5W-02

コサイン類似度に基づく分割型と凝集型ハイブリッド文書クラスタリング法

方　越洋 †　　斉藤　和巳 †
† 神奈川大学大学院 理学研究科

1　はじめに

本研究では, 大規模文書データのクラスタリングを課題とし, k-means 法に代表され分割型 (divisive) で文書クラスタを求め, 凝集型階層的 (agglomerative hierarchical) クラスタリングによりデンドログラムを構成する手法を提案する. なお, k-means 法の初期値設定に階層的クラスタリングを用いる Scatter/Gather 法 [1] を除けば, 著者らの知る限り, これら型のハイブリッドに関する研究は殆ど見受けられない.

本稿では, 球面 (spherical) k-means 法 [2] の結果に凝集型階層的クラスタリングを適用するため, この枠組みでの群平均法とウォード法を導出するとともに, 新たにコサイン法を提案する. 約 30 万文書からなる NYTimes 文書データ [3] を用いた評価実験では，これら 3 手法によるデンドログラムを定性評価するとともに, コサイン法は他と比較して, 併合非類似度が小さく, 平衡なデンドログラムを構成することを定量評価する.

2　提案手法

総数 N の文書集合を $\mathcal{N} = \{1, \cdots, n, \cdots, N\}$ とし, 語彙数 D の出現単語集合を $\mathcal{D} = \{1, \cdots, d, \cdots, D\}$ とする. また, 各文書 n の D-次元特徴ベクトルを $\mathbf{x}_n$ とし, ノルムは $\|\mathbf{x}_n\| = 1$ に正規化されているとする. 一方, 文書クラスタを $\{C_1, \cdots, C_j, \cdots, C_k\}$ とし, クラスタ j のセントロイド D-次元ベクトルを $\mathbf{y}_j = |C_j|^{-1} \sum_{n \in C_j} \mathbf{x}_n$ とすれば, . 球面 k-means 法 [2] で最大化する目的関数は

$$f(C_1, \cdots, C_k) = \sum_{j=1}^{k} \sum_{n \in C_j} \frac{\mathbf{x}_n^T \mathbf{y}_j}{\|\mathbf{y}_j\|} \tag{1}$$

となり, 各文書クラスタとの関係は

$$C_j = \{n \mid j = \arg\max_{1 \le i \le k} (\mathbf{x}_n^T \mathbf{y}_i)\} \tag{2}$$

となる. ここで $\mathbf{x}_n^T$ は $\mathbf{x}_n$ の転置をを表す.

以下では, 球面 k-means 法の結果に基づく群平均法, ウォード法, コサイン法をそれぞれ導出する. まず, クラスタ i と j 特徴ベクトル間の距離 2 乗平均は

$$\frac{1}{|C_i||C_j|} \sum_{n \in C_i} \sum_{m \in C_j} \|\mathbf{x}_n - \mathbf{x}_m\|^2 = 2(1 - \sum_{n \in C_i} \frac{\mathbf{x}_n^T}{|C_i|} \sum_{m \in C_j} \frac{\mathbf{x}_m}{|C_j|}) \tag{3}$$

となり, 群平均法でのクラスタ i と j 間の非類似度 $d_a(i, j)$ は

$$d_a(i, j) = 2(1 - \mathbf{y}_i^T \mathbf{y}_j) \tag{4}$$

となる. 一方, クラスタ i と j を併合させたクラスタを

$$\mathbf{y}_{i \cup j} = \frac{1}{|C_i| + |C_j|} \sum_{n \in C_i \cup C_j} \mathbf{x}_n \tag{5}$$

とすれば, 合併後と前でのセントロイドとの距離 2 乗総和の差は

$$\sum_{n \in C_i \cup C_j} \|\mathbf{x}_n - \mathbf{y}_{i \cup j}\|^2 - (\sum_{n \in C_i} \|\mathbf{x}_n - \mathbf{y}_i\|^2 + \sum_{n \in C_j} \|\mathbf{x}_n - \mathbf{y}_j\|^2) \tag{6}$$

となり, 式 6 を変形し, ウォード法の非類似度 $d_w(i, j)$ は

$$d_w(i, j) = |C_i| \|\mathbf{y}_i\|^2 + |C_j| \|\mathbf{y}_j\|^2 - (|C_i| + |C_j|) \|\mathbf{y}_{i \cup j}\|^2 \tag{7}$$

となる. これに対し, 合併前と後でのセントロイドとのコサイン類似度総和の差は

$$\sum_{n \in C_i} \frac{\mathbf{x}_n^T \mathbf{y}_i}{\|\mathbf{y}_i\|} + \sum_{n \in C_j} \frac{\mathbf{x}_n^T \mathbf{y}_j}{\|\mathbf{y}_j\|} - \sum_{n \in C_i \cup C_j} \frac{\mathbf{x}_n^T \mathbf{y}_{i \cup j}}{\|\mathbf{y}_{i \cup j}\|} \tag{8}$$

となり, 式 8 を変形し, コサイン法の非類似度 $d_c(i, j)$ は

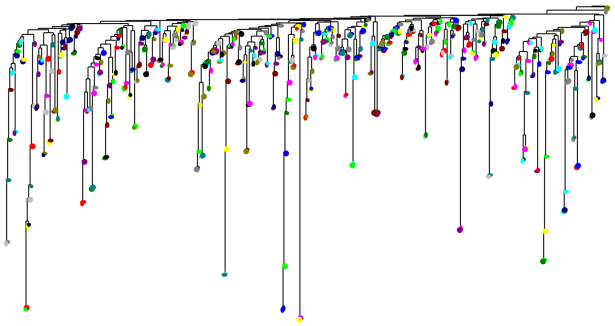
$$d_c(i, j) = |C_i| \|\mathbf{y}_i\| + |C_j| \|\mathbf{y}_j\| - (|C_i| + |C_j|) \|\mathbf{y}_{i \cup j}\| \tag{9}$$

となる. コサイン法は, 式 (1) の目的関数の減少が最も小さいクラスタペアの選択と見なせる.

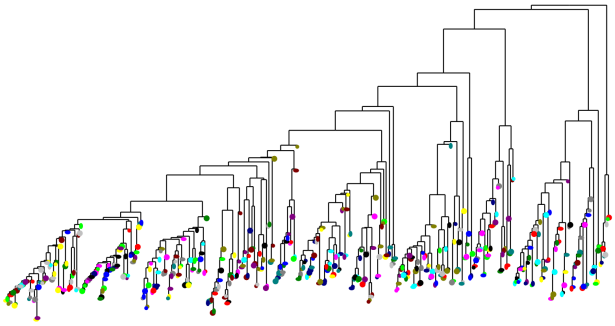
3　実験による評価

実験では，ID 付に限定し, 文書数 $N = 299,752$ で, 語彙数 $D = 102,660$ の NYTimes news articles [3] を利用した. 図 1 には，クラスタ数 $k = 500$ での, 球面 k-means 法の結果に対し, 群平均法, ウォード法, 及び, コサイン法のそれぞれで得られたデンドログラムの例を示す. ノード配色については, 無作為に設定したクラスタ番号に対し, CMYK カラーシステム 24 色をサイクリックに割り当てた. いま, 各手法での最初から最後までの併合非類似度を $d_*^{(1)}, \cdots, d_*^{(k-1)}$ と表記する. 図 1 より，群平均法の結果は, ウォード法やコサイン法と比較して, $d_a^{(k-1)}$ に近い値での併合が多く, 多くの併合点がデンドログラム上部に偏っていることが見て取れる. 一方, デンドログラム最上部の最終併合点をルートと見なし, 各手法での元クラスタ C_j までのステップ数 (深さ) を $s_*(j)$ と表記

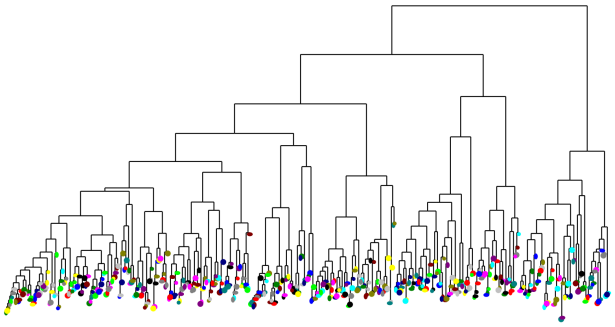
Hybrid Document Clustering of Divisive and Agglomerative Types Based on Cosine Similarity
†Yueyan FANG　†Kazumi SAITO
†Kanagawa University



(a) 群平均法



(b) ウォード法

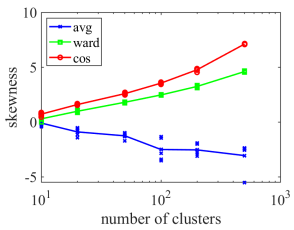


(c) コサイン法

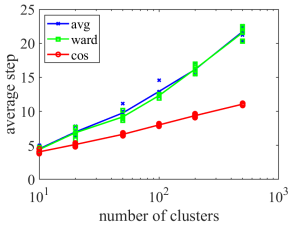
図 1: クラスター数 500 でのデンドログラムの例

する. 図 1 より，群平均法とウォード法と比較して, コサイン法の結果は, 長いステップ数 $s_c(j)$ のクラスタは少なく, 平衡 (self-balancing) なデンドログラムを構成していることが見て取れる.

このような傾向の一般性を検証するために, クラスタ数を $k \in \{10, 20, 50, 100, 200, 500\}$ に設定し, 球面 k-means 法の初期値を変えた 5 回の試行で, 非類似度の歪度 $z_*^3 = \sum_{j=1}^{k}(d_*^{(j)} - \mu_*)^3/((k-1)\sigma_*^3)$ と平均ステップ数 $\bar{s}_* = \sum_{j=1}^{k} s_*(j)/(k)$ を評価した. ここで, μ_* と σ_* は各手法での非類似度の平均と標準偏差を表す. 図 2 には，群平均 (avg) 法 , ウォード (ward) 法, 及び, コサイン (cos) 法それぞれの各試行結果を青クロス, 緑四角, 赤丸のマーカで表し, 5 回の試行の平均値を直線でつないでプロットした結果を示す. 図 2 (a) より, 群平均法で



(a) 歪度



(b) 平均ステップ数

図 2: デンドログラムの定量評価

は歪度が負となり平均より大きい併合非類似度が多く, これに対し, ウォード法とコサイン法では歪度が正となり平均より小さいものが多く, クラスタ数 k が大きくなるにつれ, この傾向は顕著になることも分かる. 一方, 図 2 (b) より, 群平均法とウォード法の平均ステップ数はコサイン法より大きく, クラスタ数 k が大きくなるにつれ, この傾向も顕著になることが分かる. すなわち, コサイン法は他と比較して, 併合非類似度が小さく, 平衡なデンドログラムを構成していることが分かる. なお, これら 3 手法の定量評価値に関しては, 比較的コサイン法の結果が安定していることも分かる.

4 おわりに

本研究では, 球面 k-means 法の結果に凝集型階層的クラスタリングする群平均法, ウォード法, 及び, コサイン法を導出した. 現実大規模データを用いた評価実験では，これら 3 手法によるデンドログラムを定性評価するとともに, コサイン法は他と比較して, 併合非類似度が小さく, 平衡なデンドログラムを構成していることを定量評価した. 今後の研究では，さらに多様なデータへの適用評価実験により, コサイン法の特性などを明らかにする.

参考文献

[1] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In Proc. of the 15th Annual ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 318?329, 1992.

[2] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. Machine Learning 42, pp.143?175, 2001.

[3] D. Dua and E. Karra Taniskidou. Bag of Words Data Set (NYTimes news articles) in UCI Machine Learning Repository. University of California, 2008