2G2-OS9-2

独立成分分析によるk-means法の初期値設定手法の提案

坂井美帆[*1]，山田誠二[*2]，小野田崇[*3]
Miho Sakai，Seiji Yamada，Takashi Onoda

[*1*3]東京工業大学 Tokyo Institute of Technology
[*2]国立情報学研究所 National Institute of Informatic

The k-means method is a widely used clustering technique because of its simplicity and speed. However, the clustering result depends heavily on the chosen initial value. In this report, we propose a seeding method with independent component analysis for the k-means method. Using a benchmark dataset, we evaluate the performance of our proposed method and compare it with other seeding methods.

1. はじめに

Web 上での情報が一般的になり，Web 上に情報が氾濫すると，その膨大な情報を整理しなければならないという問題が起こる．そのような問題に対し，教師あり学習や教師無し学習が用いられる．特に，教師無し学習ではクラスタリング手法の一つである k-means 法が非常によく用いられる．この k-means 法は簡単なアルゴリズムでかつ高速なため，Web 上の情報整理以外の場面でも頻繁に用いられている．しかし，k-means 法はそのクラスタリング結果が初期値に依存してしまうという問題点がある．本稿では k-means 法の初期値依存性を解決する一つの方法について議論する．以下，2 章で関連研究として，本稿で比較する既提案方法について紹介する．3 章で提案方法について述べ，4 章で実験結果について報告する．5 章でまとめと今後の課題について述べる．

2. 関連研究

本章では，本稿で提案する方法と比較する k-means 法と k-means++法について述べる．

2.1 k-means 法

非階層型クラスタリングの一種である 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 1 つである．(1) の評価関数を最小化することによって，任意の k 個のクラスタに分割する．

$$\phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} \| x - c \|^2 \quad (1)$$

アルゴリズムを以下に示す．

1. 任意に k 個のクラスタ中心を選ぶ.$\mathcal{C} = \{C_1, \cdots, C_k\}$.
2. 全てのデータ $\mathcal{X}$ を，最も近いクラスタ $C_i, i \in \{1, \cdots, k\}$ に割り当てる．
3. 各クラスタ C_i ごとに，含まれるデータの中心を求める：$C_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$.
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．

連絡先: 坂井美帆，東京工業大学総合理工学研究科知能システム科学専攻 sakai@ntt.dis.titech.ac.jp

2.2 k-means++法

k-means++法 [1] は，k-means 法の初期値の選択方法について改良を加えた手法である．アルゴリズムを以下に示す．

1a. 1 つ目のクラスタ中心 C_1 をデータ $\mathcal{X}$ からランダムに選ぶ.

1b. 確率 $\frac{D(x')^2}{\sum_{x \in \mathcal{X}} D(x)^2}$ を最大にするデータ点 x' を，次のクラスタ中心 C_i に選択する：$C_i = x' \in \mathcal{X}$.

1c. クラスタ中心を k 個選ぶまで 1b を繰り返す.

2-4. k-means アルゴリズムと同様の処理を行う.

このとき，$D(x)$ はデータ点 x と既に決定されたクラスタ中心との最短距離を表す．k-means++法では k-means 法に比べ良好なクラスタリングが実現しやすいことが報告されている [1].

3. 提案手法

本章では独立成分分析 (ICA；Independent Component Analysis)[2] による初期値選択を行う k-means 法を提案する．提案手法のイメージを図 1 に示す．提案手法では ICA によって与えられたデータ $\mathcal{X}$ から独立成分 (IC) を求め，k-means 法の初期値にする．実際には求められた IC にコサイン距離が最も近いデータ点を初期値として選択する．以下にアルゴリズムを示す.

1a. k 個の独立成分 $\mathrm{IC}_m, m \in \{1, \cdots, k\}$ を得る.

1b. $F = \frac{\mathbf{IC}_m \cdot \mathbf{x}}{|\mathbf{IC}_m||\mathbf{x}|}$ を最小にするデータ点 x を次のクラスタ中心 C_i に選択する：$C_i = x \in \mathcal{X}$.

1c. クラスタ中心を k 個選ぶまで 2 を繰り返す.

2-4. k-means アルゴリズムと同様の処理を行う.

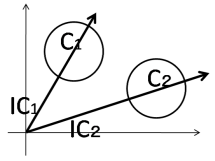

図 1: 提案手法：イメージ

4. 実験

4.1 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)[3] を用いて行う．

$$\mathrm{NMI}(\mathcal{C},\mathcal{T}) = \frac{\mathrm{MI}(\mathcal{C},\mathcal{T})}{\max(\mathrm{H}(\mathcal{C}),\mathrm{H}(\mathcal{T}))} \quad (2)$$

$\mathcal{C}$ 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，$\mathrm{H}(\mathcal{C}) = \sum_i^k -P(C_i)\log P(C_i), i \in \{1,\cdots,k\}$ で表す．また，$P(C_i) = \frac{num(C_i)}{N}$ であり，N は全データ数，$num(C_i)$ は生成されたクラスタ C_i に含まれるデータの数を示す．$\mathrm{H}(\mathcal{T})$ も同様に求める．さらに相互情報量は，$\mathrm{MI}(\mathcal{C},\mathcal{T}) = \mathrm{H}(\mathcal{C}) + \mathrm{H}(\mathcal{T}) - \mathrm{H}(\mathcal{C},\mathcal{T})$ と表す．NMI は，0 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確であることを示す．

4.2 ベンチマーク

UCI ベンチマークから，**Iris** データ（データ数：150，属性数：4，クラス数：3），**Wine** データ（データ数：178，属性数：13，クラス数：3），**Soybean-Small** データ（データ数：47，属性数：35，クラス数：4）の 3 データセットを用いた．

4.3 実験結果

各ベンチマークに k-means 法，k-means++法，ICA による初期値選択手法を実行したときの結果を表 1〜3 に示す．k-means 法，k-means++法は 100 回試行を繰り返し，平均 NMI（**avg**），最大 NMI（**max**），最小 NMI（**min**），最小分散クラスタのときの NMI（**min Sumd**）を示した．ただし，ICA は 1 回の試行のみの実行結果である．

表 1: **Iris**

	avg	max	min	min Sumd
k-means	0.70325	0.751485	0.532224	0.751485
k-means++	0.749295	0.751485	0.532471	0.751485
ICA	0.751485			

表 2: **Wine**

	avg	max	min	min Sumd
k-means	0.417794	0.428701	0.3873	0.428701
k-means++	0.418351	0.428701	0.3873	0.428701
ICA	0.428701			

表 3: **Soybean-Small**

	avg	max	min	min Sumd
k-means	0.714445	1	0.518038	0.710813
k-means++	0.806213	1	0.710813	0.710813
ICA	0.710813			

実験結果から，全データセットにおいて k-means++の方が k-means 法より平均 NMI が高くなった．**Iris** データ，**Soybean-Small** データの場合には，k-means 法に比べて解の劣化を防いでいる．このことから，k-means++法は k-means 法よりも適切な初期値を選ぶ手法であることが分かる．

Iris データ，**Wine** データにおける 3 手法の結果を比較する．提案手法は他の 2 手法の最高 NMI，最小分散時の NMI と同じ結果を示した．このことから，適切な初期値を選択できていることが分かる．加えて，平均 NMI と比較した場合，提案手法の方が優れた結果になることが分かる．

Soybeans-Small データのとき，提案手法は悪い結果を示した．これは提案手法では，図 2 のように，一つの独立成分 IC に，複数のクラスタ C が対応する場合について考慮されていないからだと考えられる．しかしこの場合でも k-means 法，k-means++法の最小分散時の NMI と同等の結果を得，著しい解の劣化を防いでいる．

以上から，提案手法の有用性を示した．加えて，k-means 法，k-means++法は，複数回実行する必要性があり，一意に適切な初期値を求められる提案手法は計算コストの面からも有用であると考えられる．

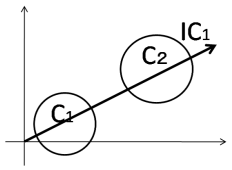

図 2: 実験結果：問題点

5. まとめと今後の方針

一般的なクラスタリング手法として挙げられる k-means 法は，簡単なアルゴリズムでかつ高速なため，多くの研究で用いられている．しかし，クラスタリング結果が初期値に依存してしまうという問題点がある．そこで一意に k-means 法の初期値を決定する手法として，独立成分分析を用いた k-means 法の初期値設定方法を提案した．実際にベンチマークデータを用いて，提案手法と，k-means 法と，k-means 法の初期値選択方法に改良を加えた k-means++法との比較実験を行い，ICA の優位性を示した．今後は，図 2 の問題に対処するため，独立成分 IC からクラスタ中心を選択する手法に改良を加えたい．

参考文献

[1] David Arthur et al., ”k-means++: The advantages of care- ful seeding,” Proc. of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp.1027-1035, 2007.

[2] A. Hyvarinen and E. Oja: “ A Fast Fixed-Point Algorithm for Independent Component Analysis ”, Neural Computation, vol9, pp. 1483-1492, 1997

[3] Hao Cheng , Kien A. Hua , Khanh Vu, Constrained locally weighted clustering, Proceedings of the VLDB Endowment, v.1 n.1, August 2008