D-004

多観点類似度を用いた凝集型階層クラスタリング
MVS Based Hierarchical Clustering

藤原勇二 † Yuji Fujiwara　**古賀久志** † Hisashi Koga　**戸田貴久** † Takahisa Toda

1. はじめに

与えられたデータ集合を教師データによる事前学習なしに分類する手法をクラスタリングと呼ぶ．一般的に文書等の高次で疎なデータに対するクラスタリングでは，類似度指標として cosine 類似度用いられる [1].

cosine 類似度を基とし，評価精度をより高めるために，Nguyen らは cosine 類似度における原点を複数用いた多観点類似度 (Multiviewpoint-Based Similarity:MVS) を提案した．そして，MVS を非階層クラスタリングに適用することで，文書データのクラスタリングにおいて優れた結果を示した [2]．ただし，非階層クラスタリングは事前にクラスタの数の指定が必要である．

一方で階層クラスタリングではその必要がなく階層的な分類構造を抽出できる．そこで我々は，MVS を適用した階層クラスタリング手法を開発する．特に MVS は cosine 類似度より計算量が大きく，クラスタリング全体の計算量を悪化させる恐れがあるが，提案手法ではクラスタ間の類似度を高速に更新する式を用いて，一般的な階層クラスタリングと同様に $O(n^2 \log n)$ の計算量を実現する．また，実験により提案手法が計算量及び分類精度の観点から文書データのクラスタリングに有用であることを示す．

本論文における構成を以下に示す．2 節では本論文が取り扱う問題の定義を行う．3 節では提案手法の基盤研究である MVS について紹介する．4 節では提案手法の説明のための前準備として凝集型階層クラスタリングについて紹介する．5 節では MVS を階層クラスタリングに適用する提案手法について論じる．6 節では実データを用いた数値実験により提案手法の優位性を示す．7 節では関連研究を紹介する．8 節では結論及び今後の展望について論じる．

2. 問題定義

本論文では，高次元単位球面上に分布するベクトル集合をクラスタ分類する問題を取り扱う．この問題の実例として，文書データに対するクラスタリングが挙げられる．文書データは，文書中における各単語の出現頻度ヒストグラムの正規化によって得られるベクトルで表現される．

3. 多観点類似度

本節では，提案手法の基盤技術となる MVS について説明する．高次元単位球面上のベクトル間の類似度には，一般的に cosine 類似度が用いられる．単位ベクトルであるデータ d_i, d_j 間の cosine 類似度 $\mathrm{CS}(d_i, d_j)$ は，式 (1) のようにベクトル内積として定義される．

$$\mathrm{CS}(d_i, d_j) = d_i^{\mathrm{T}} d_j \tag{1}$$

この式は，式 (2) のように，各データと原点 0 の差の内積としても表せる．これは cosine 類似度が原点 0 のみを唯一の基準点として類似性を評価していることを表している．

$$\mathrm{CS}(d_i, d_j) = (d_i - 0)^{\mathrm{T}} (d_j - 0) \tag{2}$$

MVS の狙いは，この基準点をデータセット中の複数の様々な点に移動させることで，cosine 類似度よりもデータ分布に適応した類似性評価を実現することである．MVS は式 (3) のように，同じクラスタ内の 2 点 d_i, d_j の類似度を n 個の全データの集合 S から d_i, d_j が所属するクラスタ r の集合 S_r を除いたものを基準点 d_h として，各 d_h とのベクトル差の内積の平均として定義される．

$$\begin{aligned}
&\mathrm{MVS}(d_i, d_j \mid d_i, d_j \in S_r) \\
&\quad = \frac{1}{n - n_r} \sum_{d_h \in S \setminus S_r} (d_i - d_h)^{\mathrm{T}} (d_i - d_h) \qquad (3) \\
&\quad = \frac{1}{n - n_r} \sum_{d_h \in S \setminus S_r} \Big\{ \mathrm{CS}(d_i - d_h, d_j - d_h) \\
&\qquad\qquad \| d_i - d_h \| \| d_j - d_h \| \Big\} \qquad (4)
\end{aligned}$$

MVS は式 (4) のように，原点の異なる cosine 類似度の重み付き平均としても表すことが出来る．ここで重み $\| d_i - d_h \| \| d_j - d_h \|$ は，d_i, d_j と d_h との距離の積である．この非類似度による重み付けは，d_i, d_j が同一クラスタに含まれることの妥当性を評価しており，クラスタリングにおける分類能力の向上に寄与すると考えられる．

式 (3) からわかるように，MVS は最大 $n - 2$ 回の内積の算術平均である．そのため，MVS の計算量は $O(n)$ であり，これは単純な cosine 類似度の約 n 倍の計算量となる．

4. 凝集型階層クラスタリング

本節では，提案手法を論じる上での事前知識として凝集型階層クラスタリングに関して説明する．凝集型階層クラスタリングは事前にクラスタ数を決めず，全てのデータが異なるクラスタに属する状態から，同一クラスタに属する状態までの分類結果を階層的に得る手法である．以後，凝集型階層クラスタリングを単に階層クラスタリングと呼称する．また，ここではデータ及びクラスタ間の類似度に cosine 類似度を用いるものとする．

†電気通信大学大学院情報理工学研究科, 調布市
Graduate School of Informatics and Engineering,
University of Electro-Communications,
1-5-1 Chofugaoka, Chofu-shi, 182-8585 Japan

アルゴリズム 1 階層クラスタリング

Input: データセット $S = \{d_1, \ldots, d_n\}$

1: $n \times n$ の類似度行列の初期化
2: **while** クラスタ数 > 1 **do**
3: 最も類似度の高いクラスタ a, b を探す
4: a, b をマージしてクラスタ c を作る
5: **for** $k \leftarrow c$ 以外のクラスタ **do**
6: クラスタ c と k の間の類似度を更新

Output: 階層的なクラスタ構造

4.1. 階層クラスタリングの計算量

階層クラスタリングの概略をアルゴリズム 1 に示す．1 行目では全てのデータ対の間の類似度を持つ $n \times n$ の類似度行列を作成する．cosine 類似度は $O(1)$ で計算可能で，類似度行列の計算量は $O(n^2)$ である．また，3 行目の最近傍のクラスタ対の探索は，ヒープを用いて $O(n \log n)$ で計算可能である．6 行目の類似度の更新は，事前に計算された値の重み付き和により $O(1)$ で計算できる．この計算が，新しいクラスタと他の全クラスタとの間で必要であるため，類似度行列の更新の計算量は $O(n)$ である．出力のクラスタ構造は，各ステップでマージされたクラスタ対とその類似度を持つ．

クラスタリング全体で，クラスタのマージは $n-1$ 回行われる．マージの発生毎に必要な操作で最も計算量が大きいものは，$O(n \log n)$ で最近傍のクラスタ対を探す部分である．そのため，一般的な階層クラスタリングは $O(n^2 \log n)$ の計算量で実行可能である．

4.2. 群平均法

階層クラスタリングは用いるクラスタ間類似度の定義によってその結果が異なる．群平均法はクラスタ間類似度を，互いのクラスタメンバ間の類似度の平均とする手法である．群平均法は，他の手法と比べて外れ値に強く，実データへの利用に適しているためよく用いられる．クラスタ a, b をマージして新しくできたクラスタ c とその他のクラスタ k との類似度 Sim_{kc} は，式 (5) のように表される．

$$Sim_{kc} = \frac{1}{n_k n_c} \sum_{d_i \in S_k} \sum_{d_j \in S_c} Sim(d_i, d_j) \qquad (5)$$

また，式 (6) のように事前に計算されているマージ前のクラスタ間類似度を用いることで，クラスタサイズに基づく重み付き平均で表現することができ，類似度の計算を $O(1)$ で行うことが可能である．

$$Sim_{kc} = \frac{n_a}{n_c} Sim_{ka} + \frac{n_b}{n_c} Sim_{kb} \qquad (6)$$

5.MVS を用いた階層クラスタリング

本節では，MVS に基づく階層クラスタリングを提案する．MVS を階層クラスタリングに適用する場合，類似度計算を含む行程を変更する必要がある．具体的には

- 事前の類似度行列の初期化
- マージに伴う類似度行列の更新

である．3 節で論じたように，MVS の計算量は $O(n)$ である．そのため，$O(1)$ で計算可能な cosine 類似度を用いた場合に比べて，計算量が非常に大きくなることが懸念される．そこで，式 (6) のように事前計算されたマージ前のクラスタとの間の類似度を用いて $O(1)$ で計算可能な更新式を導出する．これにより，全体の計算量の増加させることなく MVS に基づく高い分類精度によるクラスタリングを可能にする．

5.1.MVS を用いた場合の群平均法の更新式

定理 1. MVS を適用した群平均法による更新を行う階層クラスタリングにおいて，クラスタ a, b をマージして新しくクラスタ c を作った時，クラスタ c とその他のクラスタ k との類似度 Sim_{kc} は，式 (7) によって，$O(1)$ で計算可能である．

$$\begin{aligned} Sim_{kc} = & \frac{1}{(n_a + n_b)(n - n_k - n_a - n_b)} \\ & \Big\{ n_a (n - n_k - n_a) Sim_{ka} \\ & + n_b (n - n_k - n_b) Sim_{kb} \\ & + 2 \left(D_a^{\mathrm{T}} D_b - n_a n_b \right) \Big\} \end{aligned} \qquad (7)$$

ここで，式中の D_a, D_b はクラスタ a, b の各ベクトル和，n_a, n_b は a, b の各データ数を示す．

(証明). 式 (5) における $Sim(d_i, d_j)$ に，式 (3) に示す MVS を用いると，

$$\begin{aligned} Sim_{kc} = & \frac{1}{n_k n_c (n - n_k - n_c)} \\ & \sum_{d_k \in S_k} \sum_{d_c \in S_c} \sum_{d_h \in S \setminus S_k \setminus S_c} (d_k - d_h)^{\mathrm{T}} (d_c - d_h) \end{aligned}$$

である．Sim_{kc} について，MVS は類似度を評価するデータ対が，同一クラスタに所属していなければならない制約があるため，Sim_{kc} を計算する上で，k, c を同じクラスタとして扱っている．また，Sim_{ka}, Sim_{kb} も同様に，

$$\begin{aligned} Sim_{ka} = & \frac{1}{n_k n_a (n - n_k - n_a)} \\ & \sum_{d_k \in S_k} \sum_{d_a \in S_a} \sum_{d_h \in S \setminus S_k \setminus S_a} (d_k - d_h)^{\mathrm{T}} (d_a - d_h) \\ Sim_{kb} = & \frac{1}{n_k n_b (n - n_k - n_b)} \\ & \sum_{d_k \in S_k} \sum_{d_b \in S_b} \sum_{d_h \in S \setminus S_k \setminus S_b} (d_k - d_h)^{\mathrm{T}} (d_b - d_h) \end{aligned}$$

ここで，クラスタ c がクラスタ a, b のマージによるものであることから，

$$Sim_{kc} = \frac{1}{n_k(n_a+n_b)(n-n_k-n_a-n_b)} \sum_{d_k \in S_k} \left\{ \sum_{d_a \in S_a} \sum_{d_h \in S \setminus S_k \setminus S_a \setminus S_b} (d_k - d_h)^{\mathrm{T}}(d_a - d_h) + \sum_{d_b \in S_b} \sum_{d_h \in S \setminus S_k \setminus S_a \setminus S_b} (d_k - d_h)^{\mathrm{T}}(d_b - d_h) \right\}$$

のように書き換えることができる．この式中において Sim_{ka}, Sim_{kb} との共通部分をまとめると，

$$\begin{aligned} Sim_{kc} &= \frac{1}{(n_a+n_b)(n-n_k-n_a-n_b)} \\ &\quad \Big\{ n_a(n-n_k-n_a)Sim_{ka} \\ &\quad + n_b(n-n_k-n_a)Sim_{kb} \\ &\quad + 2\left(D_a^{\mathrm{T}} D_b - n_a n_b\right) \Big\} \end{aligned}$$

のように，Sim_{ka}, Sim_{kb} を用いた式に表すことができる．よって，$Sim_{ka}, Sim_{kb}, D_a, D_b, n_a, n_b$ が事前に求められているならば，Sim_{kc} は $O(1)$ で計算できる． □

5.2. 計算量の解析的評価

定理 2. MVS に基づく群平均法の階層クラスタリングに必要な計算量は $O(n^2 \log n)$ である．

(証明). MVS を階層クラスタリングに適用する場合，類似度行列の初期化とマージに伴う類似度行列の更新の部分のみを変更する．そのため，これらの手続きが変更前の計算量と同等であれば，階層クラスタリング全体の計算量を $O(n^2 \log n)$ より大きくすることはない．

まず，類似度行列の初期化が $O(n^2)$ で計算できることを示す．初期状態では，全てのデータがそれぞれ別のクラスタに所属ことから各クラスタ間の類似度は，

$$Sim_{ij} = \frac{1}{n-2}(n d_i^{\mathrm{T}} d_j - d_i^{\mathrm{T}} D - d_j^{\mathrm{T}} D + n)$$

により $O(1)$ で計算可能である．このことから，アルゴリズム 2 の手続きを行うことで，事前の類似度行列の初期化は $O(n^2)$ である．

アルゴリズム 2 MVS による類似度行列の初期化

Input: データセット $S = \{d_1, \ldots, d_n\}$

1: $D \leftarrow \sum_{i=1}^{n} d_i$
2: **for** $i \leftarrow 1, \ldots, n$ **do**
3: **for** $j \leftarrow 1, \ldots, n$ **do**
4: $Sim_{i,j} \leftarrow \frac{1}{n-2}(n d_i^{\mathrm{T}} d_j - d_i^{\mathrm{T}} D - d_j^{\mathrm{T}} D + n)$

Output: $n \times n$ 類似度行列 Sim

次に，マージに伴う類似度行列の更新の計算量が $O(n)$ であることを示す．5.1 節で示したように，最近傍のクラスタ対をマージしたクラスタとその他のクラスタとの類似度の更新は式 7 を用いて $O(1)$ で計算できる．また，式中にクラスタのベクトル和 D_a, D_b と要素数 n_a, n_b が現れるが，a, b をマージした新しいクラスタ c におけるこれらは，以下のように $O(1)$ で計算が可能である．

$$\begin{aligned} D_c &= D_a + D_b \\ n_c &= n_a + n_b \end{aligned}$$ □

6. 評価実験

本節では，提案手法が，階層クラスタリングに MVS を適用することで，cosine 類似度を用いた既存手法に対して優位性があることを示すために，実データに対して以下の 2 種類の実験を行った結果を示す．

- クラスタリング結果の分類精度の評価実験
- クラスタリングの実行時間の評価実験

6.1. データセット

実験には，表 1 に示す 12 種類のデータセットを用いる．これらのデータセットは，テキストデータマイニングツール CLUTO 上で提供されている [3]．これらは，[2] を含める多数の先行研究の実験評価において用いられている．

表中において，c はデータセットのもつクラス数，n はデータ数，m は単語数である．データセットはストップワード除去とステミングを含む前処理をした後，TF-IDF によるベクトルで取り扱う．TF-IDF は各単語の文書中での頻出度 (Term Frequency) とデータセット中における希少度を表す逆文書頻度 (Inverse Document Frequency) の積で特徴を表現する．前処理の詳細については [2] を参照されたい．

表 1: 実験で使用する文書データセット

データセット	情報源	c	n	m
fbis	TREC	17	2,463	2,000
hitech	TREC	6	2,301	13,170
k1a	WebACE	20	2,340	13,859
la1	TREC	6	3,204	17,273
la2	TREC	6	3,075	15,211
re0	Reuters	13	1,504	2,886
tr31	TREC	7	927	10,127
wap	WebACE	20	1,560	8,440
tr11	TREC	9	414	6,424
tr12	TREC	8	313	5,799
tr23	TREC	6	204	5,831
tr45	TREC	10	690	8,260

6.2. 分類精度評価

階層クラスタリングによる分類結果を，データセットのクラス数に等しいクラスタ数で分割した場合の所属クラスタのラベルと真のクラスのラベルの一致度から分類精度を評価する．クラスタリング結果と真値の一致度の指標には正規化相互情報量 (Normalized Mutual

Information:NMI) を用いる．真値とクラスタリング結果の間の NMI は式 (8) で評価される．

$$\mathrm{NMI} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{i,j}\log\frac{n\cdot n_{i,j}}{n_i n_j}}{\sqrt{(\sum_{i=1}^{k} n_i \log\frac{n_i}{n})(\sum_{j=1}^{k} n_j \log\frac{n_j}{n})}} \quad (8)$$

ここで，n_i は真値のクラス i のデータ数，n_j はクラスタリング結果のクラスタ j のデータ数，$n_{i,j}$ はクラス i とクラスタ j の内で共通するデータ数を示す．NMI は 0 から 1 の値を取り，大きいほど互いの説明能力が高いことを表す．

表 2: 分類精度及び処理時間の評価

データ	分類精度		処理時間 (sec)		
	MVS	CS	MVS	CS	rate
fbis	0.584	0.561	52.18	47.58	1.10
hitech	0.255	0.059	70.06	65.40	1.07
k1a	0.556	0.550	74.63	70.85	1.05
la1	0.376	0.316	171.36	163.76	1.05
la2	0.466	0.390	147.72	139.59	1.06
re0	0.312	0.296	14.39	12.75	1.13
tr31	0.670	0.527	6.98	6.29	1.11
wap	0.554	0.539	21.00	19.07	1.10
tr11	0.645	0.633	1.06	0.93	1.14
tr12	0.553	0.523	0.59	0.50	1.17
tr23	0.260	0.223	0.25	0.22	1.16
tr45	0.555	0.495	3.46	3.02	1.14

表 2 左に，クラスタリング結果の NMI による評価を示す．これらのデータセットにおいては，cosine 類似度よりも MVS のほうが一致度が高い結果となった．

6.3. 計算時間評価

ここでは，MVS を適用した階層型クラスタリングが，実データのクラスタリングに用いた場合にも従来手法と同程度の処理時間で実行できることを示す．

表 2 右に，処理時間の比較実験の結果を示す．この実験においては，最近傍のクラスタ対を探す行程を $O(n^2)$ で実装している．また，実験結果の数値は各条件ごとに 3 回試行した結果の平均である．表中の rate は，MVS の実行時間と cosine 類似度の実行時間の比を表す．この結果から，実際に MVS を用いた階層クラスタリングは，cosine 類似度を用いた従来手法と同等の時間でクラスタリングを実行可能であることがわかる．

7. 関連研究

本節では，MVS についての関連研究を紹介する．

クラスタリング問題において，データの要素 V が複数の独立な部分集合 $V^{(1)}, V^{(2)}$ からなると仮定して学習する Multi-View と呼ばれる概念に基づいた手法がある [4]．ここでは，扱われるデータが，異なる視点から得られる特徴で表現されていることを Multi-View と称している．この Multi-View は，同一のデータを複数の基準点から評価する本論文における多観点とは異なるものである．

本論文における提案手法は，MVS を凝集型階層クラスタリングに適用したものである．一方で，同じ MVS を分割型階層クラスタリングに適用する手法も存在する [5]．この手法では，分割型階層クラスタリングにおいて，クラスタを 2 分割して新たなクラスタを作る操作に，[2] に提案された分割最適化手法を用いている．

8. おわりに

本論文では，高次元で疎な単位ベクトルによるデータセットに対する階層クラスタリングの分類精度向上のため，階層クラスタリングに，Nguyen らの提案する MVS を適用した．また，MVS の導入によって懸念される計算量の増加を解消するため，マージ後のクラスタ間類似度の更新を $O(1)$ で行う式を導出した．これにより，MVS を適用した場合においても一般的な階層クラスタリングと同様に計算量 $O(n^2 \log n)$ を実現した．

さらに，実データを用いた実験により，MVS を用いた階層クラスタリングが，既存手法と同程度の計算時間で，より高い分類精度を持つことを確認した．

MVS では基準点をデータ対の所属クラスタを除いた全てのデータとしている．しかし，データ対から極端に遠い基準点との差の内積は，類似度測定全体に悪影響を及ぼす場合も考えられる．そこで，一定の条件で悪影響を与えうる基準点を除外する MVS を構築することを検討したい．

また，本論文では群平均法のみを取り扱った．本論文中で取り扱わなかった階層クラスタリングの手法についても MVS の導入を検討したい．

謝辞

本研究は科研費基盤研究 (C)15K00148 の助成を受けたものである．

参考文献

[1] I.S. Dhillon and D.S. Modha, “Concept decompositions for large sparse text data using clustering,” Machine Learning, pp.143–175, 2000.

[2] D.T. Nguyen, L. Chen, and C.K. Chan, “Clustering with multiviewpoint-based similarity measure,” IEEE transactions on knowledge and data engineering, vol.24, no.6, pp.988–1001, 2012.

[3] G. Karypis, “Cluto-a clustering toolkit,” Technical report, MINNESOTA UNIV MINNEAPOLIS DEPT OF COMPUTER SCIENCE, 2002.

[4] S. Bickel and T. Scheffer, “Multi-view clustering.,” Proc. ICDM 2004, pp.19–26, 2004.

[5] S. Jayaprada, A. Aswani, and G. Gayathri, “Hierarchical divisive clustering with multi view-point based similarity measure,” Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013, pp.483–491, 2014.