

通信ゼミナール

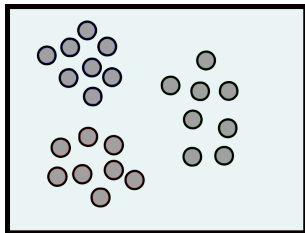
クラスタリング手法の評価

池辺 颯一

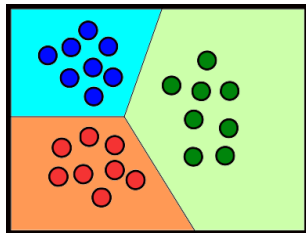
芝浦工業大学 工学部 通信工学科

2018 年 12 月 12 日

- 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

- 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

既存手法

- sFCM
- pFCM
- eFCM

比較対象手法 (上記手法にクラスサイズ調整変数を導入)

- sFCMA
- pFCMA
- eFCMA

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		

qFCMA

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

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		

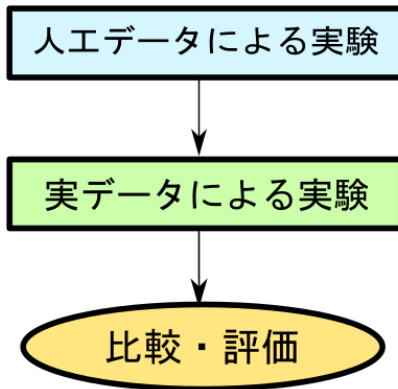
sFCMA

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

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		

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
- ARI の値が高いほど高評価

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

eFCMA

$$\lambda = 10.0$$

qFCMA

$$m = 1.1, \lambda = 10.0$$

$$m = 2.0, \lambda = 10.0$$

$$m = 2.0, \lambda = 10000$$

sFCMA

$$m = 2.0$$

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)

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

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

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 λ 50.0 から 10 刻みで 200 まで変化させた際の ARI を導出

eFCMA

最高 ARI:0.309493

qFCMA

$\lambda=10.0$ 最高 ARI:0.229751

qFCMA

m=2.0 最高 ARI:0.214619

sFCMA

最高 ARI:0.73515

各手法の最高 ARI

eFCMA	0.309493
qFCMA	0.222143
sFCMA	0.73515

考察



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



考察



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

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

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