3B3-OS-10a-1

高次独立性に基づく制約付きクラスタリング

Constrained Clustering based on high independence

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

*1東京工業大学
Tokyo Institute of Technology

*2電力中央研究所
Central Research Institute of Electric Power Industry

These days, the classified by independence semantic analysis based on independence of the data distribution have been proposed. But the clustering method is unsupervised learning, so in many cases a clustering result of a large data and user desired result will not be the same. In this paper, we propose the method of adding user constraints to the classified by independence semantic analysis. This method first classified the data without constraints then reclassified the clustering result by using constraints. We use Los Angeles Times datasets to evaluate the proposed method, and we show the proposed method is valid form results of experiments using this benchmark data.

1. はじめに

近年，Web ページや電子ニュース等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，ユーザが望む情報を探しだすことは非常に困難である．そこでユーザの目的に適合したデータの発見を容易にするために，ユーザに提示するデータをある種のグループに分割して提示する，クラスタリングと呼ばれる方法が一般的に利用されている．このクラスタリングは，多量のデータをある視点に基づいて分析し，いくつかのグループに計算機が自動的に分類する方法である．しかし多くのデータは，複数のグループに属するようなもので，容易には分類できないものが多い．特に新聞データのような文書－単語行列となったデータは，容易にまた完全には分類できないことが多い．そこでそれらの文書が，どのようなトピック（話題）から派生した文書であるのか分かれば，ユーザの求める目的の文書の発見も容易になると考えられる．そのように文書データを分析する方法として，文書データには正規分布するトピックが存在すると仮定して分析する方法 [1] や，文書データには正規分布でない独立なトピックが存在すると仮定して分析する方法が存在する．前者のデータに正規分布を仮定して分析する方法については多くの研究がなされている．

本稿では，後者のデータに正規分布でない独立なトピックが存在すると仮定して分析する方法の 1 つである，データ分布の高次独立性の視点からデータの分析を行う方法 [2, 3] について考える．この方法で得られる結果は，データ分布の独立性に基づいてデータをクラスタリングしている．しかしその結果が，常にユーザの目的に適したデータに分割を行うとは限らない．そこでこの方法に，ユーザの制約を満たしつつ，かつ高次独立なクラスタを生成するクラスタリング方法について議論する．

以下，2 章でデータ分布の高次独立性に基づいた手法について紹介し，3 章でその方法に制約を加える提案手法について述べる．4 章では提案した手法を Los Angeles Times のデータに適用した結果について示す．最後に 5 章でまとめと今後の展望について述べる．

連絡先: 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, nishigaki@ntt.dis.titech.ac.jp

2. 関連研究

本章では，データ分布の高次独立性に基づいてデータを分析・分類する手法 [2, 3] について簡単に述べる．以下，記号の小文字はスカラー，小文字太字はベクトル，大文字太字は行列を表す．

観測データ（分析・分類を行いたいデータ）$\mathbf{x}_{1,\ldots,n}$ は，属性 $\mathbf{c}_{1,\ldots,m}$ の値により表現され，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と仮定する．そして，データを表現する独立な潜在情報を求め，各データがどの潜在情報からどの程度影響を受けているのか分析し，それに基づいて分類する．

潜在情報は，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” と呼ぶ値の行列 $\mathbf{V}(\mathbf{s},\mathbf{c})$ による表現と，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” と呼ぶ値の行列 $\mathbf{U}(\mathbf{s},\mathbf{x})$ の 2 つの表現によって表す．同様に，観測データは各属性 $\mathbf{c}$ が観測データ $\mathbf{x}$ の中での強さを “観測データでの属性の強度” と呼ぶ値の行列 $\mathbf{R}(\mathbf{x},\mathbf{c})$ による表現と，各潜在情報 $\mathbf{s}$ が観測データ $\mathbf{x}$ の中での強さを “観測データでの潜在情報の強度” と呼ぶ値の行列 $\mathbf{A}(\mathbf{x},\mathbf{s})$ の 2 つの表現によって表す．このとき，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．

$$\sum_{\text{属性 }\mathbf{c}} \mathbf{R}(\mathbf{x},\mathbf{c}) \cdot \mathbf{A}(\mathbf{x},\mathbf{s}) = \sum_{\text{属性 }\mathbf{c}} \mathbf{R}(\mathbf{x},\mathbf{c}) \cdot \mathbf{V}(\mathbf{s},\mathbf{c}) \qquad (1)$$

重要度がその潜在情報での固有の属性の組み合わせに注目する，一方で，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．

また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，$\mathbf{A}(\mathbf{x},\mathbf{s})$ を用いて次のように表す．

$$\mathbf{x}_i = a_{(\mathbf{x}_i,\mathbf{s}_1)} \cdot \mathbf{S}_1 + a_{(\mathbf{x}_i,\mathbf{s}_2)} \cdot \mathbf{S}_2 + \cdots a_{(\mathbf{x}_i,\mathbf{s}_k)} \cdot \mathbf{S}_k \qquad (2)$$

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．

この手法では，この $\mathbf{A}(\mathbf{x},\mathbf{s})$ 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．この手法のアル

ゴリズムを簡単に以下に示す．この手法では，求める潜在情報の数 k は既に知っているものとする．

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. $\mathbf{R}(\mathbf{x},\mathbf{c})$ を正規化し，$\hat{\mathbf{R}}(\mathbf{x},\mathbf{c})$ 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．$\mathbf{U}^{\mathrm{T}}\cdot\hat{\mathbf{R}}\cdot\mathbf{V}=\mathbf{D}\iff\hat{\mathbf{R}}=\mathbf{U}\cdot\mathbf{D}\cdot\mathbf{V}^{\mathrm{T}}$．$\mathbf{U}$ 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 [4] に基づいたアルゴリズムによって求める．
5. ステップ 4. 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．

$$\mathcal{C}_j=\{\mathbf{x}_i\mid\arg\max_{\mathbf{s}}a(\mathbf{x}_i,\mathbf{s}_j)\},\quad i\in\{1,\cdots,n\},\quad j\in\{1,\cdots,k\}.\tag{3}$$

この手法によって観測データ集合を独立なクラスタにクラスタリングすることが可能である．

しかし，この手法では観測データ集合を独立なクラスタにクラスタリングを行っただけで，その結果がユーザの求めているクラスタと合致しているとは限らない．そこで，この手法にユーザ制約を加え，ユーザの求めている結果に近づけることを考える必要がある．

3. 提案手法

この章では，ユーザの制約を満たしつつ，独立性が高いクラスタを得る手法について提案する．ユーザ制約とはここでは must-link 制約のことを指す．must-link 制約とは，異なるクラスタに分類されている 2 つのデータをユーザが選択し，それらが最終的に同じクラスタとなるようにすることである．

前章で紹介したデータ分布の高次独立性に基づくクラスタリング手法に must-link 制約を加える．以下にそのアルゴリズムについて示す．

0. 前章で紹介したクラスタリングを行い，must-link 制約を加えたいデータ 2 つ $\mathbf{x}_i$ と $\mathbf{x}_\ell$ を選択する．
1. 制約を満たす潜在情報となる軸を生成する
   (a) $\mathbf{x}_i$ および $\mathbf{x}_\ell$ は m 個の属性 $\mathbf{c}$ によって表現されている（“観測データでの属性の強度”）ので，制約を満たす軸を $\mathbf{z}_1=(\mathbf{r}_i+\mathbf{r}_\ell)/2$ とする．ここで，$\mathbf{r}_i$ はデータ $\mathbf{x}_i$ を m 個の属性によって表現したベクトルである．
   (b) この軸が制約を満たす範囲で独立性が最大となるように更新する．更新方法は FPICA [4] を用いて行う．
2. 2 つ目から k 個までの軸は，ステップ. 0 で既に求めた潜在情報の中から，最も独立性の低い（尖度の絶対値が低い）ものを選択する．
3. 選択した軸を制約が満たされる範囲で独立性が最大となるように FPICA [4] を用いて更新する．
4. k 個の軸全てをそれぞれ，制約が満たされる範囲で独立性が最大となるように FPICA [4] を用いて更新する．
5. 全ての軸が動かなくなれば，それを新たな潜在情報として終了する．

この方法によって，ユーザによる must-link 制約を満たす範囲で，独立性が高いクラスタを得ることができる．

4. Los Angeles Times への適用

ここでは，Los Angeles Times のデータ “la12”[5, 6] に対して，データ分布の高次独立性に基づくクラスタリングに制約を加えた提案手法を適用した結果を示す．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 つのトピックに分けられており，データ数（文書数）6279 で属性数（単語数）31472 のデータである．

適用実験では，次の 2 つでそれぞれの手法の結果を比較する．1）潜在情報（トピック）の数が 6 つと分かっている場合に，a）制約なしの手法でクラスタリングしたもの（関連研究の手法）と，b）must-link 制約を加えてクラスタリングしたもの（提案手法）とでの結果の比較を行う．もう 1 つは，2）潜在情報の数が分かっていない場合に，関連研究の手法で 10 のクラスタに分類後に，a）制約なしの手法で 9 つのクラスタに分けたものと，b）must-link 制約を加えて 9 つのクラスタに分けたものとでの結果の比較を行う．

4.1 潜在情報の数が既知の場合

潜在情報の数が 6 つと分かっている場合，制約なしの手法でクラスタリングしたものと，制約ありでクラスタリングしたものを正規化相互情報量（NMI：Normalized Mutual Information）[7] を用いて比較する．NMI とは，0 から 1 の値を取り，値が大きいほど生成されたクラスタ集合が正解クラスタ集合に類似していることを示す．つまり，NMI が 1 の時，正解クラスタ集合と生成したクラスタ集合は全く同様の分類がされているということである．制約には 149 番目と 3420 番目の文書に must-link 制約をつけた．これらの文書を選択した理由は，149 番目の文書は “Financial” に属する文書なのだが，制約なしの手法でクラスタリングしたものでは “National” に誤分類されてしまっており，そこで “Financial” に最も影響を受けている 3420 番目の文書と 149 番目の文書に must-link 制約をつけた．

NMI の結果は，制約なしの手法では 0.4355，制約ありの提案手法では 0.4827 となり，制約ありの提案手法のほうが制約なしの手法よりも高い値を示していることがわかる．これはつまり，制約ありの提案手法によって得られたクラスタ集合のほうが，制約なしの手法によって得られたクラスタ集合よりも正解クラスタ集合に類似しているということを表す．

この時の制約なしの手法によって得られた “各潜在情報 $\mathbf{s}$ での属性（単語）$\mathbf{c}$ の重要度 $\mathbf{V}(\mathbf{s},\mathbf{c})$” の値が高い単語 10 個を表 1 に，制約ありの手法によって得られたものを表 2 に示す．

表 1 をみると，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

表 1: 制約なしの手法による “la12” の $\mathbf{V}(\mathbf{s}, \mathbf{c})$ の値が大きい単語の上位 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

表 2: 制約ありの提案手法による “la12” の $\mathbf{V}(\mathbf{s}, \mathbf{c})$ の値が大きい単語の上位 10 個

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

情報 $\mathbf{s}_1$ は “Foreign” を示していると考えられる．同様に他の潜在情報もそれぞれ “Entertainment”，“Metro”，“Financial”，“Sports”，“National” を示している単語で構成されていることがわかる．表 2 をみると，順番は入れ替わっているが潜在情報 $\mathbf{s}_1$ は “Financial” を示している単語で構成されており，この潜在情報でユーザの制約を満たしている．しかし，潜在情報 $\mathbf{s}_4$ と $\mathbf{s}_6$ では明らかに制約なしの手法の表 1 とは異なり，この 2 つの潜在情報は似た単語で構成されてしまっていることがわかる．これは，NMI の値が高くなるようなユーザの制約を満たすように潜在情報を変化させているため，得られたクラスタの独立性が，制約なしの手法の時よりも低くなったためであると考えられる．

4.2 潜在情報の数が未知の場合

本来ユーザは Los Angeles Times のようなデータが与えられた場合いくつのグループに分類するべきか分からないことの方が多い．そこでここでは，まず潜在情報の数を 10 として制約なしの手法のクラスタリングを行う．この時に，ユーザが与える must-link 制約を潜在情報の減少と解釈して，制約なしの手法で潜在情報の数を初めから 9 としたクラスタリング結果と，ユーザが制約を与えるデータを決め，制約ありの提案手法で潜在情報の数を 1 つ減らして行ったクラスタリング結果との比較を行う．

must-link 制約を与えるデータの選択方法は，潜在情報の数を 10 とした時の $\mathbf{V}(\mathbf{s}, \mathbf{c})$ の値が高い単語を比較して，同じトピックから派生したと思われる潜在情報 2 つを選択し，その選択した潜在情報にそれぞれ最も影響を受けている文書データ 2 つで，1904 番目と 283 番目のものでどちらも “Sports” に属するものを選択した．

初めから潜在情報の数を 9 とした制約なしの手法の時の潜在情報の属性の重要度 $\mathbf{V}(\mathbf{s}, \mathbf{c})$ の値が高い単語 10 個を表 3 に，提案手法によって制約を満たしている時の潜在情報の属性の重要度 $\mathbf{V}(\mathbf{s}, \mathbf{c})$ の値が高い単語 10 個を表 4 に示す．

表 3 では “Sports” と思われる潜在情報が $\mathbf{s}_3$ と $\mathbf{s}_8$ の 2 つにあるが，表 4 では，$\mathbf{s}_1$ の 1 つのみとなっている．このとき，制約を与えた 1904 番目と 283 番目の文書は，制約なしの方法の結果では異なる潜在情報 $\mathbf{s}_3$ と $\mathbf{s}_8$ に分類されている．制約ありの提案手法の結果では，制約を与えた 2 つのデータは同じクラスタになっているが，表 4 の潜在情報 $\mathbf{s}_4$ と $\mathbf{s}_6$ を見ると，これらの潜在情報は両方とも “Entertainment” を示すと考えられる単語で構成されてしまっている．これは提案手法では制約を満たす潜在情報を生成するために，その他の潜在情報間との独立性は制約なしの手法のものと比べて低くなっているからであると考えられる．以上より，潜在情報の数が未知の場合，ユーザが与える must-link 制約を潜在情報の減少と解釈した方法と，制約ありの提案手法で潜在情報の数を 1 つ減らして行った方法とでは，must-link 制約を優先するのか，潜在情報間の独立性を優先するのかで結果は大きく異なることが分かる．

また，潜在情報の数を 9 つでデータの分析を行うことによって得られる独立な潜在情報は，正解の 6 つをさらに細かく分類できる単語で構成されていることがわかった．例えば表 3 の潜在情報 $\mathbf{s}_2$ と $\mathbf{s}_7$ は大きく分けると “Financial” であるが，構成されている単語は大きく異なる．$\mathbf{s}_2$ は資金 “revenue” そのものに関する単語が多く，一方で $\mathbf{s}_7$ は株や投資 “stock” に関する単語が多い．同様に $\mathbf{s}_3$ と $\mathbf{s}_8$ も両方とも “Sports” を示す単語で構成されているが，同じ単語は “game” だけで，他は異なる単語で構成されている．詳しく見ると，$\mathbf{s}_3$ はサッカーの試合に関する単語 “lead” や “quarter”，“half” があり，$\mathbf{s}_8$ はサッカーそのものに関する単語 “team” や “bowl”，“coach”

表 3: 潜在情報の数を 9 とした制約なしの手法による $\mathbf{V}(\mathbf{s}, \mathbf{c})$ の値が大きい単語の上位 10 個

単語上位 $\mathbf{c}_t$	$\mathbf{s}_1$	$\mathbf{s}_2$	$\mathbf{s}_3$	$\mathbf{s}_4$	$\mathbf{s}_5$	$\mathbf{s}_6$	$\mathbf{s}_7$	$\mathbf{s}_8$	$\mathbf{s}_9$
$t = 1$	bush	million	scor	soviet	aleen	polic	stock	game	counti
$t = 2$	tower	earn	game	afghanistan	macmin	arrest	bank	team	citi
$t = 3$	senat	quarter	lead	israel	art	car	price	bowl	school
$t = 4$	reagan	revenu	rebound	foreign	entertain	offic	market	player	orang
$t = 5$	presid	net	league	afghan	report	kill	rate	coach	san
$t = 6$	budget	corpor	fullerton	union	morn	investig	bond	football	diego
$t = 7$	congress	rose	quarter	militari	nation	murder	index	season	student
$t = 8$	house	incom	half	kabul	intern	robberi	trad	plai	lo
$t = 9$	committe	billion	goal	govern	new	suspect	amp	super	angel
$t = 10$	white	loss	victori	israe	tv	driver	loan	counti	california

表 4: 潜在情報の数が 9 で制約ありの提案手法による $\mathbf{V}(\mathbf{s}, \mathbf{c})$ の値が大きい単語の上位 10 個

単語上位 $\mathbf{c}_t$	$\mathbf{s}_1$	$\mathbf{s}_2$	$\mathbf{s}_3$	$\mathbf{s}_4$	$\mathbf{s}_5$	$\mathbf{s}_6$	$\mathbf{s}_7$	$\mathbf{s}_8$	$\mathbf{s}_9$
$t = 1$	scor	counti	soviet	aleen	polic	aleen	bush	bank	million
$t = 2$	game	citi	israel	macmin	arrest	macmin	tower	loan	earn
$t = 3$	lead	orang	afghanistan	art	car	art	reagan	insur	quarter
$t = 4$	league	school	israe	entertain	kill	stock	senat	israel	revenu
$t = 5$	quarter	diego	palestinian	israel	offic	price	presid	sav	net
$t = 6$	rebound	san	afghan	stock	murder	market	aleen	amp	corpor
$t = 7$	goal	student	union	price	investig	report	macmin	deposit	rose
$t = 8$	half	lo	foreign	polic	robberi	entertain	budget	feder	incom
$t = 9$	shot	angel	govern	market	suspect	rate	house	l	loss
$t = 10$	victori	bush	militari	nation	brief	index	congress	rate	billion

などが多いことが分かる．このように，潜在情報の数を正解の数よりも多くすることで，文書をより詳しく分析することが可能である．

5. おわりに

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に，must-link 制約を加える方法について提案した．提案した方法を，Los Angeles Times のデータに適用し，潜在情報の数が既知の場合で NMI の比較と潜在情報を構成する単語の比較を行った．その結果，制約を加えたものは制約がないものよりも NMI の値は高くなっているが，潜在情報間の独立性は低下していることを示した．さらに潜在情報の数が未知であると仮定し，must-link 制約を満たしつつ潜在情報の数を減らした場合と，単に潜在情報の数を減らした場合とでの，推定した潜在情報を構成する単語の比較を行った．こちらの場合も must-link 制約によって生成される潜在情報は制約を満たすが，他の潜在情報間の独立性は低下していることがわかった．また潜在情報の数を正解の数よりも増やすとより詳細にデータを分析することができると考えられる．

今後の課題には，複数の must-link 制約や cannot-link 制約を導入するアルゴリズムについて検討することが挙げられる．

参考文献

[1] Scott Deerwester et al., “Indexing by latent semantic analysis”, Journal of the American Society for Information Science, Vol. 41, No. 6, pp.391-407, 1990.

[2] Takahiro Nishigaki and Takashi Onoda, “Indepnedence based Clusteirng”, 2012 Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), pp. 389-390, 2012.

[3] Takahiro Nishigaki and Takashi Onoda, “Clustering based on independent component”, 2012 International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 74-78, 2012.

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

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

[6] 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.

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