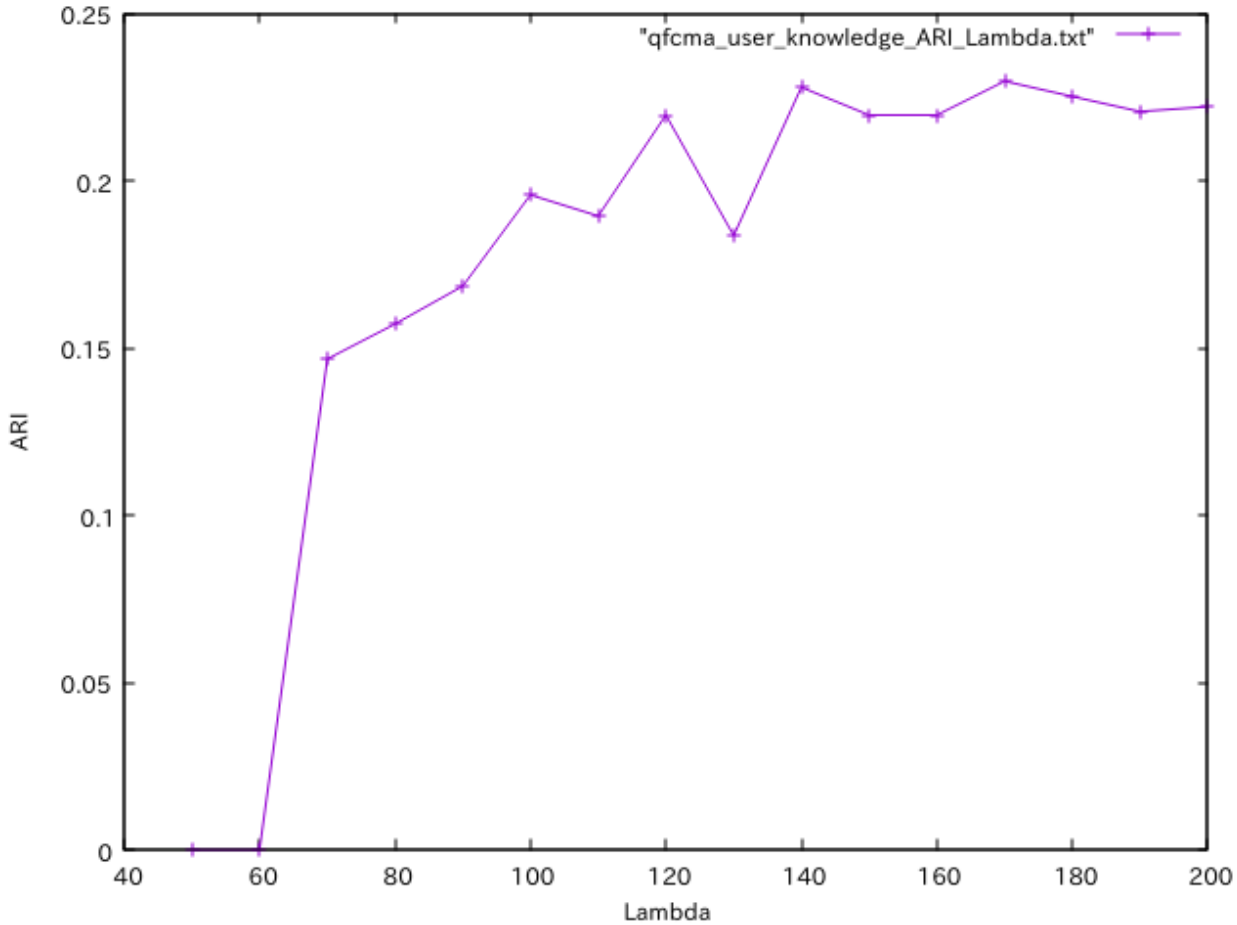
qFCMA



λ=10.0 **最高** ARI:0.229751

実験結果:実データ

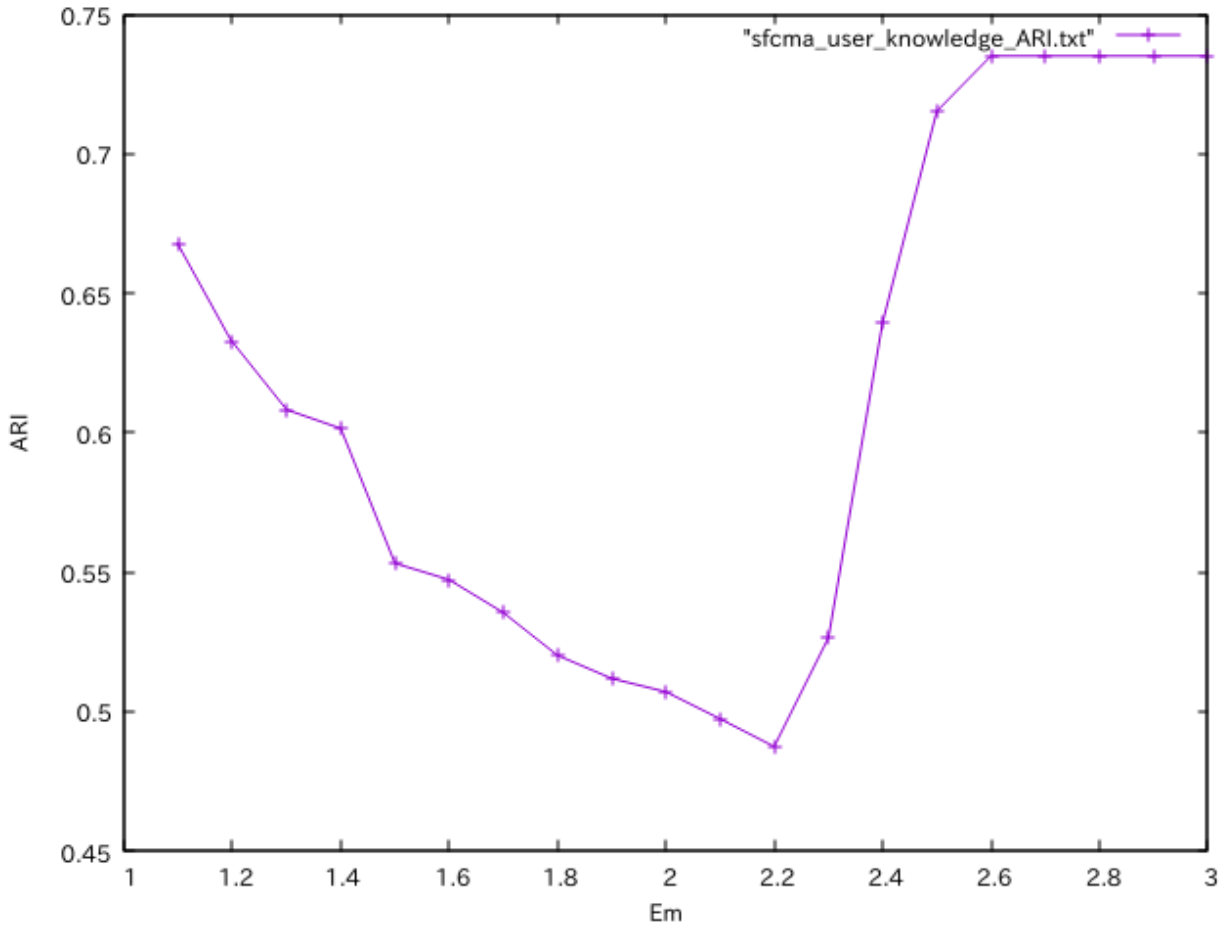
qFCMA



m=2.0 **最高** ARI:0.214619

実験結果:実データ

sFCMA



最高 ARI:0.73515

実験結果

各手法の最高 ARI

eFCMA	0.309493
qFCMA	0.222143
sFCMA	0.73515

sFCMA が最も高評価となった

考察・課題

考察

- sFCMA と eFCMA、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。

課題

- 他の実データでも同様の傾向が現れるかどうかの検証。
- エントロピー項が影響する原因及び理由の調査。

まとめ

目的

- クラスタリング手法の 1 つである Fussy c-means の中からより確実な手法を発見する

目標

- 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
- C++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

- sFCMA が高評価となった

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

補足:eFCMA の更新式

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

補足:qFCMA の更新式

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

補足:sFCMA の更新式

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