文書クラスタリングを対象とした Weighted Kernel K-means の初期値設定法

茂木哲矢　新納浩幸　佐々木稔
茨城大学工学部情報工学科

1　はじめに

本論文ではクラスタリング手法である Weighted Kernel K-means（以下 WKK と略す）の初期値依存の問題を指摘し，その対策として K-means の初期値依存の問題に効果的な KKZ を WKK に適用することを試みる．

文書クラスタリングは，文書の集合に対して，知的な処理を行う基本的な処理であり，その重要性は明らかである．例えばテキストマイニングの分野では，クラスタリングは基本的な構成要素であるし [1]、情報検索の分野では，検索された文書の集合を俯瞰的に見るためにクラスタリングするシステムが盛んに研究されている（[2] など）．

クラスタリングの標準的手法として K-means が存在する．K-means はクラスタの中心と各データの距離を基準としてクラスタリングを行う手法である．K-means は山登り法で解を求めるため，局所最適解しか求めることができない．これは初期値によって求まる解が異なることを意味する．安定した解を得るためには，ランダムな初期値を用いて，複数回 K-means を実行する必要がある．また，K-means は非線形な分離境界をもつデータに対して最適な分類を行えないという問題もある．後者の問題の対策として，データ空間を非線形関数によって高次元空間に写像し，表現力を高めてから線形分離する Kenrel K-means が提案されている．

Kernel K-means 以外にも，非線形の分離境界をもつデータに対するクラスタリング手法にスペクトラルクラスタリングがある．スペクトラルクラスタリングはデータから得られる隣接行列に関する固有値を求めることでクラスタリング結果を得る手法であり，精度の高いクラスタリング手法として知られている．一方，Kernel K-means を一般化した WKK の評価関数はスペクトラルクラスタリングの評価関数と等価であることが知られている [3]．スペクトラルクラスタリングは固有値を求める処理の負荷が大きいが，WKK は固有値を求めずに，繰り返しの処理によって，その評価関数の最適解を見つけようとする．そのため WKK により効率的かつ高精度のクラスタリングが期待できる．

しかし WKK も山登り法であるため，K-means 同様，局所最適解しか得られないという問題が存在する．ここでは K-means の初期値依存の問題に効果的な KKZ [4] を WKK に適用することを試みる．これにより初期値を変更して複数回のクラスタリングを行うことを避けることができ，クラスタリングの全体の処理が効率化される．また KKZ を用いることで，クラスタリングの精度は劣化する恐れがあるが，実験では通常の手法と同等の以上の精度が得られた．

2　Weighted Kernel K-means

非線形な分離境界をもつデータに対するクラスタリング手法に Kernel K-means とスペクトラルクラスタリングがある．Kernel K-means はデータを非線形関数を用いて，高次元空間に写像してから線形分離する方法である．スペクトラルクラスタリングはデータから得られる隣接行列の固有値を求めることでクラスタリング結果を得る手法である．スペクトラルクラスタリングで使われる基本的な評価関数として Normalized cut がある [5]．

このように非線形な分類境界をもつデータに対するクラスタリング手法のアルゴリズムは大きく異なる．しかし，Kernel K-means を一般化した WKK の評価関数と Normalized cut の評価関数が等価になることが知られている．スペクトラルクラスタリングは精度の高いクラスタリングとして知られているため，文書クラスタリングに WKK を用いることで精度の高い結果が得られることが期待できる．

本論文で使用するクラスタリングである WKK のアルゴリズムを図 1 に示す．

WKK のアルゴリズムは K-means と大部分を共有している．つまり，K-means と同様に解を山登り法で求めており，解が初期値に依存する問題がある．

1. クラスタの個数 K 及びクラスタ C_1, C_2, $\cdots$, C_K を設定する．
2. データ x とクラスタ C_i との距離 $||x - m_i||$ を計算し，以下の式で x のクラスタを設定する．

$$\boldsymbol{m}_j = \frac{\sum_{\boldsymbol{b}\in\ _j} w(\boldsymbol{b})\phi(\boldsymbol{b})}{\sum_{\boldsymbol{b}\in\ _j} w(\boldsymbol{b})} \quad (1)$$

$$C_j = \{\boldsymbol{a} : j^*(\boldsymbol{a}) = j\} \quad (2)$$

3.（a）step 2 により各データの所属するクラスタが変化しなければ終了．
（b）変化があったときには step 2 に戻る．

図 1　WKK のアルゴリズム

3　KKZ

K-means は初期値によって求まるクラスタリング結果が異なる．このため，通常，ランダムな初期値によりクラスタリングを複数回行い，そこから評価関数を用いて最適な解を選択する．しかしこれは複数回のクラスタリング処理が必要なため，処理の負荷が高い．

KKZ は K-means の初期値を設定する手法である．KKZ ではデータ間の距離を基準にして初期値を決定する．新しい代表点は既存の代表点から最も遠くなるように選択される．KKZ による初期値選択のアルゴリズムを図 2 に示す．従来安定した解を得るには，複数回クラスタリングを行う必要があったが，KKZ を用いることによって 1 回のクラスタリングで済むようになる．

1. 各データのノルムを計算し，最大となるデータを 1 番目のクラスタの代表点 c_1 とする．
2. 各データと既に決まっている代表点 c_1, c_2, $\cdots$, c_k との距離 d_{ij} を計算する
3. 各データの最小の d_{ij} を選択し，d_i とする．
4. d_i の中で最大となるデータを k 番目のクラスタの代表点 c_k に設定する．
5.（a）設定したクラスタ数だけ初期値が作成されたならば終了する．
（b）設定したクラスタ数だけ初期値が作成されていなければ step 2 に戻る．

図 2　KKZ による初期値設定手順

4　実験

本手法の有効性を確認するために，クラスタリングツール CLUTO [*1] に付属するデータセットのうち fbis，tr11，tr12，tr41，wap の 5 つを用いて，クラスタリングの実験を行う．用いる手法は以下の 3 つである．

km　データセットからランダムに初期値を選択し，K-means を用いたクラスタリングを行い得た 10 個の解の平均．

wkkm　データセットからランダムに初期値を選択し，WKK を用いてクラスタリングを行い得た 10 個の解の平均．

wkkm+kkz　データセットから KKZ を用いて初期値を選択し，WKK を用いてクラスタリング

[*1] http://glaros.dtc.umn.edu/gkhome/views/cluto/

する．

クラスタリングの評価にはエントロピーと純度を用いる [6]．エントロピーとは式 3，4 で定義され，値が低いほど結果が良好であることを意味する．純度とは式 5，6 で定義される正解のクラスタのデータをどの程度含むかという指標のことで，値が高いほど結果が良好であることを意味する．

$$\sum_{i=1}^{K} \frac{|C_i|}{N} E_i = \sum_{i=1}^{K} \frac{\sum_{j=1}^{K} x_{ij}}{N} E_i \tag{3}$$

$$E_i = -\sum_{h=1}^{K} P(A_h|C_i) \log P(A_h|C_i) \tag{4}$$

$$P_i = \frac{1}{|C_i|} \max_h |C_i \cap A_h| \tag{5}$$

$$\sum_{i=1}^{K} \frac{|C_i|}{N} P_i = \frac{1}{N} \sum_{i=1}^{K} \max_h |C_i \cap A_h| \tag{6}$$

ここで C_i は得られたクラスタ，A_i は正解クラスタ，x_{ij} は C_i と A_j に共通に属するデータの個数，K はクラスタ数，N はデータ数を表す．

また，クラスタリングが評価関数を最小にしていることを確認するために式 7 の K-means の評価関数を用いる．

$$\sum_{i=1}^{K} \sum_{\boldsymbol{x} \in C_i} ||\boldsymbol{x} - \boldsymbol{c}_i||^2 \tag{7}$$

実験の結果を表 1，2，3 及び図 3，4 に示す．

表 1　エントロピーの比較

dataset	km	wkkm	wkkm+kkz
fbis	0.356	0.364	0.366
tr11	0.307	0.272	0.215
tr12	0.425	0.348	0.299
tr41	0.283	0.297	0.203
wap	0.399	0.364	0.363

表 1，2，3 より，手法 km，wkkm を比較すると，wkkm はデータセット fbis，tr41 のエントロピーが 0.008，0.014 増加し，純度が 0.001，0.024 減少して性能が低下している．しかし，評価関数値は 117，

表 2　純度の比較

dataset	km	wkkm	wkkm+kkz
fbis	0.659	0.658	0.678
tr11	0.750	0.790	0.838
tr12	0.668	0.727	0.770
tr41	0.749	0.725	0.852
wap	0.611	0.636	0.656

表 3　評価関数値の比較

dataset	km	wkkm	wkkm+kkz
fbis	1968	1851	1515
tr11	360	317	309
tr12	274	245	246
tr41	776	733	637
wap	1432	1343	1191

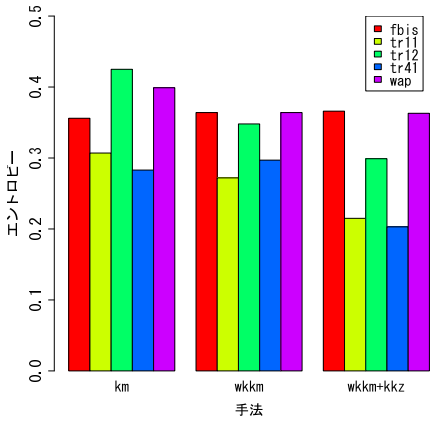

図 3　エントロピーの比較

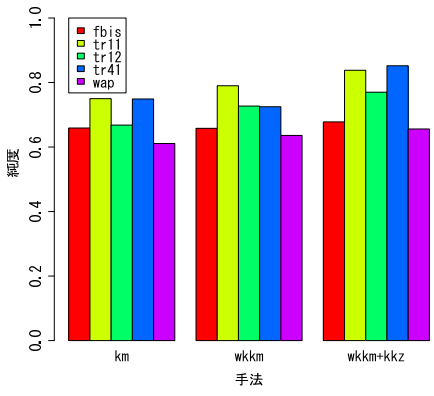

図 4　純度の比較

43 減少している．このため，クラスタリング結果は悪化したが，評価関数を最小化するという目的を達成しているといえる．手法 wkkm によって，性能が低下するデータセットはあるものの，その変化は少量であり，評価関数値が減少しているため，クラスタリングの性能は向上したことがいえる．

手法 wkkm ，wkkm+kkz を比較すると，wkkm+kkz はデータセット fbis のエントロピーが 0.002 増加し性能が低下している．ほかのデータセットにおいては，性能が向上する結果が得られた．また，データセット tr12 の評価関数値が 1 増加しているが，ほかのデータセットでは評価関数値が減少している．手法 wkkm+kkz によって，全体的にクラスタリングの性能が向上したといえる．

5 考察

手法 wkkm において求めた解の中で評価関数値が最小とする解の性能を表 4 に示す．

表 4 評価関数が最小となるときの性能

データセット	エントロピー	純度	評価関数値
bs	0.357	0.671	1754
tr11	0.246	0.826	259
tr12	0.337	0.735	198
tr41	0.286	0.708	688
wap	0.369	0.641	1217

表 1，2，3，4 より，従来の初期値設定法を使用したクラスタリングで得られた 10 個の解の中で評価関数値を最小とすると結果の性能と比較すると，データセット fbis のエントロピーが 0.009 増加し性能が低下している．ほかのデータセットでは提案手法の方が良いクラスタリング結果を得ることが出来た．

実験の結果，WKK の初期値設定法に KKZ を用いることによって，クラスタリングの性能は従来と同等，もしくは向上することが確認できた．従来の初期値設定法では初期値に依存するため複数回クラスタリングを実行する必要があったが，本論文で提案する手法では 1 回で済む．

また，fbis のエントロピーの値が K-means, WKK，提案手法と増大し性能が悪化しているが，これは WKK で使用しているカーネルが fbis に適していないことによると考えられる．

6 おわりに

本論文では WKK の初期値設定法に KKZ を使用することを提案した．従来の初期値設定法を用いた文書クラスタリング結果と初期値設定法に KKZ を使用した結果を比較すると，複数のデータセットで性能が同等，もしくは向上する結果が得られ，提案手法が有効であることが確認できた．

一方，一部のデータセットでは性能が低下することを確認した．これは WKK で使用するカーネルがデータセットに適していないことが原因であると考えられる．カーネルを適切に設定することによって，クラスタリングの性能はさらに向上すると考えられる．今後これらの点を改良した文書クラスタリングを作成する．

参考文献

[1] Michael W. Berry, editor. *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer, 2003.

[2] Hua-Jun Zeng and i-Cai He and Zheng Chen and Wei-Ying Ma and Jinwen Ma. Learning to cluster web search results. In *Proc. of the 27th annual international conference on Re- search and development in information retrieval*, pp. 210–217, 2004.

[3] I.S.Dhillon and Y.Guan and B. Kulis. Kernel k-means, Spectral Clustering and Normalized Cuts. In *KDD*, pp. 551–556. ACM Press, 2004.

[4] Ji He and Man Lan and Chew-lim Tan and Sam-yuan Sung and Hwee-boon Low. Initialization of cluster refinement algorithms: A review and comparative study. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pp. 297–302, 2004.

[5] Jianbo Shi and Jitendra Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 888–905, 2000.

[6] 新納 浩幸. R で学ぶクラスタ解析. オーム社, 2007.