4F1-OS-5-3

高次独立性に基づくクラスタリング

Clustering based on high independence

西垣 貴央[*1] 小野田 崇[*1*2]
Takahiro Nishigaki Takashi Onoda

[*1]東京工業大学 大学院 Tokyo Institute of Technology
[*2]電力中央研究所 Central Research Institute of Electric Power Industry

In general, existing clustering methods focused on the similarity of the data within the cluster. Therefore, distance and independence between clusters were not taken into account. However, users expect that the data within a cluster are similar, and data between clusters are well separated or independent from each other. In this paper, we propose a clustering method where data within a cluster are similar, and data between clusters are highly independent. We show the results of experiments using benchmark data.

1. はじめに

近年，デジタルカメラやノート PC などの安価で高性能なデバイスが簡単に手に入ることや，インターネットの普及に伴い，Web ページや電子ニュースなどの様々な電子文書データ情報の配信，交換が盛んに行われるようになってきた．そのため，Web 上または，個人のハードディスクなどには文書や画像，音楽ファイルなどが多量に蓄積されている．蓄積された文書や画像などのデータの数はあまりにも膨大であるため，その中から必要な情報を発見，検索をする事が困難になっている．必要な情報の発見や検索を容易にするために，多量のデータは何らかの方法でグループ化されていることが必須である．こういったデータをグループ化する方法として，一般にクラスタリングが適用されている．このクラスタリングには，クラスタリング結果がユーザに分かりやすいことと，ユーザの欲しているグループに一致していることが求められる．

Web ページなどデータの更新頻度が高いデータをグループ化する際に，用いられるクラスタリング手法として，簡単なアルゴリズムである k-means 法がある．この k-means 法によって，生成されるクラスタ内は類似したデータで構成されている．しかし，生成されたクラスタ間には何の制約もないため，クラスタ間がどのような関係にあるのかが分からない．

そこで，本稿ではクラスタ間の独立性が高く，クラスタ内の類似性が保証されたクラスタリング結果を得るクラスタリング手法について述べる．以下，2 章で関連研究とその課題について紹介し，3 章では提案手法について述べる．4 章では実験についてとその結果について報告し，5 章で今後の課題について述べる．

2. 関連研究

本章では，関連研究として k-means 法の初期値を独立成分分析 (Independent Component Analysis: ICA)[A.Hyvarinen 97][村田 05] によって選択する手法について述べる [坂井 10].

k-means 法は，式（1）の評価関数を最小化することによって，観測されたデータを，任意の k 個のクラスタに分割する．

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

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 $\mathbf{X}$ 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．

このように，非常に簡単なアルゴリズムなため，クラスタリング手法として多く利用されている．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った．

そこで坂井ら [坂井 10] は，k-means 法の初期値を選ぶ際に，独立成分分析で推定した独立成分に最も近いデータを初期値に設定することにより，クラスタリング結果が初期値に依存するという問題を解決した．

坂井らが提案した初期値設定法について簡単に図 1 を用いて説明する．図 1 左側では，初期値に依存するという問題点を解決するために，独立成分分析によって推定した独立成分（図

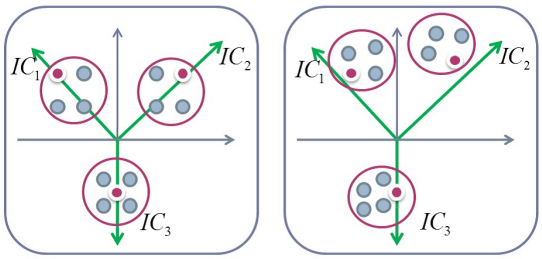

図 1: ICA による k-means 法の初期値設定法（左）とその課題（右）

連絡先: 西垣貴央，東京工業大学大学院 総合理工学研究科 知能システム科学専攻，〒 226-8502 神奈川県横浜市緑区長津田町 4259 J2-53，045-924-5205，nishigaki@ntt.dis.titech.ac.jp

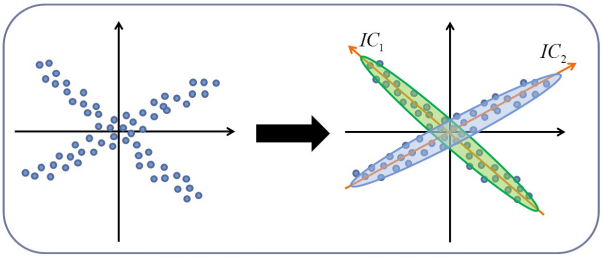

図 2: 提案手法のイメージ

では軸）に最も近いデータを初期値に選択することにより，クラスタリング結果が初期値に依存しないようにした．

しかしこの手法は，独立成分分析によって独立成分を抽出したにも関わらず，その後に k-means 法を行うことで，生成されたクラスタ間での独立性は必ずしも保障されているわけではないという問題点（図 1 右側) があった．

3. 提案手法

本章では，独立成分分析の概念に基づきクラスタ同士の独立性を保つとともに，クラスタ内の類似性を維持するクラスタリング手法を提案する．それにより生成されるクラスタは，クラスタ間の独立性および，クラスタ内の類似性が保証されるため，ユーザにとって直感的に理解しやすいものになるはずである．

提案手法では，まず与えられたデータから独立成分分析によってクラスタの数だけ独立成分を推定する．推定した独立成分がクラスタの中心となるので，各データはその独立成分との関係によってクラスタリングされる．関係というのは，求めた独立成分の属性と各データの属性の相関係数の値が，最も高い独立成分のクラスタにクラスタリングされる．

以下にアルゴリズムを示す．

1. 独立成分分析に基づいて，全データ $\mathbf{X}$ から k 個の独立成分 $IC_m, m \in \{1, \cdots, k\}$ を得る．
2. $r = \dfrac{cov(x_i, IC_m)}{\sqrt{cov(x_i, x_i)cov(IC_m, IC_m)}}, i \in \{1, \cdots, N\}, m \in \{1, \cdots, k\}$ を最大とする IC_m のクラスタに x_i がクラスタリングされる．$cov(x, y)$ は x と y の共分散行列を出力する．
3. 全てのデータ x_i がいずれかのクラスにクラスタリングされるまで，ステップ 2. を繰り返す．

提案手法のイメージを図 2 に示す．図 2 の左側のデータが与えられた時，独立成分分析を行うと，図の右側に示すような独立成分が得られる．この独立成分がクラスタ中心となり，これと各データとの相関係数を計算してクラスタリングを行なった結果が図 2 である．

また本提案手法の特徴として，データが複数のクラスタに分類されることがあるという点である．図 2 を見ると，座標軸の中心付近のデータが IC_1 と IC_2 のどちらにも分類されていることがわかる．

4. 実験

本章では，4 つの人工データと CLUTO のコーパスデータを用いて，既存手法と提案手法のそれぞれを適用した結果の比較を行う．

4.1 評価手法

クラスタリング結果の評価は，式 (2) に示す正規化相互情報量（NMI；normalized mutual information）[Hao CHeng 08] を用いて行う．

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

$\mathcal{C}$ 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，$H(\mathcal{C}) = \sum_i^k -P(C_i)\log P(C_i), i \in \{1, \cdots, k\}$ 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．$H(\mathcal{T})$ も同様に求める．さらに相互情報量は，$MI(\mathcal{C}, \mathcal{T}) = H(\mathcal{C}) + H(\mathcal{T}) - H(\mathcal{C}, \mathcal{T})$ となる．このとき，$H(\mathcal{C}, \mathcal{T}) = \sum_i^k \sum_j^k -P(C_i, T_j)\log P(C_i.T_j), j \in \{1, \cdots, k\}$ である．NMI は，0 から 1 の間の値を取り，値が大きいほど生成されたクラスタが正解であることを示す．

4.2 人工データ

図 3 から図 6 のような，基本的に 2 つのクラス（塗りつぶした丸と，塗りつぶしていない丸）に分けることができる 4 つの人工データを作成した．以下に，各人工データの詳細について述べる．

図 3 提案手法と既存手法のそれぞれが上手くいくもので，$y = x, (50 \leq x \leq 100)$ の点 51 個と，$y = -x, (50 \leq x \leq 100)$ の点 51 個にそれぞれ，± 20 の一様乱数を足したものである．

図 4 提案手法では上手くいくが既存手法では上手くいかないもので，$y = x, (-100 \leq x \leq 100)$ の点 201 個と，$y = -x, (-100 \leq x \leq 100)$ の点 201 個にそれぞれ，± 20 の一様乱数を足したものである．このデータだけ，どちらのクラスタに分類されても良い部分があり，図では三角で書かれているちょうどデータがクロスしている部分，図では三角で表している部分である．

図 5 提案手法では上手くいかないが既存手法では上手くいくもので，$y = -x, (-100 \leq x \leq -50)$ の点 51 個と，$y = -x, (50 \leq x \leq 100)$ の点 51 個にそれぞれ，± 20 の一様乱数を足したものである．

図 6 両手法とも上手くいかないもので，$x^2 + y^2 = 25^2$ 上の点 200 個と，$x^2 + y^2 = 10^2$ の点 200 個にそれぞれ，± 7 の一様乱数を足したものである．

4.2.1 実験結果

前節で説明した 4 つの人工データに対して，既存手法を適用した場合の結果を図 7 に，提案手法を適用した場合の結果を図 8 にそれぞれ示す．提案手法の図にある 2 本の直線は，それぞれの独立成分でクラスタの中心を表している．また，既存手法および提案手法を適用した時の NMI の値を表 1 に示す．

4.2.2 考察

実験結果の図 7 より既存手法の，図 8 より提案手法の得手不得手がわかる．既存手法は，単純にデータ間の距離によってクラスタを決定する．そのため，図 7 のような結果になる．

しかし提案手法は，独立成分分析によってクラスタの中心を求め，そのクラスタの中心を元にしてクラスタリングを行う

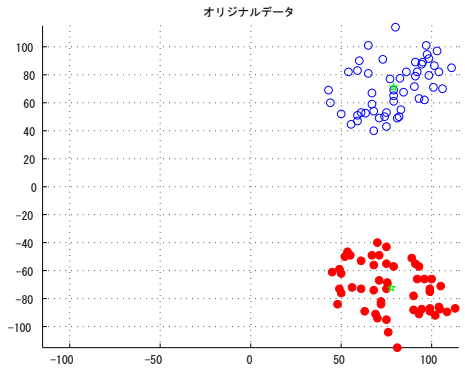

図 3: 人工データ（どちらも可）

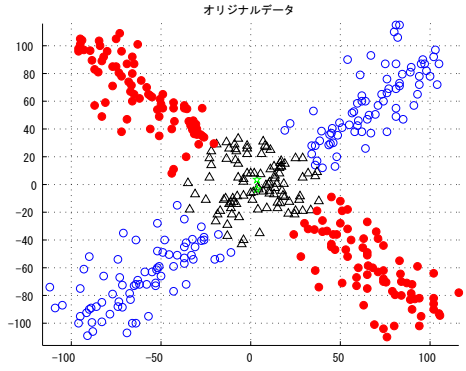

図 4: 人工データ（提案可、既存不可）

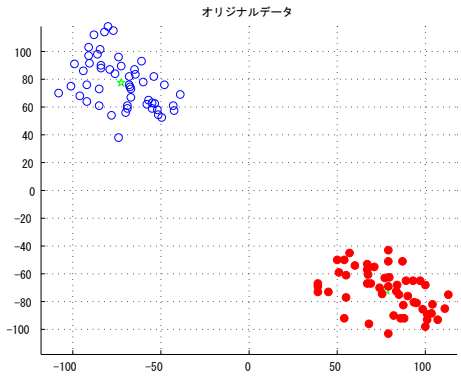

図 5: 人工データ（提案不可、既存可）

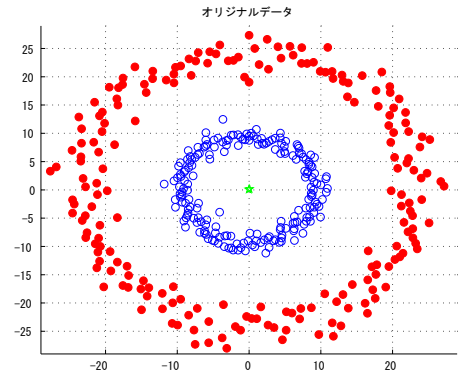

図 6: 人工データ（どちらも不可）

ものである（図 8）．そのため，求めたクラスタの中心となる軸によって与えられたデータを上手く分けることができる上段左の図や，上段右の図のようなものは非常に得意である．しかし，求めたクラスタの中心となる 1 つの軸上に 2 つ以上クラスタがある下段左の図や，下段右の図ように求めたクラスタの中心となる軸全てを横断するようなデータは，上手くクラスタリング出来ないことがわかる．また，上記の図 7 と図 8 からわかる結果は，表 1 に示した NMI の結果とも一致する．

4.3 CLUTO のコーパスデータ

大規模コーパスデータとして CLUTO のコーパスデータ [Gerge 02] を扱う．CLUTO は大規模データ用クラスタリングツールであり，CLUTO のデータはそれを評価するためのデータセットである．

CLUTO コーパスは，Web 文書データである．文書の表現方法として，ベクトル空間モデルを用いる．このとき文書単語行列 $\mathbf{X}$ は，全文書データ数 n，前単語数 $n \times m$ 行列で表現され，各単語に重み付けされる．重み付けには一般的に tfidf 法 [岸田 03] が用いられる．

CLUTO のコーパスデータのうち,「cacmcisi」「tr45」「wap」のデータを用いた．それぞれのデータセットのデータ数，属性数，クラス数を表 2 に示す．

表 1: 人工データ（図 3 から図 6）の NMI の結果

	上段左	上段右	下段左	下段右
提案手法	1	0.9229	0	0.0226
既存手法	1	0.2969	1	0.0002

4.3.1 実験結果

前節で説明した CLUTO のコーパスデータに対して，既存手法および提案手法のそれぞれを適用した NMI の結果を表 3 に示す．

4.3.2 考察

表 3 に示した NMI の値から，CLUTO のデータの一部において，提案手法が既存手法より有効な場合があることがわかった．CLUTO のデータはデータ数が多く，また属性の数も非常に多い．そのようなデータは現実世界にも多く，新聞データや Web ニュースなどが該当する．そのようなデータは，はっきりとクラスタリング出来ない場合が多く，どのクラスタに入るのか分からないものや，どのクラスタにでも入る可能性があるものも少なくない．提案手法は，そういったどのクラスタに入るのかわからないものや，どのクラスタにでも入る可能性があるものが含まれるデータをクラスタリングすることが，既存手

表 2: CLUTO のコーパスデータの詳細

	データ数	属性数	クラス数
cacmcisi	4663	14409	2
tr45	690	8261	10
wap	1560	8460	20

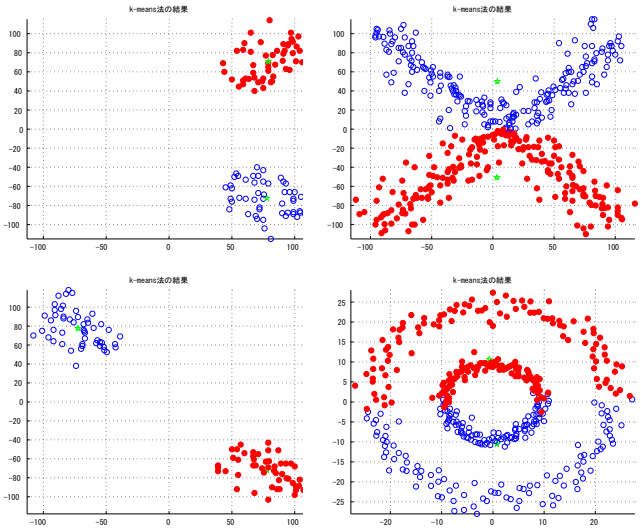

図 7: 既存手法の結果

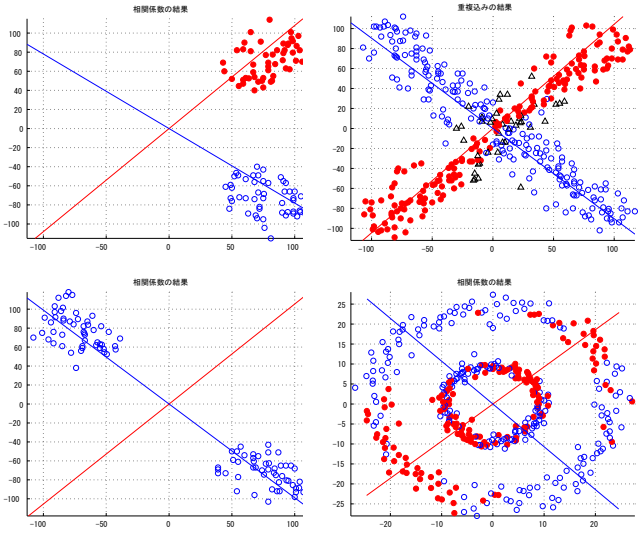

図 8: 提案手法の結果

法よりも向いている可能性が高い.

5. 今後の課題

今後の課題として，3 点ある．まず 1 点目，評価手法の検討である．提案手法はクラスタの重複を許す手法であるが，評価手法である NMI は 1 つのデータが 1 つのクラスタにクラスタリングされた場合の評価手法である．そのため，クラスタの重複を考慮した評価手法を検討しなければならない．

次に 2 点目，重複を許したデータの作成の検討である．今回実験に用いた CLUTO データは重複を許したものではないので，重複を許したデータの作成を行い，それに対して実験考察を行うつもりである．

最後の 3 点目は，本提案手法にユーザ制約（must-link など）を導入することである．既存のクラスタリング手法を用いてデータをクラスタリングする際，ユーザの視点が導入されているケースはなく，採用したクラスタリング手法の基準に従って，自動的にクラスタリングされてしまうため，生成された結果がユーザの欲しているグループになっていない場合も多い．そのため，本提案手法にユーザ制約を導入することで，ユーザが欲している結果に近づけることができる．

表 3: CLUTO のコーパスデータの NMI の結果

	cacmcisi	tr45	wap
提案手法	0.2664	0.6699	0.5751
既存手法	0.1239	0.6219	0.5250

6. おわりに

既存の k-means クラスタリングの初期値を ICA によって選択する手法 [坂井 10] では，クラスタ内のデータは類似しているものの，クラスタ間の独立性については，保証の限りではなかった．提案手法である独立成分分析の概念を用いたクラスタリングでは，クラスタ間の高い独立性を維持しつつ，クラスタ内の類似性も保証するものである．提案手法によるクラスタリング結果は，上記の関係を満たしている結果として，ユーザに安心感を与えると考えられる．

本稿では，作成した人工データに対して提案手法を適用し，提案手法の得意なデータ不得意なデータについて明確にした．また，提案手法を CLUTO のコーパスデータに適用し，データ量の多い，クラスタの重複がありそうなデータでの本提案手法の有効性を示した．今後は上述した課題の解決に取り組んでいく．

参考文献

[A.Hyvarinen 97] A.Hyvarinen, E.Oja, ”A Fast Fixed-Point Algorithm for Independent Component Analysis”, Neural Computation, Vol.9, pp. 1483-1492, 1997.

[村田 05] 村田昇, ”入門 独立成分分析”, 東京電機大学出版,2005.

[坂井 10] 坂井美帆, ”独立成分分析による安定な k-means 法の初期値設定手法の提案”, 2010.

[A.Hyvarinen 05] A.Hyvarinen, E. Oja:著, 根本幾・川勝真喜:訳, ”詳解 独立成分分析 信号解析の新しい世界”, 東京電機大学出版, 2005.

[A.Hyvarinen 99] A.Hyvarinen, ”Fast and Robust Fixed-Point Algorithms for Independent Component Analysis”,IEEE Trans. on Neural Networks, 1999.

[Hao CHeng 08] Hao Cheng, Kien A. Hua, Khanh Vu, ”Constrained locally weighted clustering”, Proceedings of the VLDB Endowment, Vol.1 No.1, August 2008.

[Gerge 02] George Karypis, CLUTO -A Clustering Toolkit, Dept. of Computer Science, University of Minnesota, May 2002, http://glaros.dtc.umn.edu/gkhome/views/cluto/

[岸田 03] 岸田和明, ”文書クラスタリングの技法：文献レビュー”, Proc. Library and Information Science No.49, pp. 33-75, 2003.