The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

MT 法を用いた k-means 法の初期値設定法の提案

Proposal of seeding methods for k-means using Mahalanobis Taguchi Method

西垣 貴央 *1　田口 隼平 *1　小野田 崇 *1
Takahiro NISHIGAKI　Jumpei TAGUCHI　Takashi ONODA

*1 青山学院大学 理工学研究科
Department of Science and Engineering, Aoyama Gakuin University

The k-means method is a generally used clustering technique. The method is widely used for clustering to a large number of data because of its simplicity and speed. However, the clustering result depends heavily on the chosen initial clustering centers, which are chosen uniformly at random from the data points. In addition, the clustering result is not good, if an outlier for initial clustering centers is selected at initial clustering centers. Many methods have been proposed to solve these two problems. However, there is no way to solve these two problems at the same time. We propose a seeding method based on the Mahalanobis Taguchi System and KKZ for k-means clustering method. We evaluate the performance of our proposed method and compare it with other seeding methods by using benchmark datasets.

1. はじめに

近年のインターネット利用の一般化や高性能デバイスの普及に伴って，膨大な情報が氾濫している．例えば，ニュース，製品情報，レストラン情報などの Web ページや，個人所有のハードディスクドライブ (HDD) には旅行で撮った写真や動画など，様々な情報が様々な場所に溢れている．この溢れる情報の中からユーザが欲しいときに欲しい情報を探し出すことができれば，ユーザは様々な作業を効率的に進めることができる．膨大なデータからユーザが欲しい情報を探し出す方法として，キーワード検索やグループ化された情報の探索が利用されている．

グループ化された情報の探索とは，類似した内容という視点から，グループ化された情報からユーザの欲しい情報を探し出す方法である．例えば，Yahoo!のニュースのカテゴリの下に収集された情報が，グループ化された情報に該当する．グループ化することで，膨大なデータを類似した内容の情報ごとに整理でき，情報の概観を把握しやすくできる．グループ化された情報の検索は，近年注目されている方法である．この方法では，類似した情報ごとにグループを作成する必要がある．しかし Web ページのような膨大なデータに対する人手でのグループ化は実質不可能である．一般に，膨大なデータに対するグループ化にはクラスタリング手法が用いられ，計算機による自動グループ化が行われている [元田 06]．

クラスタリング手法とは，Web ページなどのテキストや写真のような画像データを，自動的にグループ化する一種の教師なし学習法である．このクラスタリングには，クラスタとデータとの類似度を測るある種の評価関数を用いることで，直接データをクラスタと呼ばれるグループに分割する．この方法は計算コストが比較的小さいため，Web ページなどの大規模なデータのクラスタリングに適している．一般に，クラスタリング手法として k-means 法 [MacQueen 67] が多く用いられている．この k-means 法は，クラスタ内の中心点（クラスタ中心）とクラスタ内のデータ間の二乗距離の総和が最小となるようにクラスタ中心を逐次的に求めることでクラスタを生成する方法である．この方法は，簡単なアルゴリズムのため，データマイニングや画像処理などの様々な分野の研究でよく用いられている．しかし k-means 法には，クラスタリング結果がランダムに選択される初期のクラスタ中心に依存（初期値依存）してしまうという問題や，初期値として選択したデータによってはクラスタ内のデータ間の二乗距離の総和が非常に大きなクラスタを生成してしまう可能性がある（初期値外れ値）という問題がある．これらの問題を解決するために KKZ 法 [Katsavounidis 94] や k-means++法 [Arthur 07] など様々な方法が提案されている．しかし，これらの方法では初期値依存や初期値外れ値の問題の両方を同時に解決することができない．本稿では，この初期値依存や初期値外れ値の問題の両方を同時に解決することができる k-means 法の新たな初期値設定の方法を提案する．

連絡先: 西垣 貴央，青山学院大学理工学部経営システム工学科，〒 252-5258 神奈川県相模原市中央区淵野辺 5-10-1, nishigaki@sei.aoyama.ac.jp

2. 関連研究

k-means 法の初期値設定の方法の研究は数多く行われているが，本稿では主要な研究として Katsavounidis らによって提案された KKZ 法 [Katsavounidis 94]，David Arthur によって提案された k-means++法 [Arthur 07] について説明する．最初に k-means 法のアルゴリズムについて述べ，その後既存の 2 つの初期値設定法について述べる．

2.1 k-means 法

k-means 法はクラスタリング手法として最も広く使われる手法の一つである．この手法は，式 (1) の評価関数 ϕ を最小化するクラスタ中心を見つけることによって，データ $\mathbf{X}$ を任意の k 個のクラスタに分割する．

$$\phi = \sum_{\mathbf{x}_j \in \mathbf{X}} \min_{i \in k} \|\mathbf{x}_j - \mathbf{c}_i\|^2 \tag{1}$$

$\mathbf{x}_j, j \in \{1, \cdots, n\}$ は各データ，n はデータの総数を示す．また，$\mathbf{c}_i$ はクラスタ $i \in \{1, \cdots, k\}$ の中心である．つまり，k-means 法は，各データ点から最も距離が近いクラスタ中心との距離の総和が，最小となるようなクラスタ中心を求めることで，クラスタリングを行う．この手法のアルゴリズムを以下に示す．

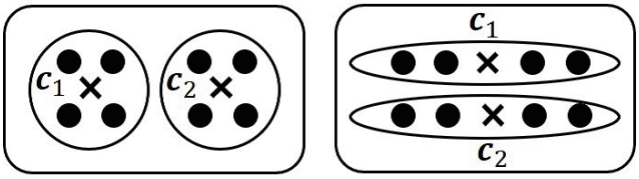

図 1: 初期値によってクラスタリングが変わってしまう初期値依存問題

1. 任意の k 個のクラスタ中心 $\mathbf{c}_i$ をランダムに選択する．
2. すべてのデータを，各データ点 $\mathbf{x}_j, j \in \{1, \cdots, n\}$ から最も近いクラスタ i に割り当てる．
3. 各クラスタごとに，式 (2) にしたがってクラスタ中心を求める．

$$\mathbf{c}_i = \frac{1}{|\mathcal{C}_i|} \sum_{\mathbf{x}_j \in \mathcal{C}_i} \mathbf{x}_j \tag{2}$$

   $\mathcal{C}_i$ は各クラスタ i に含まれるデータ集合であり，$|\mathcal{C}_i|$ はクラスタ $\mathcal{C}_i$ に含まれるデータ数である．
4. クラスタに変化がなくなるまで，ステップ 2, 3 を繰り返す．

k-means 法は，適切なクラスタ中心を求めて分割するという簡単なアルゴリズムであり，よく用いられている．しかし，k-means 法にはランダムに決定されるクラスタ中心の初期値にクラスタリングの結果が依存してしまうという問題がある．このクラスタ中心の初期値にクラスタリング結果が依存してしまう例を図 1 に示す．左右とも同じデータが与えられ，それぞれに k-means 法を用いた場合のクラスタリング結果を示している．この図が示すように，k-means 法では初期に選択されたクラスタ中心によって，最終的に得られるクラスタ中心が変わってしまう．クラスタリングには結果を評価する明確な指標がないため，複数のクラスタリング結果から適切な結果を見つけ出すことは容易ではない．

2.2 KKZ 法

KKZ 法は，Katsavounidis らによって提案され，初期のクラスタ中心として，最も離れているデータ同士をクラスタ中心の初期値として選択する手法である．具体的には，逐次的にクラスタ中心 $\mathbf{c}_i \in \{\mathbf{c}_1, \cdots, \mathbf{c}_k\}$ を選んでいく過程で，既存のクラスタ中心から最も遠いデータを選択する．この手法のアルゴリズムを以下に示す．

a. 与えられたデータ $\mathbf{X}$ からデータ同士の距離が最大となる 2 つのデータを，初期値 $\mathbf{c}_1, \mathbf{c}_2$ に設定する．

b. 全データに対して $D(\mathbf{x}_j), j \in \{1, \cdots, n\}$ を求める．$D(\mathbf{x}_j)$ はデータ点 $\mathbf{x}_j$ とすでに決定されたクラスタ中心との最短距離を表す (図 2)．

c. 最大となる $D(\mathbf{x}'_j)$ のデータ $\mathbf{x}'_j$ を次のクラスタ中心 $\mathbf{c}_l$ に選択する．

d. クラスタ中心を k 個選ぶまで，ステップ 2,3 を繰り返す．k 個選択した後，k-means 法アルゴリズムのステップ 2, 4 と同様の処理を行う．

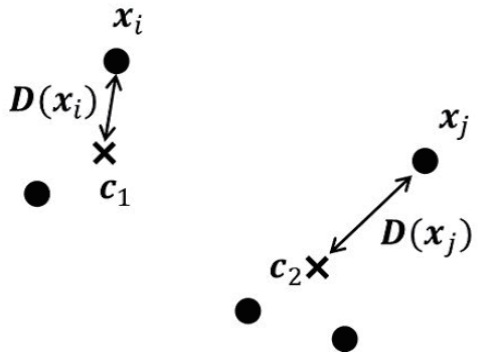

図 2: $D(\mathbf{x})$:各データ点から最も近い既存のクラスタ中心との距離

この手法は，入力順序に依存せず，クラスタ同士の距離が離れたクラスタリング結果を得ることができるが，初期値として選択したデータによってはクラスタ内のデータ間の二乗距離の総和が非常に大きなクラスタを生成してしまう初期値外れ値の問題がある．

2.3 k-means++法

k-means++法は，David Arthur によって提案され，最も注目されている k-means 法の初期値設定法である．この方法は，KKZ 法の初期値外れ値に弱い特性を改良したものである．具体的には，KKZ 法では $D(\mathbf{x}_j)$ の値が最も大きいデータを選択するが，k-means++法では必ずしも最も大きな $D(\mathbf{x}_j)$ となるデータが選択されるわけではない．k-means++法のアルゴリズムを以下に示す．

a. 一つ目のクラスタ中心 $\mathbf{c}_1$ をデータ $\mathbf{X}$ からランダムに選ぶ．

b. 全データに対して $D(\mathbf{x}_j), j \in \{1, \cdots, n\}$ を求める．

c. 次式を満たす実数値 L をランダムに求める．

$$0 \leq L \leq \sum_{j=1}^{n} D(\mathbf{x}_j)^2 \tag{3}$$

d. 次式を満たす $\mathbf{x}'_j$ を次のクラスタ中心 $\mathbf{c}_l$ に選択する．

$$\sum_{j=1}^{l-1} D(\mathbf{x}'_j)^2 \leq L \leq \sum_{j=1}^{l} D(\mathbf{x}'_j)^2 \tag{4}$$

e. クラスタ中心を k 個選ぶまで，ステップ 2,3,4 を繰り返す．k 個選択した後，k-means 法アルゴリズムのステップ 2, 4 と同様の処理を行う．

式 (4) からわかるように，$D(\mathbf{x}_j)$ が大きいデータほど次のクラスタ中心に選ばれる可能性が高い．つまりこの手法では，すでに決定されたクラスタ中心からより遠いデータ点をクラスタ中心と決定することができる．このため，クラスタ同士を話すことができ，式 (4) 中の L がランダムに選択されることから，必ず最も遠いデータ点を選択するわけではないので，初期値外れ値に弱い問題に対処している．しかし，上述したアルゴリズムからわかるように，k-means++法は最初のデータ点をランダムに選択するなど初期値依存の問題は残っている．

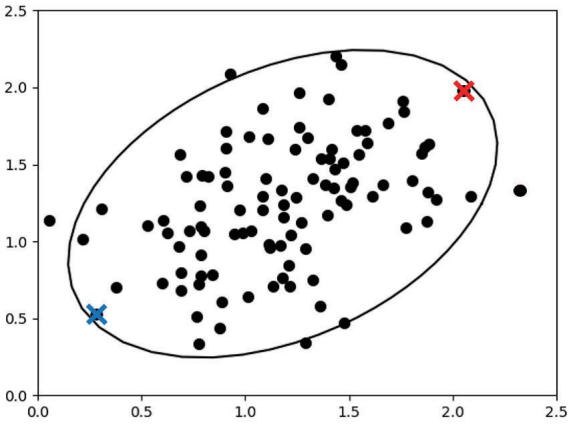

図 3: 楕円より外のデータは初期値外れ値として，それ以外から初期クラスタ中心を選択

3. 提案する初期値設定方法：KKZ-EX

本節では，MT 法に基づく k-means 法の初期値設定方法について述べる．この手法では，データ集合が正規分布していると仮定することで，MT 法 [長谷川 06] を使用して初期値外れ値を求める．その後，初期値外れ値とされたデータ以外を用いて最も離れたデータ同士を初期値として設定する．MT 法は異常検知手法の一つであり，データ集合の平均からのマハラノビス距離の二乗があるしきい値を超えたデータを以上と判断する方法である．MT 法でデータが異常であるかの判断をするためのしきい値の決定方法として，χ^2 分布が用いられることが多く [兼高 87]，本提案手法でも χ^2 分布を用いる．しきい値の χ^2 分布の自由度と有意確率を決める．自由度には適用するデータの属性数を，有意確率には，今回 4.55% を用いた．その理由は，すべてのデータが正規分布であると仮定すると 2σ 区間に収まる確率は 95.45% となる．そのため，初期値外れ値となる確率を $100\% - 95.45\% = 4.55\%$ とした．図 3 にその様子を示す．たくさんの黒い丸で示されているものがデータ点である．黒い大きな楕円で表現されているものが全データ点の中心からマハラノビス距離で $\chi^2(2, 0.0455) = 6.18$ を示している．その楕円の外にあるものが，提案手法により初期値外れ値と選択され，初期値として選択されない．初期値として選択されるのは楕円の内側に存在するデータで最も離れたデータ同士が選択される．図 3 中では，バツ印で示されたものである．以下にアルゴリズムを示す．

a. 全てのデータ集合 $\mathbf{X}$ の平均から各データのマハラノビス距離の二乗を求める．

b. 求めたマハラノビス距離の二乗が χ^2 分布の 4.55% 時の値を超えたデータ $\mathbf{x}_m$ を初期値外れ値とみなし，k-means 法の初期値として選択しない．

c. ステップ 2 で選択しないとしたデータ以外の全てのデータ $\mathbf{X} \setminus \mathbf{x}_m$ の距離が最大となる 2 つのデータを，初期値 $\mathbf{c}_1, \mathbf{c}_2$ に設定する．

d. 残った全データに対して $D(\mathbf{x}_j), j \in \{1, \cdots, n\}$ を求める．$D(\mathbf{x}_j)$ はデータ点 $\mathbf{x}_j$ とすでに決定されたクラスタ中心との最短距離を表す．

e. 最大となる $D(\mathbf{x}'_j)$ のデータ $\mathbf{x}'_j$ を次のクラスタ中心 $\mathbf{c}_l$ に選択する．

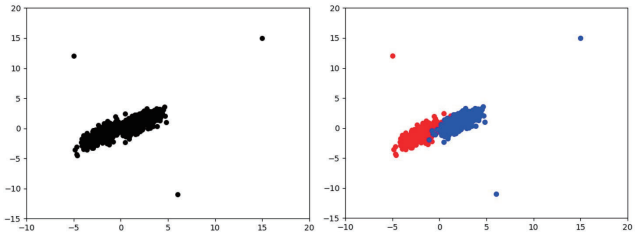

図 4: 左図：作成した人工データ，右図：人工データの正解クラスタ

f. クラスタ中心を k 個選ぶまで，ステップ 3,4,5 を繰り返す．k 個選択した後，k-means 法アルゴリズムのステップ 2, 4 と同様の処理を行う．

4. 実験条件

4.1 実験データ

前述した手法の有効性を検証するために，人工的に作成したデータとベンチマークデータを用いて実験を行った．

人工的に作成したデータは，平均 $(x_1, x_2) = (-2, -1)$ および，$(x_1, x_2) = (2, 1)$ の正規分布で，x_1 と x_2 の相関係数が 0.7 となるように 400 データずつ作成した．作成したデータ集合に，全てのデータの平均からのマハラノビス距離の二乗の値が 11.829 を超えたデータを 3 つ加えて作成した．図 4 の左は作成した人工データを，右は正解クラスタに色分けを行ったものを示す．

ベンチマークデータとして，UCI 機械学習レポジトリ [Newman 98] から abalone データを使用する．abalone データは，直径，直径から垂直の長さ，高さの属性数 3 のデータで，データ数 4177 があり，3 クラスに分類する．

4.2 評価方法

クラスタリング結果の評価は，式 (5) に示すクラスタ内の各データとクラスタ中心との距離の二乗和（クラスタ内残差二乗和）および，式 (6) に示す全てのクラスタ中心の平均と各クラスタ中心との距離の二乗和（クラスタ間残差二乗和）を用いて行う．式 (5) の $1[\mathbf{x}_j \in \mathcal{C}_i]$ はデータ $\mathbf{x}_j$ がクラスタ $\mathcal{C}_i$ に所属しているなら 1，所属していないなら 0 を返す．また Δ はすべてのクラスタ中心の平均を示している．

$$\text{クラスタ内残差二乗和} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\mathbf{c}_i - \mathbf{x}_j)^2 1[\mathbf{x}_j \in \mathcal{C}_i] \tag{5}$$

$$\text{クラスタ間残差二乗和} = \sum_{i=1}^{k}(\Delta - \mathbf{c}_i)^2 \tag{6}$$

クラスタ内残差二乗和の値が小さいということは，クラスタ内のデータはまとまっている．つまり似たデータが集まっていることを示し，クラスタ内残差二乗和の値が大きいということはクラスタ内のデータは似ていないデータの集まりということを示す．クラスタ間残差二乗和の値が小さいということは，クラスタ間は近い位置に生成されている．つまり異なるクラスタ同士が似ているということを示し，クラスタ間残差二乗和の値が大きいということは，異なるクラスタ同士は似ていないということを示す．k-means 法のアルゴリズムではクラスタ内残差二乗和が小さくなるようにクラスタが生成される．したがっ

表 1: 人工データの実験結果

初期値設定方法	一致率	内残差二乗和	間残差二乗和
k-means (max)	**98%**	**2373**	**5.58**
k-means (min)	50%	8203	14.3
k-means++ (max)	**98%**	**2373**	**5.58**
k-means++ (mix)	50%	8203	14.3
KKZ	50%	8203	14.3
KKZ-EX	**98%**	**2373**	**5.58**

表 2: abalone データの実験結果

初期値設定方法	内残差二乗和	間残差二乗和
k-means (Min)	**2981**	**4.28**
k-means (Max)	4326	31.98
k-means++ (Min)	**2981**	**4.28**
k-means++ (Max)	4326	31.98
KKZ	4326	31.98
KKZ-EX	**2981**	**4.28**

て，クラスタ内残差二乗和が小さく，クラスタ間残差二乗和の値は大きいほうが良いクラスタができていると判断する．

さらに，人工的に作成したデータには正解クラスタが存在するため，人工データに対しては生成されたクラスタがどの程度正解クラスタと一致しのかの一致率の比較も行う．

5. 実験結果

5.1 人工データの結果

表 1 に人工データの各初期値設定法の結果を示す．一致率の値が最も大きかった初期値設定方法の数値は太字にした．表 1 中の k-means は k-means 法のランダムで初期クラスタ中心を選択する方法のことである．また内残差二乗和と間残差二乗和はクラスタ内残差二乗和とクラスタ間残差二乗和を示す．k-means 法および k-means++法は初期値がランダムに選択されるので，10 回繰り返して生成されたクラスタの一致率が最大の時を (max) として，最小の時を (min) として示した．この表を見ると，KKZ は初期値外れ値を初期クラスタ中心に選択してしまったので，正解クラスタとの一致率が著しく低くなってしまったと考えられる．最も離れたデータ同士のデータを初期クラスタ中心に設定するためにクラスタ間残差二乗和は最も大きくなっている．また，クラスタ内残差二乗和が大きな理由は，生成された一つのクラスタが初期値外れ値となったデータともう一つのデータのみで一つのクラスタを生成し，残りすべてのデータで二つ目のクラスタを生成していた．その結果，クラスタ内残差二乗和が大きくなってしまっている．一方で，KKZ-EX は初期値外れ値となったデータを初期クラスタ中心に選択しないので，正解クラスタと殆ど同じクラスタを生成できた．また従来の k-means 法や k-means++法はランダムで初期クラスタ中心が選択されるので，一致率が最も高い場合のクラスタ結果を得ることができる．しかし一致率が最も低い場合のクラスタが生成されることもある．この実験から，KKZ-EX は初期値依存問題と初期値外れ値の問題に有効であることが確認できた．

5.2 ベンチマークデータの結果

abalone データの属性数は 3 であるため，しきい値は $\chi^2(3, 0.0455) = 8.025$ とした．abalone データのクラスタ内残差二乗和とクラスタ間残差二乗和を表 2 に示す．k-means 法および k-means++法は初期値がランダムに選択されるので，10 回繰り返して生成されたクラスタのクラスタ内残差二乗和が最小の時を (Min)，最大の時を (Max) として示した．また，クラスタ内残差二乗和が最小の時を太字としている．

abalone データも人工データと同様にクラスタ間残差二乗和は KKZ-EX より KKZ が大きくなり，クラスタ内残差二乗和の値は KKZ-EX より KKZ が大きくなった．KKZ-EX より KKZ のクラスタ間残差二乗和が大きくなった理由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するため，最終的なクラスタ中心も自然と離れたからである．

6. まとめと今後の課題

k-means 法の初期値依存の問題や初期値外れ値の問題を解決するために数多くの研究が行われている．数多く研究されているが，初期値依存の問題と初期値外れ値の問題の両方を同時に解決する方法は存在しなかった．本稿では，初期値依存の問題と初期値外れ値の問題の両方を同時に解決する方法を提案した．提案手法は MT 法を用いて初期値外れ値を除き，最も遠いデータを初期クラスタ中心に選択することで，二つの問題を解決した．人工データとベンチマークデータを用いて提案手法の有効性を検証した．今後の課題として，本稿の提案手法では初期値外れ値の判断方法に MT 法を用いて，そのしきい値として χ^2 分布を適用したが，しきい値設定には F 分布を用いたしきい値設定などもあるため，他のしきい値設定で比較検討を行う必要がある．

参考文献

[元田 06] 元田浩, 山口高平, 津本周作, 沼尾正行, “データマイニングの基礎”, オーム社, 2006.

[MacQueen 67] J.B. MacQueen, “Some methods for classification and analysis of multivariate observations”, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297, 1967.

[Katsavounidis 94] I. Katsavounidis, C.C.J. Kuo, Z. Zhang, “A new initialization technique for generalized Liyoyd iteration”, IEEE Signal Processing Letters, 1(10), 144-146, 1994.

[Arthur 07] David Arthur, “k-means++: The advantages of careful seeding”, Proc. of the eighteenth annual ACM-SIAM symposium on Discrete algorithm, 1027-1035, 2007.

[長谷川 06] 長谷川良子, “マハラノビス・タグチ（MT）システムのはなし”, 日科技連出版社, 2006.

[兼高 87] 兼高達貳, “マハラノビスの汎距離の応用例 特殊健康診断の事例”, 標準化と品質管理, 40(10), pp.57-64, 1987.

[Newman 98] C. B. D. J. Newman, S. Hettich, C. Merz, “UCI Repository of Machine Learning Databases”, http://www.ics.uci.edu/mlearnMLRepository.html , 1998.