LF-015

# 高次元データに対して頑健な文書クラスタリング手法
# A Robust Text Clustering Method for High-Dimensional Data

金田 有二 † 上田 修功 †
Yuji Kaneda Naonori Ueda

## 1. はじめに

近年の電子テキストの増大に伴い，文書群を類似文書からなる複数のクラスタに分割する文書クラスタリング手法が多数提案されている．代表的な文書クラスタリング手法の 1 つに Spherical $K$-means (SpKmeans) [1] がある．SpKmeans は，コサイン類似度による $K$-means である．$K$-means は，文書のクラスタへの分割と，クラスタの代表ベクトルを逐次的に更新する手法である．$K$-means には，データが高次元の場合，低品質な局所最適解に陥りやすい問題点がある [2]．この問題点を解決するために，代表ベクトルに制約を設け，解の探索空間を限定する手法を提案する．精度の高いクラスタリングを実現するには，適切な制約を導入する必要がある．提案法では，文書の特性を利用した制約を導入する．

## 2. Spherical $K$-means

SpKmeans を簡単に説明する．

$N$ 個の文書 $\{d_n\}_{n=1}^N$ を，$K$ 個のクラスタに分割する文書クラスタリング問題を考える．文書 $d_n$ が $k$ 番目のクラスタに属するとき $y_{n,k}=1$, 属さないとき $y_{n,k}=0$ とする．また，$\mathbf{Y}=(y_{n,1},\cdots,y_{n,K},\cdots,y_{N,K})$ とする．

文書の表現として，BOW(Bag-of-Words) 表現を用いる．すなわち，文書 $d_n$ を，単語出現頻度ベクトル $\mathbf{x}_n=(x_{n,1},\cdots,x_{n,V})$ により表す．ここで，$V$ は予めた定めた語彙 $\mathcal{W}=\{w_1,\cdots,w_V\}$ の語彙数を，また，$x_{n,i}$ は，文書 $d_n$ における，単語 $w_i$ の出現頻度を表す．SpKmeans では，$\mathbf{x}_n$ を $L_2$ ノルムにより正規化する．以後，簡単のため，この正規化が施された単語出現頻度ベクトルは $\mathbf{x}_n$ で表し，単語出現頻度ベクトルと呼ぶこととする．

各クラスタに対して，代表ベクトル $\{\boldsymbol{\theta}_k\}_{k=1}^K$ を考え，$\boldsymbol{\Theta}=(\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_K)$ とする．そして，次の目的関数 $J(\mathbf{Y},\boldsymbol{\Theta})$ を最大にする $\{\mathbf{Y},\boldsymbol{\Theta}\}$ を求める．

$$\mathcal{J}(\mathbf{Y},\boldsymbol{\Theta})=\sum_{n=1}^{N}\sum_{k=1}^{K} y_{n,k}\; s(\mathbf{x}_n,\boldsymbol{\theta}_k). \tag{1}$$

ただし，$s$ は，単語出現頻度ベクトルと代表ベクトル間のコサイン類似度であり，$s(\mathbf{x}_n,\boldsymbol{\theta}_k)=\cos(\mathbf{x}_n,\boldsymbol{\theta}_k)$ とする．

$J$ の最大化問題の局所最適解 $\{\hat{\mathbf{Y}},\hat{\boldsymbol{\Theta}}\}$ を，$\mathbf{Y}$ と $\boldsymbol{\Theta}$ の逐次更新により求める．$\mathbf{Y}$ の更新では，各文書を，その文書との類似度が最も高いクラスタに分類する．また，$\boldsymbol{\theta}_k$ の更新では，$k$ 番目のクラスタに属する文書群の単語出現頻度ベクトルの平均ベクトルを，新たな $\boldsymbol{\theta}_k$ とする．

## 3. 提案アプローチ

高次元データにおいて，$K$-means の局所最適解が低品質になりやすい主な原因として，解 $\{\mathbf{Y},\boldsymbol{\Theta}\}$ の探索空間が，次元と共に大きくなり，それに伴って，局所最適解の数が増大する点がある．

そこで，この問題を克服するため，代表ベクトル集合 $\boldsymbol{\Theta}$ に対して制約を導入し，探索空間を小さくすることを考える．具体的には，代表ベクトルが取り得る値を，$M(>K)$ 個の候補ベクトル $\boldsymbol{\Phi}\in\mathbf{R}^{V\times M}$ に限定する．すなわち，

$$\boldsymbol{\theta}_k\in\boldsymbol{\Phi},\quad k=1,\cdots,K \tag{2}$$

とする．SpKmeans では，$\boldsymbol{\theta}_k$ は任意の $V$ 次元ベクトルを取り得たが，この制約の下では，$\boldsymbol{\theta}_k$ は $M$ 個のベクトルから選択される．従って，解の探索空間が限定され，低品質な局所最適解に陥りにくくなると考える．

このアプローチによって，精度の高いクラスタリングを実現するには，適切な $\boldsymbol{\Phi}$ を構成する必要がある．そこで，文書が持つ特性を利用した $\boldsymbol{\Phi}$ の構成法を提案する．具体的には，ある単語 $v$ が出現した文書群 $\mathcal{D}_v\subset\{d_n\}_{n=1}^N$ を考え，文書群 $\mathcal{D}_v$ に属する文書の単語出現頻度ベクトルの平均ベクトル

$$\mathbf{m}_v=\frac{1}{|\mathcal{D}_v|}\sum_{\{n|d_n\in\mathcal{D}_v\}}\mathbf{x}_n \tag{3}$$

を候補ベクトルとする．つまり，ある単語集合 $\mathcal{V}=\{v_1,\cdots,v_M\}$ を用いて，$\boldsymbol{\Phi}$ を次のように構成する．

$$\boldsymbol{\Phi}=\{\mathbf{m}_{v_1},\cdots,\mathbf{m}_{v_M}\} \tag{4}$$

$\mathbf{m}_v$ を候補ベクトルとして用いるのは，単語 $v$ が出現した文書群 $\mathcal{D}_v$ では，文書間の類似度が高い傾向が有ると考えるからである．まず，$\mathcal{D}_v$ に属する文書間の類似度が高いと考える理由を述べる．例として，$v$ が「野球」であるとする．単語間には依存性があり,「野球」が出現した文書では,「野球」に関連した単語 (例えば「バット」や「ストライク」) が出現しやすい．このような依存性により，共通の語が出現した文書間では，単語出現頻度ベクトル間の類似度が高いと考える．

次に，$\mathcal{D}_v$ に属する文書間の類似度が高ければ，$\mathbf{m}_v$ が，候補ベクトルとして望ましいと考える理由を述べる．クラスタリングの目的は，類似文書を同一のクラスタにまとめることである．従って，同一クラスタに属する文書間では，類似度が高い方が良い．逆にいえば，良い代表ベクトルであるための必要条件は，文書間類似度の高い文書群の平均ベクトルであることである．これより，$\mathbf{m}_v$ が候補ベクトルとして望ましいと考える．

## 4. 提案法の詳細

提案法の詳細を述べる．提案法では，まず，$M$ 個の単語群 $\mathcal{V}=\{v_1,\cdots,v_M\}$ を定める．本稿では，tf-idf の値が上位 $M$ 個の単語を $\mathcal{V}$ とした．そして，式 (4) より，候補ベクトル集合 $\boldsymbol{\Phi}$ を得る．これらの候補ベクトルの中から，$\boldsymbol{\theta}$ を選択するという制約の下で，式 (1) の目的

†日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

関数 $\mathcal{J}(\mathbf{\Theta}, \mathbf{Y})$ を最大にする，$\{\hat{\mathbf{Y}}, \hat{\mathbf{\Theta}}\}$ を求める．すなわち，次の最適化問題を解く．

$$\begin{aligned} &\underset{\mathbf{Y},\mathbf{\Theta}}{\text{maximize}}\ J(\mathbf{Y}, \mathbf{\Theta}) \\ &\text{subject to}\ \ y_{n,k} \in \{0,1\},\ n = 1, \cdots, N,\ k = 1, \cdots, K \\ &\sum_{k=1}^{K} y_{n,k} = 1,\ n = 1, \cdots, N, \quad \boldsymbol{\theta}_k \in \mathbf{\Phi},\ k = 1, \cdots, K \end{aligned} \tag{5}$$

式 (5) の離散最適化問題を解くには，greedy search アルゴリズムや，アニーリング法などの，一般の最適化手法が適用できる．本稿では，$\mathbf{\Theta}$ に関する greedy search を用いた．greedy search の手順を示す．

(i) $\mathbf{\Theta}$ を初期化する．

(ii) 目的関数の値を大きくするように $\mathbf{\Theta}$ を更新する．但し，$\mathbf{\Theta}$ 全体を一度に更新するのは困難なので，各代表ベクトル $\{\boldsymbol{\theta}_k\}_{k=1}^{K}$ を，１つずつ順番に更新する．$\boldsymbol{\theta}_k$ を更新する際には，

$$F(\mathbf{\Theta}) = \max_{\mathbf{Y}} \mathcal{J}(\mathbf{Y}, \mathbf{\Theta}) \tag{6}$$

として，次式により $\boldsymbol{\theta}_k$ を更新する．

$$\boldsymbol{\theta}_k \leftarrow \underset{\boldsymbol{\theta}_k \in \mathbf{\Phi}}{\operatorname{argmax}} \{F(\mathbf{\Theta})\} \tag{7}$$

このとき，$\boldsymbol{\theta}_k$ 以外の代表ベクトル $\boldsymbol{\theta}_{k'}, k' \neq k$ は固定し，$F(\mathbf{\Theta})$ を最大にする $\boldsymbol{\theta}_k \in \mathbf{\Phi}$ を求める．

(iii) (ii) の更新を，収束するまで繰り返す．

(iv) 得られた解 $\hat{\mathbf{\Theta}}$ を用いて，$\hat{\mathbf{Y}} = \operatorname{argmax}_{\mathbf{Y}} \mathcal{J}(\mathbf{Y}, \hat{\mathbf{\Theta}})$ を得る．そして，$\hat{\mathbf{Y}}$ に従って，文書群を $K$ 個のクラスタに分割する．

文書と候補ベクトル間の類似度を，予め全て計算しておけば，提案法の計算量は $O(NVM + NK^2M)$ である．

(ii) の更新では，$F(\mathbf{\Theta})$ の変化量を考慮すれば十分である．なぜなら，$\boldsymbol{\theta}_k^{(t)}$ を $\boldsymbol{\theta}_k^{(t+1)}$ に更新する際には，$F(\mathbf{\Theta})$ の変化量を最大にする $\boldsymbol{\theta}_k^{(t+1)}$ を選択すれば良いからである．$\boldsymbol{\theta}_k$ の更新によって，$\mathbf{\Theta}^{(t)}$ が $\mathbf{\Theta}^{(t+1)}$ に変化したとする．ただし，$\boldsymbol{\theta}_l^{(t)} = \boldsymbol{\theta}_l^{(t+1)}, l \neq k$ であることに注意．$F(\mathbf{\Theta})$ の変化量は次式のようになる．

$$\begin{aligned} &F(\mathbf{\Theta}^{(t+1)}) - F(\mathbf{\Theta}^{(t)}) \\ &= \sum_{n \in \mathcal{D}_1} \left[ \max_l \left\{ s(\mathbf{x}_n, \boldsymbol{\theta}_l^{(t+1)}) \right\} - s(\mathbf{x}_n, \boldsymbol{\theta}_k^{(t)}) \right] \\ &+ \sum_{n \in \mathcal{D}_2} \left[ s(\mathbf{x}_n, \boldsymbol{\theta}_k^{(t+1)}) - \max_l \left\{ s(\mathbf{x}_n, \boldsymbol{\theta}_l^{(t)}) \right\} \right] \end{aligned} \tag{8}$$

$\mathcal{D}_1$ は， 更新前に $k$ 番目のクラスタに分類された文書群であり，$\mathcal{D}_1 = \{n | y_{n,k}^{(t)} = 1\}$ である．ただし，$\mathbf{Y}^{(t)} = \operatorname{argmax}_{\mathbf{Y}} \mathcal{J}(\mathbf{Y}, \mathbf{\Theta}^{(t)})$ とする．また，$\mathcal{D}_2$ は，この更新によって， 新たに $k$ 番目のクラスタに分類される文書群であり，$\mathcal{D}_2 = \{n | n \notin \mathcal{D}_1,\ y_{n,k}^{(t+1)} = 1\}$ である．$|\mathcal{D}_1| + |\mathcal{D}_2| \leq N$ であるから，単純に $n$ に関する和を計算するよりも，計算量を削減できる．

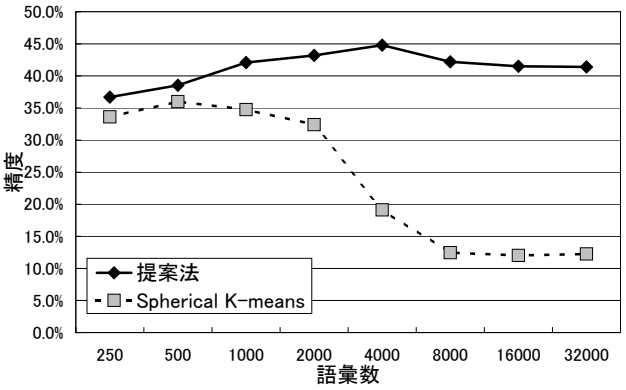


図 1: 語彙数 $V$ を変化させた場合の，提案法と SpKmeans の精度の比較

## 5. 評価実験

評価用データとして，文書クラスタリングの評価に広く用いられている 20newsgroups データ ‡を用いた．このデータでは，各文書は 20 のカテゴリのいずれかに分類されている．

この既存の分類と，クラスタリング結果との間の一致度により，クラスタリングの性能を評価する．評価尺度として，精度 (Micro Averaged Precision [3]) を用いた．精度は，既存の分類結果と，クラスタリング結果とが完全に一致した時のみ 100 % となる．

1000 個の文書をクラスタリングした際の，提案法と SpKmeans の精度を比較した．このとき，データの次元の影響を調べるため，用いる語彙の数 $V$ の値を変化させた場合のクラスタリング精度の変化を調べた．語彙を選択する規準として，相互情報量基準 [3] を用いた．なお，提案法では $M = 250$ とした．

図 1 にクラスタリング精度を比較した結果を示す．図 1 より，提案法が SpKmeans を上回るクラスタリング精度を達成し，その差はデータが高次元の場合顕著であることが分かる．また，再現率による評価でも同様に結果が得られた．なお，Xeon 3.2GHz，メモリ 2GB の計算機において，$V = 32000$ のとき，SpKmeans の計算時間は 0.28[s]，提案法の計算時間は 1.33[s] であった．

## 6. まとめ

文書の持つ特性を利用した文書クラスタリング手法を提案した．今後は，提案法の様々なテキスト・マイニング問題への応用を進めていきたい．

## 参考文献

[1] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1):143–175, 2001.

[2] Tao Liu, Shengping Liu, Zheng Chen, and Wei-Ying Ma. An evaluation on feature selection for text clustering. In *Proc. ICML '03*, 2003.

[3] N. Slonim, N. Friedman, and N. Tishby. Unsupervised document classification using sequential information maximization. In *Proc. SIGIR '02*, 2002.

‡入手先: `http://www.ai.mit.edu/~jrennie/20Newsgroups/`