4U-03

k-meansのための初期値外れ値を考慮した初期値設定手法の提案

田口　隼平　　西垣　貴央　　小野田　崇
青山学院大学大学院理工学研究科

1　はじめに

昨今、データマイニングの重要性が高まっている。クラスタリングはデータマイニングの一般的な手法であり、k-means はクラスタリングの一般的な方法である。k-means は各データを最も近いクラスタ中心に所属させてクラスタを生成する方法である。k-means を行うにあたって、クラスタ中心の初期値を設定する必要がある。通常の k-means はクラスタ数だけデータからランダムに選択し、それをクラスタ中心の初期値に設定している。k-means はクラスタ中心の初期値に依存してクラスタを生成するため、k-means はクラスタリング結果が毎回異なるという課題を抱えている。k-means の初期値設定の研究は k-means++や KKZ など盛んに行われているが，初期値を一意に決定でき，かつ初期値外れ値問題を解決する設定手法は提案されていない。本稿では，k-means の初期値設定において，初期値を一意に決定し，かつ初期値外れ値問題を解決する手法を提案する。

2　関連研究

KKZ は初期値を一意に設定できる初期値設定であり、最も離れたデータ同士となるデータをクラスタ中心の初期値に設定する [1]。KKZ のアルゴリズムを説明する。

1. クラスタ中心の初期値の集合 $\mathbb{C} = \phi$ を作成する。
2. クラスタの番号 $k = 1$ を作成する。
3. 各データ $\boldsymbol{x_i}$ のノルムを計算する。ノルムが最大となるデータをクラスタ中心の初期値 $\boldsymbol{c_k}$ とし、$\boldsymbol{c_k}$ を $\mathbb{C}$ に所属させる。
4. クラスタ番号 k に 1 を加える。
5. 各データ $\boldsymbol{x_i}$ に対して $\mathbb{C}$ との最短距離を計算し、値が最も大きくなったデータ $\boldsymbol{x_i}$ をクラスタ中心の初期値 $\boldsymbol{c_k}$ とし、$\boldsymbol{c_k}$ を $\mathbb{C}$ に所属させる。
6. k が任意のクラスタ数 q に達していれば、終了する。達していなければステップ 4 に戻る。

A proposal for generation of initial value except for outlier for k-means
Jumpei Taguchi　Takahiro Nishigaki　Takashi Onoda
Aoyamagakuin University

KKZ は初期値外れ値問題という課題を抱えている。初期値外れ値とは、そのデータをクラスタ中心の初期値に選択してしまうと、クラスタ内のデータ同士の類似度が小さいクラスタを生成してしまうデータである。KKZ は初期値外れ値を選択する可能性が非常に高い。本報告では初期値外れ値問題を解決できる初期値設定を提案する。

3　提案手法

提案する手法は初期値を一意に設定でき、かつ、初期値外れ値を避けたデータの中で KKZ を行う初期値設定である。初期値外れ値の判断には MT 法の考えを用いる。MT 法はマハラノビス距離を適用してパターン認識や予測を行う手段であり、異常データの判断などに用いられている [2]。MT 法は全データの平均からのマハラノビス距離の二乗が特定のしきい値より離れたデータを異常データと判断している。今回は [3] と同様にカイ二乗分布を使ってしきい値を決定する。MT 法はマハラノビス距離を利用しているため、多次元データを統合した 1 つの尺度でデータを測れる。そのため、属性同士の関係を考慮して異常データの判断を行えるという利点を持つ。次に、提案手法のアルゴリズムを説明する。

1. i 個あるデータ $\boldsymbol{x_i}$= $(x_{i1}, ..., x_{im})$ を標準化し、標準化したデータを $\boldsymbol{z_i}$= $(z_{i1}, ..., z_{im})$ とする。標準化は式 1 のように行う。$j \in (1, ..., m)$ は属性である。$\overline{x_j}$ は平均、σ_j は標準偏差である。

$$z_{ij} = \frac{(x_{ij} - \overline{x_j})}{\sigma_j} \tag{1}$$

2. 全データの平均と各データのマハラノビス距離の二乗を求める。各データのマハラノビス距離の二乗 MD_i^2 の式は式 (2) である。$\boldsymbol{R}^{-1}$ はデータ同士の相関行列 $\boldsymbol{R}$ の逆行列である。

$$MD_i^2 = \boldsymbol{z_i}^T \boldsymbol{R}^{-1} \boldsymbol{z_i} \tag{2}$$

3. MD_i^2 の値が、カイ二乗分布で全体データの 95.45% が正常データと考えた場合の 4.55% 時の値、もしくは全体データの 99.73% が正常データと考え

た場合の 0.27% 時の値を超えたデータ $\boldsymbol{x}_i$ は初期値外れ値とみなし、初期値に設定しない。

4. 初期値に設定しないデータ以外で KKZ のアルゴリズムを行う。

4 実験条件

人工データと実データに KKZ、提案手法（しきい値は 0.27% 時の値）と通常の k-means の初期値設定の random で実験を行う。人工データは平均が $\boldsymbol{\mu}_1 = (-2,-1), \boldsymbol{\mu}_2 = (2,1)$ で相関係数が 0.7 の 339 個と 344 個という 2 つの集団からなるデータ集合である。そのデータ集合には全体の平均から 2σ から 3σ の範囲で離れた 30 個のデータと全体の平均から 3σ を超えて離れたデータ 4 個が含まれている。このデータから 2 個のクラスタを生成する。実データとして、UCI レポジトリの abalone データを使用する。属性は 3 個で、データ数は 4177 個である。このデータから 3 個のクラスタを生成する。今回の提案手法のしきい値は $\chi^2(0.0027, 2) = 11.829$ と $\chi^2(0.0027, 3) = 14.156$ である。評価には以下の内的結合と外的分離の指標を用いる。内的結合 V_{in} は式 (3) で求められ、値が小さいほどクラスタ内のデータ同士の類似度が大きいことを意味する。データ数を n、クラスタ数を q とし、データを $\boldsymbol{x}_i, i \in (1,...,n)$、クラスタ中心を $\boldsymbol{c}_k, k \in (1,...,q)$、クラスタを $C_k, k \in (1,...,q)$ とする。$1[\boldsymbol{x}_i \in C_k]$ は $\boldsymbol{x}_i$ が C_k に所属しているなら 1、所属していないなら 0 を意味する。クラスタ中心 $\boldsymbol{c}_k$ の式は式 (4) である。$|C_k|$ はクラスタ k のデータ数を表す。

$$V_{in} = \sum_{k}^{q}\sum_{i}^{n}(\boldsymbol{c}_k - \boldsymbol{x}_i)^2 1[\boldsymbol{x}_i \in C_k], \boldsymbol{c}_k = \frac{1}{|C_k|}\sum_{\boldsymbol{x}_i \in C_k}\boldsymbol{x}_i \quad (3)$$

$$\boldsymbol{c}_k = \frac{1}{|C_k|}\sum_{\boldsymbol{x}_i \in C_k}\boldsymbol{x}_i \quad (4)$$

外的分離 V_{out} は式 (5) で求められ、値が大きいほど生成されたクラスタが他のクラスタとの類似度が小さいことを意味する。$\boldsymbol{\Delta}$ は全てのクラスタ中心の重心を意味する。

$$V_{out} = \sum_{k=1}^{q}(\boldsymbol{\Delta} - \boldsymbol{c}_k)^2, \quad \boldsymbol{\Delta} = \frac{1}{q}\sum_{k=1}^{q}\boldsymbol{c}_k \quad (5)$$

random は毎回異なるクラスタを生成するため 10 回繰り返したときの最大と最小の内的結合と外的分離の値を評価する。

5 実験結果

図 1 は人工データの KKZ と提案手法のクラスタリング結果である。KKZ は最も離れたデータ同士のデータをクラスタ中心の初期値に設定しているため、最終的なクラスタ中心同士も離れて外的分離は 11.6 と大きい値であった。しかし、KKZ では内的結合が 6128 と大きくなっている。提案手法は初期値外れ値を避けてクラスタ中心の初期値を設定しているので、外的分離は 4.78 と KKZ より小さいが内的結合は 3283 と KKZ より小さくなった。random の最大内的結合と外的分離は 6128、11.6 で最小内的結合と外的分離は 3283 と 4.78 であった。abalone データでも同じ傾向が見られた。外的分離は提案手法より KKZ の方が大きくなり、内的結合は KKZ より提案手法の方が小さくなった。

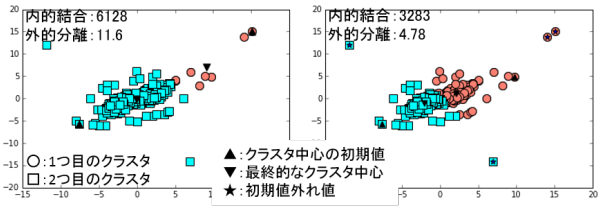

図 1: クラスタリング結果 : 左が KKZ、右が提案手法。

6 まとめと今後の課題

k-means の初期値設定方法である KKZ は初期値外れ値を選んでしまいクラスタの内的結合を大きくしてしまう問題が存在していた。そこで本稿では初期値外れ値を選択しない初期値設定手法を提案した。提案手法を、人工データと実データに適用した。その結果、提案手法は初期値を一意に設定でき、かつ、クラスタの内的結合を KKZ より小さくすることができた。今後、様々なデータに提案手法の有効性を検証するために提案手法を分析する必要がある。

参考文献

[1] I Katsavounidis, C.-C. J. K. a. B. Z, ”A New Initialization Technique for Generalized Lloyd Iteration” ,IEEE SIGNAL PROCESSING LETTERS, 1994.

[2] 立林和夫・手島昌一・長谷川良子, ” 入門 MT システム”, 日科技連出版社, 2008 ’

[3] 兼高達貳,” マハラノビス汎距離の応用例 特殊健康診断の事例”,DREG 研究報告-医療のための実験計画 [1],1987