1J1-OS9-1

k-means法の様々な初期値設定によるクラスタリング結果の実験的比較

Experimental Comparison of Clustering Results for k-means by using different seeding methods

小野田 崇[*1*2] 坂井 美帆[*2] 山田 誠二[*2*3]
Takashi Onoda Miho Sakai Seiji Yamada

[*1](財) 電力中央研究所
Central Research Institute of Electric Power Industry

[*2]東京工業大学大学院
Tokyo Institute of Technology

[*3]国立情報学研究所
National Institute of Informatics

The k-means method is a widely used clustering technique. The method is widely used for clustering information on theWeb 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. We propose a seeding method based on the independent component analysis for the 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

近年インターネットの普及に伴い，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，個人のハードディスクには旅行で撮った写真など，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せれば，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，一般的に「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!の映画，ニュースなど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探索は,「carrot2.org」や「Clusty.jp」などでも開発が行われており，近年注目されている方法である [1]．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 [2]．

クラスタリング手法とは，web ページなどのテキストや記念写真の画像のようなデータを，自動的にグループ化する一種の教師なし学習手法である．このクラスタリングには，階層型クラスタリングと非階層型クラスタリングがある．

階層型クラスタリングは，最も類似したデータから逐次的に一つのグループであるクラスタへまとめていくことで階層的なクラスタ構成を得る手法である．例えば図 1 のようにデータが与えられた場合，階層型クラスタリングはそれぞれのデータ同士の類似度に基づき，図 2 のようなデンドログラムを得る．このデンドログラムは，直感的にデータ構造を理解しやすくなっている．この方法では全体的な類似関係から，データ一つ一つの類似関係まで把握できることから，データ全体の構造を理解しやすい．しかしこの方法では，全てのデータ間の類似度の計算が必要になるなど，計算コストが膨大となる．このため，web ページのように大規模で日々膨大な量が増加するデータに対して，階層型クラスタリングの適用は現実的ではない．

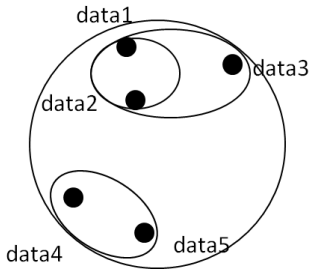

図 1: 元データ

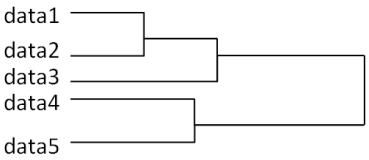

図 2: デンドログラム

非階層型クラスタリングは，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に分ける．この方法は，階層型クラスタリングと比べ計算コストが小さいため，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，非階層型クラスタリングとして，k-means 法が多く用いられている．この k-means 法は，クラスタの中心と各データとの距離が最小になるようなクラスタ中心を逐次的に求めることで，クラスタリングを行う手法である．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）してしまうという重大な問題がある．本稿では，この k-means 法の初期値設定の違いによるクラスタリング結果の実験的な比較について報告する．

連絡先: 小野田崇，(財) 電力中央研究所システム技術研究所，〒 201-8511 東京都狛江市岩戸北 2-11-1，03-3480-2111（内線 1633），onoda@criepi.denken.or.jp

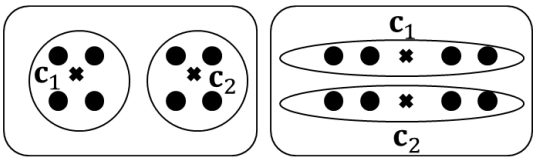

図 3: k-means 法の初期値依存

2. 関連研究

k-means 法の初期値設定手法の研究は古くから行われているが [3]，本節では，主要な研究として Katsavounidis らによって提案された KKZ 法 [4]，David Arthur によって提案された k-means++法 [5]，について説明する．最初にオリジナルの k-means 法について述べ，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 法は非階層型クラスタリングの一種であり，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.

1. 任意の k 個のクラスタ中心 $\mathbf{c}_i$ を一様ランダムに選択する.
2. 全てのデータを，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.

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

   C_i は各クラスタ i に含まれるデータの集合であり，$|C_i|$ はクラスタ 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，計算コストが小さいというメリットから，よく用いられる．しかし，kmeans 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結果が依存してしまう問題がある．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 3 に示す．左右共に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 法では初期の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なことではない.

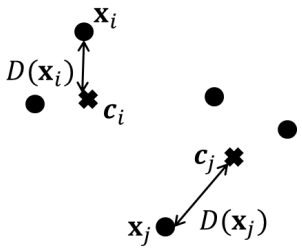

図 4: $D(\mathbf{x})$: 各データ点から最も近い既存のクラスタ中心との距離

2.2 KKZ(Katsavounidis 1994)

KKZ 法は，Katsavounidis ら (1994) 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 $\mathbf{x}_j$ と既に決定されたクラスタ中心との最短距離を表す (図 4).

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，b.-c. を繰り返す．k 個選択した後，k-means 法アルゴリズム 2-4 と同様の処理を行う.

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，外れ値に敏感であるという問題がある.

2.3 k-means++法 (David Arthur 2007)

k-means++ は，David Arthur によって提案された，近年最も注目されている k-means 法の初期値設定手法である．この方法は，KKZ 法の外れ値に弱いといった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したが，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. 1 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.

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

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.

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

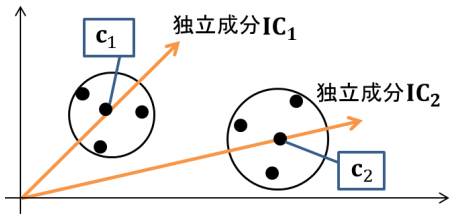

図 5: 提案方法のイメージ

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

式（4）から分かるように，$D(\mathbf{x}_j)$ 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．つまりこの手法では，既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，クラスタ同士の距離を離すことができる．また，式（4）中の L がランダムに選ばれることから，必ず一番遠いデータ点を選択する訳ではなく，外れ値に依存するといった傾向が弱くなる．このため kmeans++ 法は，ランダムに初期値を選択する k-means 法よりもより良いクラスタリングが実現できることが報告されている．しかし上述したアルゴリズムから分かるように，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．

3. 提案する初期値設定法

本節では，提案する独立成分分析 (ICA) に基づく k-means 法の初期値設定手法について述べる．この手法では，データの集合を観測信号と考え，ICA によって求められる復元信号 Y が，クラスタ同士を独立させるようなクラスタの特徴ベクトルであると考える．つまり，この手法では，与えられたデータ $\mathbf{X}$ から ICA によって，各クラスタを特徴づける独立成分ベクトル (IC) を求め，IC からコサイン距離が最も近いデータ点を初期値として設定する（図 5）．以下にアルゴリズムを示す．

a. データ $\mathbf{X}$ から，ICA を用いて k 個の独立成分 $\mathbf{IC}_t, t \in \{1, \cdots, k\}$ を得る．

b. 全ての独立成分 $\mathbf{IC}_t$ に対して，次式を最小にするデータ点 $\mathbf{x}_j$ をクラスタ中心 $\mathbf{c}_i$ に選択する．

$$F_{ICA} = \frac{\mathbf{IC}_t \cdot \mathbf{x}_j}{\|\mathbf{IC}_t\| \|\mathbf{x}_j\|} \tag{5}$$

クラスタ中心を k 個選択した後，k-means 法アルゴリズム 2-4 と同様の処理を行う．

この方法を用いることで，クラスタ同士の独立性が高い k-means 法のクラスタ中心の初期値を設定できる．

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，小規模なデータセットと，大規模なデータセットを用いて実験を行った．実験では小規模データとして，UCI ベンチマークデータ [6] と Open Directory Project(ODP)[7] コーパスを用いた．

4.1.1 UCI ベンチマークデータ

実験では UCI ベンチマークデータから，Iris，Wine，Soybean-Small，Breast-Cancer の 4 つのベンチマークデータを採用する．

4.1.2 Open Directory Project(ODP) コーパス

ODP コーパスは，ボランティア方式で運営される世界最大の Web ディレクトリである．実験では，ODP コーパスのうち「Agriculture」「Physics」「Social Sciences Linguistics」「Technology Structural Engineering」の 4 つのクラスについて考える．

4.2 評価方法

クラスタリング結果の評価は，次式に示す正規化相互情報量 (NMI；normalized mutual information)[8] を用いて行う．

$$NMI(\mathcal{C}, \mathcal{T}) = \frac{MI(\mathcal{C}, \mathcal{T})}{\max(H(\mathcal{C}), H(\mathcal{T}))} \tag{6}$$

$\mathcal{C}$ は生成されたクラスタ集合，$\mathcal{T}$ は正解クラスタ集合であり，MI は相互情報量，H はエントロピーを表す．このとき，

$$H(\mathcal{C}) = \sum_{i}^{k} -P(C_i) \log P(C_i), i \in \{1, \cdots, k\} \tag{7}$$

で表される．また，

$$P(C_i) = \frac{|C_i|}{N} \tag{8}$$

であり，N は全データ数，$|C_i|$ は生成されたクラスタ i に含まれるデータの数を示す．$H(\mathcal{T})$ も同様に求める．さらに相互情報量は，

$$MI(\mathcal{C}, \mathcal{T}) = H(\mathcal{C}) + H(\mathcal{T}) - H(\mathcal{C}, \mathcal{T}) \tag{9}$$

となる．このとき，

$$H(\mathcal{C}, \mathcal{T}) = \sum_{i}^{k} \sum_{j}^{k} -P(\mathcal{C}_i, \mathcal{T}_j) \log P(\mathcal{C}_i, \mathcal{T}_j), \tag{10}$$

$$j \in \{1, \cdots, k\}. \tag{11}$$

である．NMI は，00 から 1 の間の値をとり，値が大きい程生成されたクラスタが正解クラスタ集合に類似していることを示す．

4.3 比較手法

本研究の実験では，オリジナルの k-means 法，KKZ 法，k-means++法，提案方法および提案方法の独立成分分析を主成分分析に換えた方法のクラスタリング結果を比較する。独立成分分析の部分を主成分分析に換えたアルゴリズムを以下に示す。

a. k 個の主成分ベクトル $\mathbf{PC}_t, t \in \{1, \cdots, k\}$ を得る．

b. 全ての主成分に対して以下を最小にするるデータ点 $\mathbf{x}_j$ をクラスタ中心 $\mathbf{c}_i$ に選択する．

$$F_{PCA} = \frac{\mathbf{PC}_t \cdot \mathbf{x}_j}{\|\mathbf{PC}_t\| \|\mathbf{x}_j\|} \tag{12}$$

クラスタ中心を k 個選択した後，k-means 法アルゴリズム 2-4 と同様の処理を行う．

表 1: UCI ベンチマークデータ実験結果

Iris

	最小分散時 NMI	最大 NMI
k-means 法	0.751	0.751
k-means++法	0.751	0.751
KKZ 法	0.751	-
PCA	0.751	-
提案方法	0.751	-

Wine

	最小分散時 NMI	最大 NMI
k-means 法	0.428	0.428
k-means++法	0.428	0.428
KKZ 法	0.387	-
PCA	0.428	-
提案方法	0.428	-

Soybean-Small

	最小分散時 NMI	最大 NMI
k-means 法	0.710	1.000
k-means++法	0.751	1.000
KKZ 法	0.751	-
PCA	0.751	-
提案方法	0.751	-

Breast-Cancer

	最小分散時 NMI	最大 NMI
k-means 法	0.743	0.743
k-means++法	0.743	0.743
KKZ 法	0.743	-
PCA	0.743	-
提案方法	0.743	-

4.4 類似度

本実験では，k-means 法に用いる類似度として，UCI ベンチマークデータに対してはユークリッド距離，文書データ（ODP コーパス）に対しては，コサイン距離を用いる．

5. 実験結果

5.1 UCI ベンチマークデータ

UCI ベンチマークデータに k-means 法，kmeans++法，KKZ 法，PCA に基づく手法 (以下，PCA)，提案手法による初期値選択手法を実行したときの結果を表 1 に示す．表では，k-means 法と k-means++法での比較する NMI として，最小分散時 NMI，最大 NMI の値を示した．KKZ 法，PCA，提案手法では，一度に得たクラスタリング結果の NMI の値を示している．実験結果から，提案手法と，k-means 法，k-means++法とを比較するると，全ての UCI ベンチマークデータで，提案手法の NMI は，一般的な k-means 法と k-means++法の最小分散時の NMI と同じ値であった．Iris，Wine，Breast-Cancer データでは，提案手法の NMI は，一般的な k-means 法と k-means++法の最大 NMI とも同じ結果を示した．また，提案手法と PCA を比較した場合，全ての UCI ベンチマークデータで同じ NMI を示した．提案手法と KKZ 法とを比較すると，Iris，Soybean-Small，Breast-Cancer データでは，KKZ 法は提案手法と同じ NMI を示したが，Wine データでは，提案手法の方が高い NMI を示した．このときの KKZ 法の NMI の値は，k-means 法が示す最小の NMI の値であった．

表 2: ODP コーパス実験結果

	最小分散時 NMI	最大 NMI
k-means 法	0.555	0.589
k-means++法	0.555	0.589
KKZ 法	0.531	-
PCA	0.500	-
提案方法	0.638	-

5.2 ODP コーパス

ODP コーパスに各手法を適用した結果を表 2 に示す．実験結果から，提案手法と k-means 法，k-means++法と比較した場合，提案手法の NMI の値は，両手法の最小分散時の NMI の値だけでなく，最大 NMI の値より高い NMI の値を示した．また，KKZ 法と PCA と比較した場合でも，提案手法が最も高い NMI の値を示した．

6. まとめと今後の展開

本論文では，k-means 法における様々な初期値設定法について，UCI ベンチマークデータならびに ODP コーパスを用いて，実験的にその性能を比較した。小規模なデータに対してではあるが，提案方法が他よりも良い性能を示した。また，提案方法はクラスタ内には類似データが集められ，クラスタ間は独立のようなクラスタが生成されやすくなっており，生成されたクラスタにユーザが制約を加える際の事前知識として利用できる。

今後は，より大きいデータに提案方法を適用し，その特性を分析する。

参考文献

[1] Mari A. Hearst, "Clustering versus facted categories for information exploration", Commum. ACM, 49(4):59-61,2006.

[2] 元田浩ほか, データマイニングの基礎, オーム社, 2006.

[3] Douglas Steinley, Michael J. Brusco, "Initializing k-means Batch Clustering: A Critical Evaluation of Several Techniques", Journal of Classification, Vol.24, No.1, 99-121, 2007.

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

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

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

[7] Open Directory Project, http://dmoz.org/, 2002.

[8] Hao Cheng, Kien A. Hua, Khanh Vu, "Constrained locally weighted clustering", Proc. of the VLDB Endowment, Vol.1, No.1, 2008.