

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

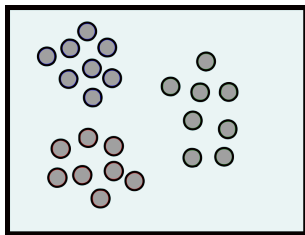
池辺 颯一

芝浦工業大学 工学部 通信工学科

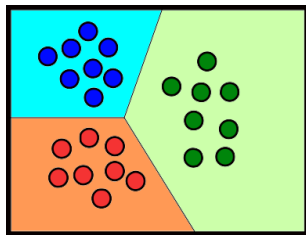
2018 年 1 月 16 日

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



目標



実験対象

提案手法

- eFCMA
- qFCMA
- sFCMA

研究評価方法



クラスタリングの最適化問題

eFCMA

$$\begin{aligned} \underset{u,v,\alpha}{\text{minimize}} \quad & \sum_{i=1}^C \sum_{k=1}^N u_{i,k} \|x_k - v_i\|_2^2 + \lambda^{-1} \sum_{i=1}^C \sum_{k=1}^N u_{i,k} \log\left(\frac{u_{i,k}}{\alpha_i}\right) \\ \text{subject to} \quad & \sum_{i=1}^C u_{i,k} = 1, \sum_{i=1}^C \alpha_i = 1 \text{ and } \lambda > 0, \alpha_i > 0 \end{aligned}$$

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

$$\begin{aligned} & \underset{u,v,\alpha}{\text{minimize}} \sum_{i=1}^C \sum_{k=1}^N (\alpha_i)^{1-m} (u_{i,k})^m \|x_k - v_i\|_2^2 + \frac{\lambda^{-1}}{m-1} \sum_{i=1}^C \sum_{k=1}^N (\alpha_i)^{1-m} (u_{i,k})^m \\ & \text{subject to } \sum_{i=1}^C u_{i,k} = 1, \sum_{i=1}^C \alpha_i = 1 \text{ and } \lambda > 0, m > 1, \alpha_i > 0 \end{aligned}$$

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

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

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		

実験方法

- 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

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

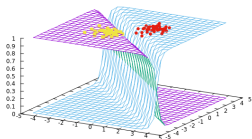
- 平均値 $(-1, -1)$, 標準偏差 $(0.5, 0.5)$ 及び平均値 $(-1, -1)$, 標準偏差 $(0.5, 0.5)$ のガウスサンプリングで生成
- データ数: 100
- クラス数: 2

人工データ実験のアルゴリズム

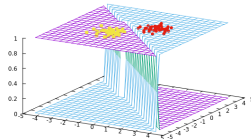
- ① クラスタ中心 V をランダムに与える。
- ② V を用いて帰属度 U を更新する。
- ③ U を用いて V 及びクラスタサイズ調整変数 A を更新する。
- ④ 収束条件すれば終了し、そうでない場合は2に戻る。

実験結果:人工データ

eFCMA



$\lambda = 1$



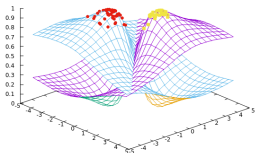
$\lambda = 10000$

eFCMA の特徴

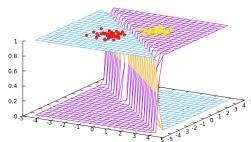
パラメータ λ を無限大に近づけるほど HCM に近づく

実験結果:人工データ

sFCMA



$m = 2$



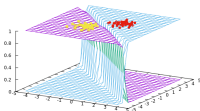
$m = 1.01$

sFCMA の特徴

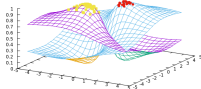
パラメータ m を 1 に近づけるほど HCM に近づく

実験結果:人工データ

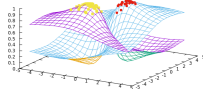
qFCMA



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



$$m = 2, \lambda = 10$$

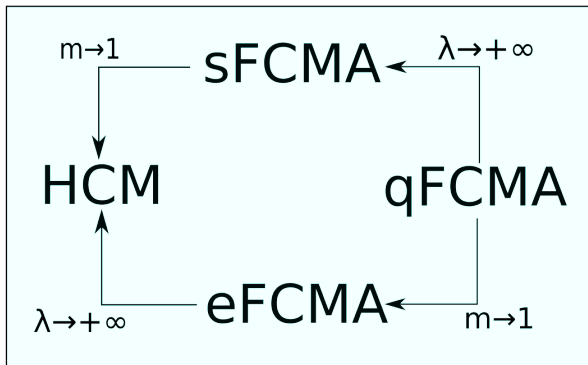


$$m = 2, \lambda = 10000$$

qFCMA の特徴

- パラメータ λ を無限大に近づけると sFCMA に近づく
- パラメータ m を 1 に近づけると eFCMA に近づく

各手法間の関係



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 5.0 から 0.1 刻みで 10.0 まで変化させる

qFCMA

- パラメータ λ を 5.0 から 0.1 刻みで 10.0 まで変化させる
- パラメータ m を 2.0 から 0.01 刻みで 1.01 まで変化させる

sFCMA

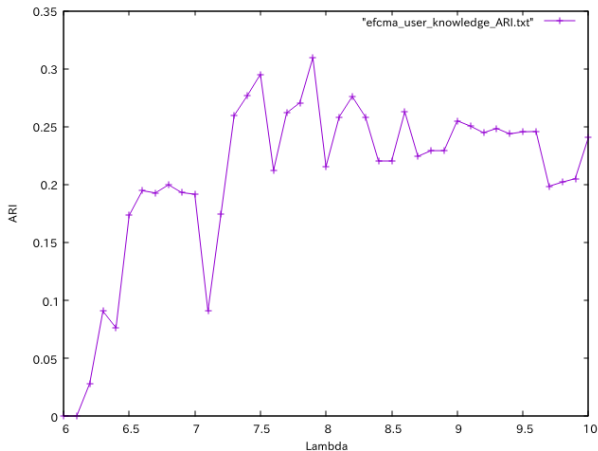
パラメータ m を 2.0 から 0.01 刻みで 1.01 まで変化させる

実データ実験のアルゴリズム

- ① 正解帰属度を用いて U を初期化する。
- ② U を用いてクラスタ中心 V 及びクラスタサイズ調整変数 A 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:実データ

eFCMA



最高 ARI:

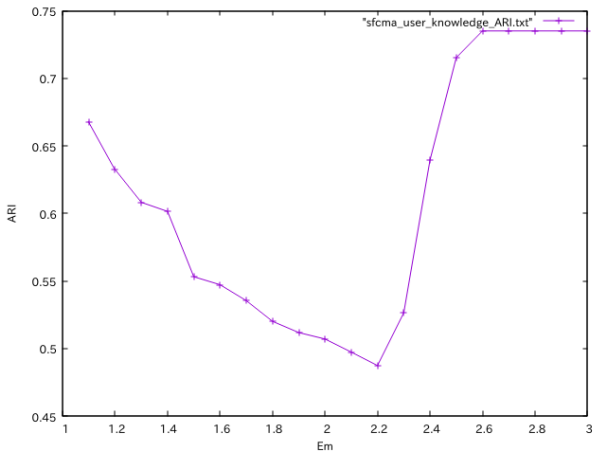
実験結果:実データ

qFCMA

最高 ARI:

実験結果:実データ

sFCMA



最高 ARI:

実験結果

各手法の最高 ARI

eFCMA		$\lambda =$
qFCMA		$\lambda = , m =$
sFCMA		$m =$

評価

sFCMA が最も高評価

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

-

目標

-

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

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

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

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