2F4-OS-04-8

データ分布の独立性に基づくクラスタリングの実験的特性分析

Experimental characteristic analysis of clustering based on the independence of the data distribution

西垣 貴央[*1] 小野田 崇[*2]
Takahiro Nshigaki 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

We have proposed a clustering method based on Independent Component Analysis. This algorithm estimates the each independent semantic information from the observation data, and classifies based on distance metric defined by the independent semantic information. In this paper, we use Los Angeles Times dataset to evaluate the proposed method, and we show the feature of the independent semantic information. Consequently, we show that the proposed clustering method obtained higher clustering accuracy than k-means clustering initialized by Independent Component Analysis.

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，Web ページや電子ニュースなどのインターネット利用の一般化に伴い，個人のハードディスクや Web 上には文書や画像などの電子データが多量に蓄積されている．このような膨大な量のデータから，ユーザが欲しいデータを探し出すことは非常に困難である．膨大な量のデータからユーザが必要なデータの発見を容易にするためには，ユーザに提示するデータが何らかの方法でグループ化されていることが望ましい．データのグループ化をする方法として，一般に類似した内容のデータを同じグループにするクラスタリング手法が利用されている．膨大な量のデータを人が一つ一つ確認し，類似の内容と判断してデータをグループ分けすることは事実上不可能である．特に Web ページのように，日々膨大な量のデータが増加する場合，人の手でクラスタリングを行うことは非常に難しい．このような状況の下，現在では日々増加し続ける膨大な量のデータに対するクラスタリングは計算機によって行われており，k-means 法 [1] が広く用いられている．

k-means 法を用いて新聞記事のような文書データをグループ化する場合を考える．文書データはその文書で使用されている単語の頻度が行列で表現されており，多くの文書で使用される単語に該当する値は高い値に，あまり使われない単語に該当する値は低い値となる．そういったデータを k-means 法でグループ化した結果のイメージを図 1 に示す．図 1 中の小さい丸で示したものがグループ化したい単語データを示し，この単語データを 2 つのグループに分ける．図 1 では，左側のグループに多くの単語データが含まれていることが分かる．このグループが，ほとんどの文書で使用されている単語データを示しており，右側のグループに含まれるデータが，あまり使われることのない珍しい単語データを示している．このように，k-means 法では多くの文書で使用される単語データが集まったグループと，あまり使われることのない珍しい単語データが集まったグループに分けられる．

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

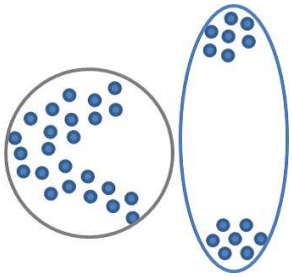

図 1: k-means 法のグループ化のイメージ

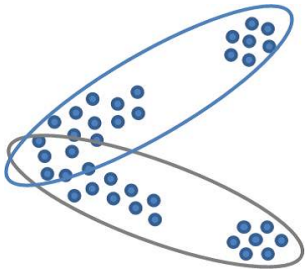

図 2: 異なる情報に基づくグループ化のイメージ

一方で，我々が提案したデータ分布の独立性に基づいてクラスタリングを行う方法 [2] を適用した場合を図 2 に示す．これは与えられた文書単語データから，2 つの異なる話題を独立な潜在情報として推定し，推定した話題に基づいて単語データをグループ化している．得られたグループは，2 つの話題の共通の単語データを含み，かつその話題でのみ使われる珍しい単語データも含むものとなっている．

本論文では，提案したデータ分布の独立性に基づいてクラスタリングを行う手法 [2] と，参考文献 [3, 4] で提案された手法を，Los Angeles Times に適用しその結果の比較を行い，提案手法の特徴を報告する．

以下，2 章で関連研究として k-means 法の初期値を ICA によって決定する手法を紹介し，3 章では提案した手法について述べる．4 章で提案手法を Los Angeles Times のデータに適用した結果と考察を行い，最後に 5 章で本研究のまとめを行う．

2. 関連研究

本論文で提案手法と比較を行う手法として参考文献 [3, 4] で提案された k-means 法の初期値を ICA によって決定する方法について紹介する．

k-means 法のクラスタ中心の初期値に，独立成分分析 (ICA：Independent Component Analysis) [5, 6] によって推定された独立成分と観測データ点とのコサイン類似度が最も大きいデータを選択することで，生成されるクラスタがより独立なものになることを期待した方法である．ここでは，参考文献

[3, 4] で提案された k-means 法の初期値を ICA で決定する手法について簡単に紹介する．

k-means 法は，任意の k 個の初期クラスタ中心を一様にランダムに選択する．参考文献 [3, 4] で提案された k-means 法の初期値を ICA で決定する方法では，以下のアルゴリズムに従って初期クラスタ中心を決定する．これによって，最終的に得られる各クラスタの独立性を期待した．

1. 観測データ集合 $\mathbf{X}$ から独立成分 $\mathbf{ic}_j, j \in \{1, \cdots, k\}$ を ICA によって推定する．
2. ステップ 1. で推定した独立成分 $\mathbf{ic}_j$ と最もコサイン類似度の値が大きい観測データを初期クラスタ中心 $\mathbf{c}_j$ とする．コサイン類似度は次の式で表される．

$$Sim(\mathbf{x}_i, \mathbf{ic}_j) = \frac{\mathbf{x}_i \cdot \mathbf{ic}_j}{||\mathbf{x}_i|| \cdot ||\mathbf{ic}_j||}, \quad i \in \{1, \cdots, n\}, \quad j \in \{1, \cdots, k\}. \tag{1}$$

3. ステップ 2. をクラスタの数 k 回だけ行い，初期クラスタ中心 $\mathbf{c}_j, j \in \{1, \cdots, k\}$ を得る．
4. 後は通常の k-means 法を行う．

しかし，初期値を決定した後は通常の k-means 法を行うために，最終的に得られるクラスタ間に独立性は保証されていない．

3. 提案手法

本章では，[2] で提案したデータ分布の独立性に基づいてクラスタリングを行う方法について簡単に述べる．以下、記号の小文字はスカラーを、小文字太字はベクトルを、大文字太字は行列を表す。

3.1 提案手法の概要

観測されたデータ（クラスタリングを行いたいデータ）$\mathbf{x}_{1,\ldots,n}$ は各属性 $\mathbf{c}_{1,\ldots,m}$ の値による表現され，またこの観測データは独立な潜在情報 $\mathbf{s}_{1,\ldots,k}$ の線形和で表現される．データを表現する潜在情報を求め，各データが最も属する可能性が高い潜在情報に分類する．

3.2 提案手法における潜在情報と観測データの表現方法

潜在情報は，属性による表現方法と，観測データによる表現方法が存在する．属性による潜在情報の表現方法は，各属性 $\mathbf{c}$ が潜在情報 $\mathbf{s}$ を特定する力を “潜在情報での属性の重要度” と呼ぶ値 $\mathbf{V}(\mathbf{s}, \mathbf{c})$ によって表す．この潜在情報での属性の重要度は，属性 $\mathbf{c}$ がその潜在情報をどの程度特定するのかを示す行列である．観測データによる潜在情報の表現方法は，各観測データ $\mathbf{x}$ が潜在情報 $\mathbf{s}$ を特定する力を “潜在情報での観測データの重要度” と呼ぶ値 $\mathbf{U}(\mathbf{s}, \mathbf{x})$ によって表す．この潜在情報での観測データの重要度は，観測データ $\mathbf{x}$ がその潜在情報をどの程度特定するのかを示す行列である．観測データの表現には，属性による表現と，潜在情報による表現方法が存在する．属性による観測データの表現方法は，各観測データ $\mathbf{x}$ を，観測データ中に含まれる属性 $\mathbf{c}$ が観測データ中でどの程度の強さを持つのかを示す数値 “観測データ中での属性の強度”$\mathbf{R}(\mathbf{x}, \mathbf{c})$ で表現される．この観測データ中での属性の強度は，属性 $\mathbf{c}$ がその観測データでどの程度の強さなのかを示す行列である．

潜在情報による観測データの表現方法は，各観測データ $\mathbf{x}$ 中での各潜在情報 $\mathbf{s}$ の強さを示す数値 “観測データにおける潜在情報の強度”$\mathbf{A}(\mathbf{x}, \mathbf{s})$ で表現される．この観測データにおける潜在情報の強度は，潜在情報 $\mathbf{s}$ がその観測データでどの程度の強さなのかを示す行列である．このとき，観測データは潜在情報の線形和で表現できる．潜在情報数を k 個とすると，各観測データ $\mathbf{x}$ は，“観測データにおける潜在情報の強度”$\mathbf{A}(\mathbf{x}, \mathbf{s})$ を用いて以下のように表現される．

$$\mathbf{x}_i = a_{(\mathbf{x}_i, \mathbf{s}_1)} \cdot \mathbf{s}_1 + a_{(\mathbf{x}_i, \mathbf{s}_2)} \cdot \mathbf{s}_2 + \cdots + a_{(\mathbf{x}_i, \mathbf{s}_k)} \cdot \mathbf{s}_k$$

ここで $a_{(\mathbf{x}_i, \mathbf{s}_1)}$ は，観測データ $\mathbf{x}_i$ における潜在情報 $\mathbf{s}_1$ の強度を示す値である．

観測データの属性による表現と，観測データの潜在情報による表現の間には，次の関係がある

$$\sum_{\text{属性 } \mathbf{c}} \mathbf{R}(\mathbf{x}, \mathbf{c}) \cdot \mathbf{A}(\mathbf{x}, \mathbf{s}) = \sum_{\text{属性 } \mathbf{c}} \mathbf{R}(\mathbf{x}, \mathbf{c}) \cdot \mathbf{V}(\mathbf{s}, \mathbf{c}) \tag{2}$$

重要度がその潜在情報での固有の属性の組み合わせに着目するのに対して，強度は観測データ中での潜在情報の組み合わせの多さを示すものといえる．

提案手法では，この “観測データにおける潜在情報の強度”$\mathbf{A}(\mathbf{x}, \mathbf{s})$ に基づいて各観測データがどの潜在情報から派生しているのかを決定する．

3.3 潜在情報の独立性と潜在情報の観測データへの集中度

潜在情報間には独立性を仮定しているが，独立性を評価する指標は複数存在する．ここでは，独立性を評価する指標として代表的な指標の一つである高次統計量の尖度（同一の平均・分散を持つ正規分布との 4 次モーメントの差） [6] を使用する．

“潜在情報の観測データへの集中度” を以下で定義する．

$$\sum_{i=1}^{n} u^4_{(\mathbf{s}_j, \mathbf{x}_i)} \cdot \mathbf{r}_i - 3 \cdot \left(\sum_{i=1}^{n} u^2_{(\mathbf{s}_j, \mathbf{x}_i)} \cdot \mathbf{r}_i\right)^2 \tag{3}$$

この値が大きいということは，大半の観測データの重要度は 0 の近くにあり，少数の観測データの重要度のみが大きい値を持つことを示す．つまり，少数の観測データのみでその潜在情報が特徴付けられることを示している．

潜在情報における相互の独立性の強さは，各潜在情報における集中度の二乗和によって表現できる．この値が大きい場合，潜在情報間の独立性（＝非関連性）が高くなる．

3.4 提案手法のアルゴリズム

提案手法では，求められる潜在情報の本数 k が与えられており，各潜在情報は正規直交性を満たしている．この時，各潜在情報の重要度を座標軸とする k 次元空間を考えて，観測データに対応する点の近くに，観測データ中に現れる属性に対応する点もあるという最適な至近配置を実現する．そして，最適至近配置の中で，各潜在情報の独立性が最大となる配置が提案手法で求める「独立な潜在情報」となる．最後に，求めた独立な潜在情報による観測データの表現 “観測データにおける潜在情報の強度”$\mathbf{A}(\mathbf{x}, \mathbf{s})$ に基づいて観測データをクラスタリングする．この配置決定およびクラスタリングのアルゴリズムを以下に示す．

1. 観測データ集合 $\mathbf{X}$ を，観測データを行に，その属性を列にとった行列 $\mathbf{R}(\mathbf{x}, \mathbf{c})$ として整理する．
2. “観測データ中の属性の強度”$\mathbf{R}(\mathbf{x}, \mathbf{c})$ によって表現された観測データ集合を，各属性の平均値をそれぞれの値から引くことにより正規化した $\hat{\mathbf{R}}(\mathbf{x}, \mathbf{c})$ を求める．

3. ステップ 2. で求めた $\hat{\mathbf{R}}(\mathbf{x},\mathbf{c})$ を次のように分解する．$\mathbf{U}^{\mathrm{T}}\cdot\hat{\mathbf{R}}\cdot\mathbf{V}=\mathbf{D}\Longleftrightarrow\hat{\mathbf{R}}=\mathbf{U}\cdot\mathbf{D}\cdot\mathbf{V}^{\mathrm{T}}$．$\mathbf{U}$ と $\mathbf{V}$ は潜在情報での観測データの重要度と属性の重要度を示す行列である．また $\mathbf{D}$ は特異値の対角行列であり，その大きさの順に k 個の成分を抜き出し，$\mathbf{U}_k$，$\mathbf{D}_k$，$\mathbf{V}_k$ を作成する．

4. ステップ 3. で得られた $\mathbf{U}_k$，$\mathbf{D}_k$ を用いて，各潜在情報間の独立性が最大となるときの，“観測データにおける潜在情報の強度”$\mathbf{A}(\mathbf{x},\mathbf{s})$ を，FPICA [7] に基づいた以下のアルゴリズムによってを求める．

   4.1 初期化：$\mathbf{W}_{old}=\mathbf{U}_k\cdot\mathbf{D}_k$

   4.2 終了判定：$||\mathbf{W}-\mathbf{W}_{old}||$ もしくは $||\mathbf{W}+\mathbf{W}_{old}||$ が閾値以下ならば，終了ステップ 4.5. へ

   4.3 $\mathbf{W}$ の更新：$\mathbf{W}_{old}=\mathbf{W}$,

   $$\mathbf{W}=\hat{\mathbf{R}}\cdot\left(\mathbf{W}_{old}^{\mathrm{T}}\cdot\hat{\mathbf{R}}\right).^{3}-3\cdot\mathbf{W}_{old}. \quad (4)$$

   $\mathbf{W}^{\mathrm{T}}$ は行列の転置を表し，$\mathbf{R}.^{3}$ は行列要素の 3 乗を表す．

   4.4 $\mathbf{W}$ の正規化：$\mathbf{W}=\mathbf{W}/||\mathbf{W}||$ ステップ 4.2. へ

   4.5 終了：求めた $\mathbf{W}$ を $\mathbf{W}^{*}$ として返す．

5. ステップ 4. で求めた $\mathbf{W}^{*}$ の逆行列が，“観測データにおける潜在情報の強度”$\mathbf{A}(\mathbf{x},\mathbf{s})$ である．$(\mathbf{A}=(\mathbf{W}^{*})^{-1})$

6. ステップ 5. で求めた “観測データにおける潜在情報の強度”$\mathbf{A}(\mathbf{x},\mathbf{s})$ の値によって，各観測データがどの潜在情報から派生しているのかを決定する．それにより得られるクラスタを次式によって表す．

$$\mathcal{C}_j=\{\mathbf{x}_i\mid\arg\max_{\mathbf{s}}a(\mathbf{x}_i,\mathbf{s}_j)\},$$
$$i\in\{1,\cdots,n\},\quad j\in\{1,\cdots,k\}. \quad (5)$$

4. Los Angeles Times への適用

ここでは，Los Angeles Times のデータ “la12” [8] に対して提案手法と，k-means 法の初期値を ICA で決定する方法 [3, 4] のそれぞれを適用した．その結果の比較分析を行う．このデータは，Los Angeles Times の 1989 年と 1990 年の記事で，“Entertainment”，“Financial”，“Foreign”，“Metro”，“National”，“Sports” の 6 つのカテゴリに分けられている．このデータは，データ数（文書数）6279 で属性数（単語数）31472 のデータである．

4.1 提案手法によるクラスタリング結果

提案手法によって推定した潜在情報集合 $\mathbf{S}$ は “潜在情報での観測データの重要度” $\mathbf{U}(\mathbf{s},\mathbf{x})$ による表現される．各潜在情報 $\mathbf{s}_j, j\in\{1,\cdots,6\}$ で，$\mathbf{U}(\mathbf{s},\mathbf{x})$ の値が最も大きい値を示す際の観測データ，ここでは文書 $\mathbf{x}$ を表 1 に示す．また，この文書 $\mathbf{x}$ がどの正解クラスタに分類されているのかについても表 1 に示す．表 1 より，推定した “潜在情報での文書の重要度”$\mathbf{U}(\mathbf{s},\mathbf{x})$ が各潜在情報で最も大きい時の文書 $\mathbf{x}$ の正解クラスタは，1 つも重複がなく全て別のクラスタに分かれていることがわかる．このことから推定した各潜在情報はそれぞれのクラスタを示していることが考えられる．つまり，潜在情報 $\mathbf{s}_1$ は “Foreign” を，潜在情報 $\mathbf{s}_2$ は “Entertainment” を示しており，同様に潜在情報 $\mathbf{s}_3$ は “Metro”，潜在情報 $\mathbf{s}_4$ は “Financial”，潜在情報 $\mathbf{s}_5$ は “Sports”，潜在情報 $\mathbf{s}_6$ は “National” とそれぞれのクラスタを示していると考えられる．

表 1: “la12” の各潜在情報 $\mathbf{s}$ における $\mathbf{U}(\mathbf{s},\mathbf{x})$ の値が最大となる際の $\mathbf{x}_i$ とその正解クラスタ

潜在情報 $\mathbf{s}_j$	$\max\mathbf{U}(\mathbf{s},\mathbf{x})$ 時の $\mathbf{x}_i$	$\mathbf{x}_i$ の正解クラスタ
$\mathbf{s}_1$	$\mathbf{x}_{1537}$	“Foreign”
$\mathbf{s}_2$	$\mathbf{x}_{1203}$	“Entertainment”
$\mathbf{s}_3$	$\mathbf{x}_{6141}$	“Metro”
$\mathbf{s}_4$	$\mathbf{x}_{3420}$	“Financial”
$\mathbf{s}_5$	$\mathbf{x}_{2931}$	“Sports”
$\mathbf{s}_6$	$\mathbf{x}_{4200}$	“National”

また，潜在情報の別の表現方法である “潜在情報での属性の重要度”$\mathbf{V}(\mathbf{s},\mathbf{c})$ から潜在情報について分析する．ここでは属性は単語を示す．各潜在情報 $\mathbf{s}_j, j\in\{1,\cdots,6\}$ の時，$\mathbf{V}(\mathbf{s},\mathbf{c})$ の値が高い 10 個の単語 $\mathbf{c}_t, t\in\{1,\cdots,10\}$ を表 2 に示す．表 2 より，潜在情報 $\mathbf{s}_1$ では，$\mathbf{V}(\mathbf{s},\mathbf{c})$ の値が高い単語に soviet, afghanistan, israel, foreign などが含まれており，潜在情報 $\mathbf{s}_1$ は “Foreign” を示していると考えられる．同様に，潜在情報 $\mathbf{s}_2$ は，art や music，film などが単語上位にあることから “Entertainment” を示していると考えられ，また，潜在情報 $\mathbf{s}_3$ は，単語上位に car，arrest，diego，san などがあり，Los Angeles 近郊のことであると考えられるので “Metro” を示していると考えられる．潜在情報 $\mathbf{s}_4$ は単語上位に million，earn，bank など含まれていることから “Financial” を示していると推測できる．さらに潜在情報 $\mathbf{s}_5$ は，単語上位に game，team，coach などが含まれていることから “Sports” を，潜在情報 $\mathbf{s}_6$ は，1989 年や 1990 年当時のアメリカの大統領は George Herbert Walker Bush を示すと考えられる bush という単語が 1 位にあることや，budget，president を示すであろう presid，insurance を示すであろう insur などが含まれているため，“National” を示していると考えられる．

表 2 の分析により，表 1 から得られた各潜在情報 $\mathbf{s}$ はどのクラスタを示すのかという推測は正しいと考えられる．

4.2 k-means 法の初期値を ICA で決定する方法によるクラスタリング結果

k-means 法の初期値を ICA で決定する方法 [3, 4] によって，最終的に得られたクラスタ中心からのユークリッド距離が最も小さい文書データとその文書の正解クラスタ名を表 3 に示す．生成されたクラスタ中心に最もユークリッド距離が小さい文書データは，そのクラスタを代表するデータを示していると考えられる．クラスタを代表する文書データの正解クラスタ名を調べることで，そのクラスタがどのようなデータの集合であるのか推測できる．

表 3 より，最終的に得られたクラスタ中心からのユークリッド距離が，最も小さい文書データの正解クラスタはクラスタ $\mathcal{C}_4$ とクラスタ $\mathcal{C}_6$ において “Financial” が重複してしまっており，“National” が存在しない．正解クラスタが重複していない他の 4 つのクラスタである，クラスタ $\mathcal{C}_1$ は “Foreign” を，クラスタ $\mathcal{C}_2$ は “Entertainment” を示し，同様にクラスタ $\mathcal{C}_3$ は “Metro”，クラスタ $\mathcal{C}_5$ は “Sports” のクラスタを示していると考えられる．しかし，クラスタ $\mathcal{C}_4$ とクラスタ $\mathcal{C}_6$ は “Financial” のデータと正解クラスタが存在しない “National” のデータとが混在したクラスタとなっていると考えることができる．

この結果を表 1 と比較すると，生成したクラスタの代表

表 2: “la12” の各潜在情報 $\mathbf{s}$ において $\mathbf{V}(\mathbf{s},\mathbf{c})$ の値が大きい単語 $\mathbf{c}$ の上位 10 個

単語上位 $\mathbf{c}_t$	$\mathbf{s}_1$	$\mathbf{s}_2$	$\mathbf{s}_3$	$\mathbf{s}_4$	$\mathbf{s}_5$	$\mathbf{s}_6$
$t=1$	soviet	aleen	polic	million	game	bush
$t=2$	afghanistan	macmin	bush	earn	scor	counti
$t=3$	israel	art	counti	quarter	lead	presid
$t=4$	foreign	entertain	car	bank	team	budget
$t=5$	govern	report	arrest	rose	plai	citi
$t=6$	militari	morn	offic	compani	season	propos
$t=7$	bush	nation	kill	revenu	rebound	insur
$t=8$	afghan	music	diego	billion	coach	school
$t=9$	datelin	intern	orang	corpor	league	house
$t=10$	israe	film	san	net	fullerton	feder

となるデータのクラスタは提案手法では 1 つも重複がなく全て別のクラスタに分かれているのに対し，k-means 法の初期値を ICA で決定する方法によるクラスタリングした場合では，6 つのクラスタに分けることができていない．このことは，NMI [9] の結果からも見ることができ，提案手法が NMI の値が 0.4355 に対して，k-means 法の初期値を ICA で決定する方法の NMI の値は 0.3472 であった．また，提案手法では “潜在情報での観測データの重要度”$\mathbf{U}(\mathbf{s},\mathbf{x})$ や “潜在情報での属性の重要度”$\mathbf{V}(\mathbf{s},\mathbf{c})$ を仮定することによって，独立な潜在情報を推定するので，生成したクラスタや潜在情報の分析が可能である．一方で，k-means 法の初期値を ICA で決定するクラスタリング方法ではそういった分析を行うことはできない．

以上のことから，提案手法は k-means 法の初期値を ICA で決定する方法より，新聞情報のようなオーバーラップが多いようなデータの分析に向いていると考えることができる．

5. まとめ・今後の課題

観測したデータのデータ分布から独立な潜在情報を推定し，その推定した潜在情報に基づいてクラスタリングを行う方法を提案した．その提案した方法を，Los Angeles Times のデータに適用し，提案手法によって推定した潜在情報がどのような特性を持っているのかを分析した．さらにデータ分布の独立性に着目した関連研究である，k-means 法の初期値を ICA によって決定する方法と結果を比較した．その結果，Los Angeles Times などのような新聞データには，k-means 法の初期値を ICA で決定する方法を用いるよりも提案手法が有効であることを示した．

今後の課題には，現在の手法にユーザによる制約（must-link や cannot-ling）を導入し，その制約を満たす中で，最もデータの独立性が高くなるクラスタを生成するアルゴリズムについて検討することである．

表 3: “la12” に参考文献 [3, 4] の手法で得られたクラスタ中心に最も近い文書データとその正解クラスタ

クラスタ $\mathcal{C}_j$	中心から最近傍の $\mathbf{x}_i$	$\mathbf{x}_i$ の正解クラスタ
$\mathcal{C}_1$	$\mathbf{x}_{1537}$	“Foreign”
$\mathcal{C}_2$	$\mathbf{x}_{1203}$	“Entertainment”
$\mathcal{C}_3$	$\mathbf{x}_{4448}$	“Metro”
$\mathcal{C}_4$	$\mathbf{x}_{3420}$	“Financial”
$\mathcal{C}_5$	$\mathbf{x}_{2931}$	“Sports”
$\mathcal{C}_6$	$\mathbf{x}_{4504}$	“Financial”

参考文献

[1] Christopher M. Bihop, “Pattern Recognition and Machine Learnin”, Springer 2006.

[2] 西垣 貴央, 小野田 崇, “高次独立性に基づくクラスタリング”, Annual Conference of the Japanese Society for Artificial Intelligence, 4F1-OS-5-3, 2012.

[3] Takashi Onoda, Miho Sakai, Seiji Yamada, “Careful Seeding based on Independent Component Analysis for k-means Clustering”, International Conference on Web Intelligence and Intelligent Agent Technology, 2010.

[4] 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”, 知能システム科学専攻修士論文, 2011.

[5] A. Hyvärinen, E. Oja, “Independent component analysis: algorithms and applications”, Neural Networks, vol.13, p.411-430, 2000.

[6] A. Hyvärinen, J. Karhunen, E. Oja, “Independent Component Analysis”, John Wiley & Sons, 2001.

[7] A. Hyvarinen, E.Oja, “A Fast Fixed-Point Algorithm for Independent Component Analysis”, Neural Computation, vol.9, no.7, p. 1483-1492, 1997.

[8] George Karypis, “CLUTO - A Clustering Toolkit”, http://glaros.dtc.umn.edu/gkhome/views/cluto/ , Department of Computer Science and Engineering, University of Minnesota, 2002.

[9] Hao Cheng, Kien A. Hua, Khanh Vu, “Constrained locally weighted clustering”, Proceedings of the VLDB Endowment, vol.1 no.1, 2008.

[10] S. Zhong, J. Ghosh, “A comparative study of generative models for document clustering”, Data Mining Workshop on Clustering High Dimensional Data and Its Applications, 2003.

[11] TREC. Text REtrieval conference. http://trec.nist.gov, 1999.