クラスタリング手法の評価に向けて

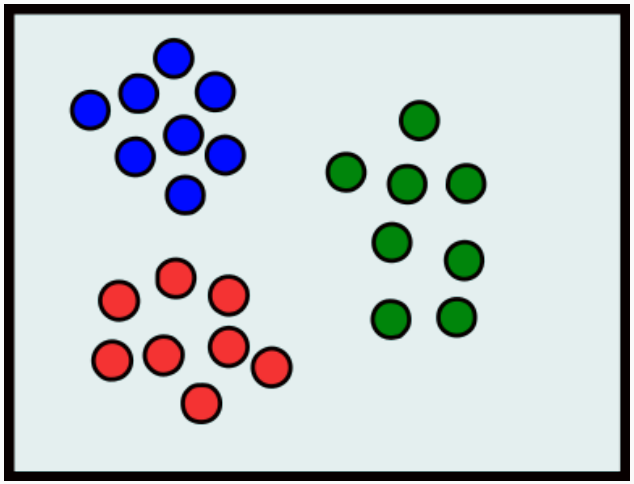
池辺 颯一

2018 年 12 月 15 日

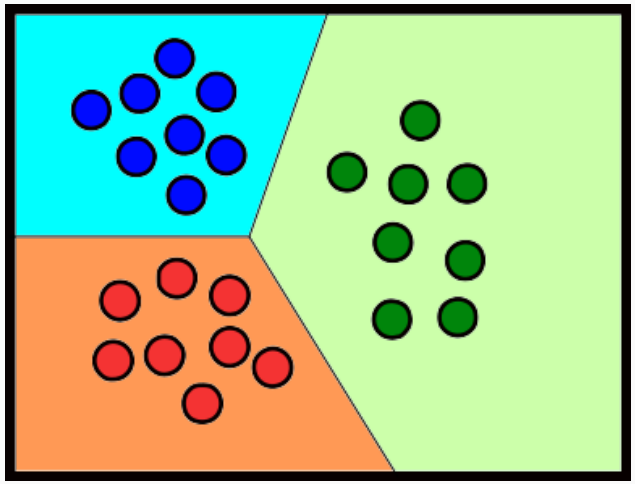
芝浦工業大学

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
- 機械学習における教師なし学習



クラスタリング前



クラスタリング後

目的

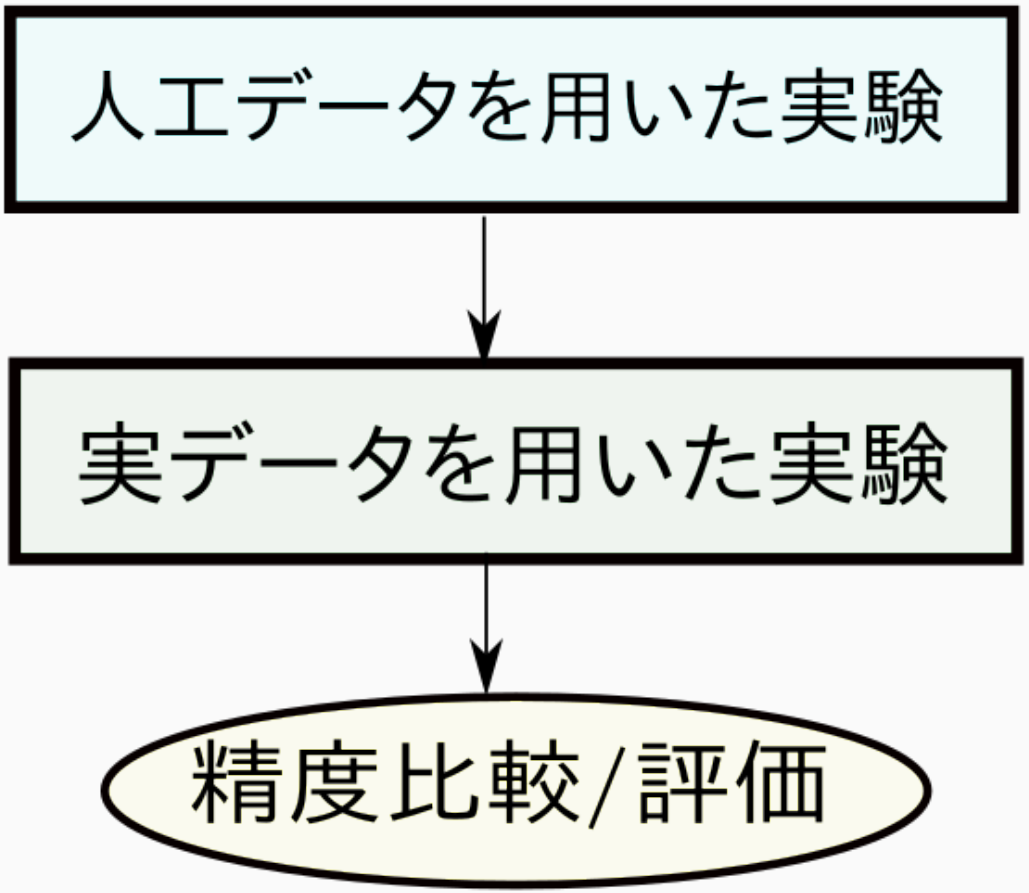
目的

- FCM の基本的な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

実験方法



実験評価方法

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
- マイナスの値はランダムの期待値を下回る
- ARI の値が高いほど高評価

クラスタリング手法

eFCMA

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

$$d_{i,k} = \|x_k - v_i\|_2^2,$$

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

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

クラスタリング手法

qFCMA

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

$$d_{i,k} = \|x_k - v_i\|_2^2,$$

$$u_{i,k} = \frac{\alpha_i(1+\lambda(1-m)\|x_i - v_k\|_2^2)^{\frac{1}{1-m}}}{\sum_{j=1}^{C}\alpha_j(1+\lambda(1-m)\|x_j - v_k\|_2^2)^{\frac{1}{1-m}}},$$

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

クラスタリング手法

sFcma

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

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

$$d_{i,k} = \|x_k - v_i\|_2^2 = \left(\sqrt{\sum_{\ell=1}^{m} (x_{k,\ell} - v_{i,\ell})^2} \right)^2 ,$$

$$u_{i,k} = \frac{1}{\sum_{j=1}^{c} \frac{\alpha_j}{\alpha_i} \left(\frac{d_{j,k}}{d_{i,k}} \right)^{\frac{1}{1-m}}}, \quad v_i = \frac{\sum_{k=1}^{n} (u_{i,k})^m x_k}{\sum_{k=1}^{n} (u_{i,k})^m},$$

使用する実データ

Yeast Data Set

実験結果

考察

-

まとめ

-