ファジィクラスタリングを用いた階層型文書クラスタリング

村上浩司　乾孝司　橋本泰一　石川正道
東京工業大学統合研究院

1　はじめに

電子化文書の増加に伴い，膨大な文書を俯瞰的，効率的に分類，整理することが非常に重要になっている．これを実現する 1 つの技術として文書クラスタリングがある．結果として得られたクラスタに属する文書は，出現する単語の分布が類似していることからクラスタ別に文書を分析することで，それぞれのクラスタが示すトピックなどを大まかに捉えることができる．階層型クラスタリング手法では，クラスタ間の類似度によりクラスタの結合 (もしくは分割) が行われ，その全過程がデンドログラムによって示される．これにより，対象の文書群中において各クラスタが示すトピックの関連性なども，俯瞰的に捉えることが可能となる．

階層型クラスタリングに関する研究は数多く行われているが [6]，その多くは文書が複数のクラスタへの所属を許さない，ハードクラスタリングが対象である．我々は現在，新聞記事からの社会課題抽出 [11] を検討しているが，新聞記事などは一般的に，ただ 1 つのトピックについてのみ記述されているだけでなく，複数のトピックについて書かれていることが多い．実際に新聞記事群をクラスタリングし，デンドログラムを参考にしながら得られたクラスタ内の文書を分析する際に，ある文書が所属するクラスタ A だけでなくクラスタ B にも属することで，クラスタ B の分析が容易になるという例は少なくない．

クラスタリング手法には，クラスタへの所属の重複を許すファジィ(ソフト) クラスタリング手法があり [1]，多重トピックの文書群の分類に有効であると考えられる．しかしファジィクラスタリングは非階層型のクラスタリング手法であり，クラスタリング結果の分析に有益なデンドログラムを出力することができない．

そこで我々は，複数のトピックを持つ文書を含む文書群から俯瞰的な階層を得ることを目的として，これら 2 つの異なるクラスタリング手法を組み合わせた新しい階層型ファジィクラスタリング手法を提案する．

提案手法は，対象の文書群を N 個のクラスタに分けたいとするとき，まずはじめに文書-単語行列を入力としてファジィクラスタリングにより文書集合を M 個 $(M > N)$ のクラスタに分類する．そして，ファジィクラスタリングによって計算された各クラスタの中心ベクトルから，クラスタ-中心ベクトル行列を作成し，これを新たな入力として階層型ハードクラスタリング手法により，必要な N 個のクラスタに分類する．

2　関連研究

階層型ハードクラスタリングは，クラスタリングの中でも中心的な手法であり，これまで多くの研究がなされてきた．Ying らは文書-単語行列のような高次元かつスパースである大規模データセットに対して高精度で分類のできる階層型のハードクラスタリングを提案し [6, 7]，更にこれらの成果をツールとして公開している [4]．

ファジィクラスタリングは，複数のクラスタに対象のインスタンスが所属することを許すクラスタリング手法であり，Fuzzy C-Means 法 (FCM)[1] は最も広く利用されているアルゴリズムの 1 つである．Mendes らは，このアルゴリズムを文書分類に適した形へと改良し，非階層型クラスタリングの 1 つである k-means 法よりも高い分類精度を示した [5]．

こうした階層的クラスタリングとファジィクラスタリングの概念を結合させた研究には，例えば Ying ら試みがある [8]．これは，階層型クラスタリングの枠組みで，一般的な評価関数とそれらをファジィに拡張したものとを比較し，ファジィ化した評価関数の優位性を示したものである．しかしながら彼らは文書の各クラスタへの帰属度を最適化し，ハードクラスタリングとして手法を評価しているため，複数のクラスタに対象の文書が所属することを許していない．また遠藤らはクラスタへの帰属度からクラスタ数の自動推定ができる階層的ファジィクラスタリング手法を提案した [9]．彼らは，星図データに対象に有効性を示したが，文書分類への評価は行われていない．Ricardo らは，Web ページ検索を目的とした，階層的なファジィクラスタリング [3] を用いたメタ検索エンジンを提案している [2]．しかしながら，用いられているクラスタリング自身は，ルールを獲得するためのもので直接文書分類には用いられていない．

3　提案手法

本手法は，2 つの異なったタイプのクラスタリング手法を組み合わせて，複数のトピックを持つ文書を含む

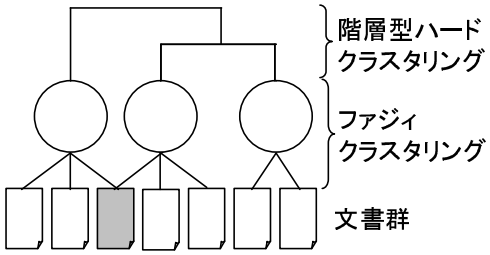

図 1: 提案手法の構成

文書群から俯瞰的な階層を得ることが目的である．それぞれのクラスタリングにおける役割と考え方を簡単に説明する．

3.1 ファジィクラスタリング

ファジィクラスタリングの目的は，文書が複数のクラスタに所属することを許しながら対象の文書群を複数のクラスタに分類することである．これにより，大まかに文書群のトピックを分類することができる．ファジィクラスタリングの代表的な手法のひとつであり，実験にも用いた Fuzzy C-Means 法について簡単に説明する．この手法は，クラスタリングの対象 x_j が単一のクラスタのみに属するのではなく，各クラスタ C_k への帰属度 $u_{kj} \in [0,1]$ を求めるものである．

文書集合を $\mathbf{d_j}$=$\{d_1,\ldots,d_n\}$，文書に含まれる単語集合を $\mathbf{t_i}$=$\{t_1,\ldots,t_m\}$ とすると，文書 $\mathbf{d_j}$ の特徴を表す文書ベクトル $\{d_{j1},\ldots,d_{jm}\}$ のベクトルは，TFIDF を用いて表すと $d_{ji} = tf_{ji} \cdot idf_i$ ， $idf_i = log(n/n_i)$ となる．tf_{ji} は文書 $\mathbf{d_j}$ における単語 t_i の出現回数，n_i は単語 t_i が出現する文書数とする．

このベクトルを用いて，Fuzzy C-Means 法により各文書 $\mathbf{d_j}$ のクラスタ C_k への帰属度 u_{kj} を求める．アルゴリズムは以下のようになる．

入力 クラスタ数 c，ファジィ化パラメータ $m(>1)$，文書ベクトル $\mathbf{d_j}(j=1,\ldots,n)$

出力 帰属度 $u_{kj}(j=1,\ldots,n;k=1,\ldots,c)$，クラスタ中心 $\mathbf{v_k}(k=1,\ldots,c)$

1. 文書 $\mathbf{d_j}$ のクラスタ C_k への帰属度 $u_{kj}(j=1,\ldots,n;k=1,\ldots,c)$ の初期値をランダムに定める．

2. 各クラスタの中心 $\mathbf{v_k}$ を下の式で求める．

$$\mathbf{v_k} = \frac{\sum_{j=1}^{n}(u_{kj})^m \mathbf{d_j}}{\sum_{j=1}^{n}(u_{kj})^m} \tag{1}$$

3. 帰属度 $u_{kj}(j=1,\ldots,n;k=1,\ldots,c)$ を下の式で求める．$sim(\mathbf{d_j},\mathbf{v_k})$ は文書 $\mathbf{d_j}$ とクラスタ中心 $\mathbf{v_k}$ との類似度を示す．

$$u_{kj} = \left[\sum_{k'=1}^{c}\left(\frac{sim(\mathbf{d_j},\mathbf{v_{k'}})}{sim(\mathbf{d_j},\mathbf{v_k})}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{2}$$

4. クラスタ中心が変化しなくなるまで 2. と 3. を繰り返す．

3.2 階層型ハードクラスタリング

階層型ハードクラスタリングの目的は，クラスタ間の類似度によってクラスタを結合させ，階層を構築することである．これにより，ファジィクラスタリングによって得られた各クラスタの関連を俯瞰的に示すことが可能になる．階層型ハードクラスタリングには階層併合方法と階層分割方法があるが，本論文中の実験では，階層併合方法を選択した．アルゴリズム中の類似度計算法には，単一リンク法，完全リンク法，UPGMA(群平均法) などがある [10]．アルゴリズムは以下のようになる．ここで，$C=\{C_1,\ldots,C_c\}$ をクラスタ集合とする．

入力 クラスタ数 c，文書ベクトル $\mathbf{d_j}(j=1,\ldots,n)$

出力 デンドログラムを作成するためのクラスタ結合の順序，クラスタ間の類似度などの情報

1. 個々の文書をクラスタとする．すべての文書 $\mathbf{d_j}(j=1,\ldots,n)$ について，$|C| := n$，$C_i := \mathbf{d_j}$，$sim(C_i,C_j) := sim(\mathbf{d_i},\mathbf{d_j})$，$1 \le i,j \le n, i \ne j$ とする．

2. 最大類似度のクラスタ対を検索して結合する．$s(C_q,C_r) = \max_{i,j} sim(C_i,C_j)$，$C_q$ と C_r を C から除き，$C' = C_q \cup C_r$ を C に追加，$C := C-1$，$C=1$ ならば終了する．

3. すべての $C_i \in C, C_i \ne C'$ についてクラスタ間の類似度を再計算する．

3.3 クラスタリングの結合

提案手法では，ファジィクラスタリングで得られた結果であるクラスタの情報を擬似的な文書として扱い，疑似文書を階層型ハードクラスタリングの入力として分類することで，最終的な結果とする．図 1 に，本手法の全体的な構造を示す．得られる結果には，ファジィクラスタリングにより，文書が複数のクラスタに属する情報を有したまま，デンドログラムによりクラスタ間の関連が得られることになる．ファジィクラスタリングの出力を反映した階層型ハードクラスタリングの入力には，再度，各クラスタ毎に所属する文書からベクトルを作成することも考えられるが，本論文では，手法全体で用いる情報を一貫させるために，階層型ハー

ドクラスタリングへの入力には，ファジィクラスタリングによって得られるクラスタの中心 u_k を用いることとした．

4 評価実験

提案手法を評価するために，文書群に対して直接，階層型ハードクラスタリングを行った場合 (AHC) と，提案手法であるファジィクラスタリングにより得られたクラスタの中心を入力として，階層型ハードクラスタリングを行った場合 (FCM+AHC) とを比較する．

4.1 実験条件

実験のためのテキストデータには，Reuters-21578 コレクション[1]を用いた．まず ModApte 分類 (LEWIS-SPLIT が”TEST” もしくは”TRAIN”，TOPIC が”YES” である 9603 文書) を切り出した．切り出された文書群から，より多くの文書が存在するトピックを選択しデータセットを作成した．作成したデータセットは 2 種類である．データセット 1 (reuter1) は，トピックラベルが “acq”，“earn” の 2 種類を対象とした．これら 2 種類のラベルが付与されているデータと，どちらのラベルも含むデータをまとめた．データセット 2 (reuter2) では，対象としたトピックラベルは money-fx”，”ship”，”interest”，”trade”，”crude” の 5 種類であり，これらのいずれか 1 つのトピックラベルが付与されている文書と，“money-fx+interest” と “crude+ship” の 2 種類の複数トピックラベルが付与された文書を組み合わせた．reuter1 においては複数のトピックラベルが付与されている文書が全体の 2.8%，reuter2 においては 28.7%となっている．各文書セットで用いた文書数の詳細は表 1 に示す．括弧内はそれぞれ文書数および異なり単語数である．

4.2 クラスタリングアルゴリズム

本手法は，2 つの異なるクラスタリング手法を組み合わせることで実現される．階層型ハードクラスタリング，ファジィクラスタリングともに，多くのアルゴリズムが提案されている．本実験ではまず，提案手法の概念の可能性を検証することを目的とする．そのため，種々のクラスタリングアルゴリズムの組合わせによる性能評価はここでは行わない．我々は，最も広く使われているアルゴリズムのひとつである，UPGMA 法 (群平均法) と，Fuzzy C-Means 法をそれぞれ，階層型ハードクラスタリング，ファジィクラスタリングのアルゴリズムとして用いることとした．また文書群に対して直接，階層型クラスタリング行う場合も同様に UPGMA 法を用いた．

[1]Reuters-21578 test collection:
http://www.daviddlewis.com/resources/testcollections/

表 1: データセットの詳細

文書セット	トピックラベル	文書数
reuter1 (639*4076)	acq	229
	earn	392
	acq+earn	18
reuter2 (700*5084)	money-fx	143
	interest	114
	ship	70
	crude	172
	money-fx+interest	163
	crude+ship	38

4.3 評価尺度

本実験では，ある正解クラスの文書がそれぞれのクラスタにどれくらい適切に所属するかに着目する必要がある．そこで評価尺度として，Entropy，Purity，Recall，Precision，F-score の 5 つを用いた．Recall と Precision，F-score は一般的に用いられる評価尺度であるため，ここでは説明を割愛する．Entorpy は各クラスの文書がどのようにそれぞれのクラスタに割り当てられたかを表し，Purity は，クラスタ C_r に属する文書が，クラスタ内の多数を占める正解クラスに分類されたと考えたときの正解率である．Entropy，Purity ともに範囲は [0,1] であり，最適なクラスタリング行われた場合，ただ 1 つのクラスの文書が 1 つのクラスタに所属することになる．この場合は Entoropy は 0，Purity は 1 となる．

クラスタ C_r の大きさを n_r としたときの Entropy は，

$$E(C_r) = -\frac{1}{log(c)} \sum_{i=1}^{c} \frac{n_r^i}{n_r} log \frac{n_r^i}{n_r} \tag{3}$$

で定義され，c はクラス数，n_r^i は i 番目のクラスの文書のうち，r 番目のクラスタに所属する文書の数を表す．クラスタリング全体の Entropy は，下の式で示すように，重みを付けた各クラスタのエントロピーの合計となる．

$$Entropy = \sum_{r=1}^{k} \frac{n_r}{n} E(C_r) \tag{4}$$

また，クラスタ S_r および，全体の Purity は以下のように定義される

$$P(C_r) = \frac{1}{n_r} \max_i (n_r^i) \tag{5}$$

$$Purity = \sum_{r=1}^{k} \frac{n_r}{n} P(C_r) \tag{6}$$

表 2: 比較実験結果

データ	手法	Entropy	Purity	Recall	Precision	F-score
reuter1	AHC	0.70	0.62	0.35	0.51	0.42
	FCM+AHC	0.68	0.64	0.49	0.54	0.51
reuter2	AHC	0.88	0.45	0.27	0.51	0.35
	FCM+AHC	0.82	0.54	0.48	0.30	0.37

4.4 実験結果と考察

表 2 に，階層型ハードクラスタリング (AHC) を直接行った場合と，ファジィクラスタリング (FCM) と階層型ハードクラスタリング (AHC) を組み合わせた場合のクラスタリング結果を示す．FCM+AHC においては第 2 位の帰属度が平均 (1/クラスタ数) より高い場合，帰属度が第 1，2 位のクラスタに文書が所属すると見なし，AHC で得られる最終的なクラスタを評価した．

reuters1 セットにおいては，どちらの手法を用いても Entropy と Purity には大きな差はないが Recall や Precision，F-score では FCM+AHC の方が高いスコアであることがわかる．これは，複数のクラスタへの重複を許すことにより，第 1 位の帰属度のクラスタへの分類が誤りであった場合に，第 2 位の帰属度のクラスタへの分類が正しければ，そのインスタンスが評価の対象になるためであると考えられる．

また reuter2 セットでは，FCM+AHC において，Entropy および Purity はセット 1 と同様改善されたものの，Precision が大幅に下がっている．これは，第 2 位の帰属度のクラスタへの分類も誤った場合が多いのが原因であると考えられる．

どちらのセットにおいても精度全体では改善が見られたことから，クラスタリングを行う際に複数の帰属度を利用することは一定の有効性があることが確認された．しかしながら両データセットともに，直接 AHC を行った場合でも Entropy が高いことから，クラスタリング自体の精度が低いことが分かる．理由として，分割したいクラス数に対してベクトル行列の次元が大きいために，ノイズとなる素性が分類の精度に影響を与えていると考えられる．より適切な評価を行うためには行列のスパースネスを考慮した上で，よりクラス数の多いデータセットの再構築が必要である．

5 まとめと今後の課題

複数のクラスタへの所属を許すファジィクラスタリングと，デンドログラムを生成する階層型ハードクラスタリングを併用する新しいクラスタリング手法を提案した．新聞記事を用いた評価実験により，階層型ハードクラスタリングを直接用いる場合と比較した結果，提案手法はクラスタリングとしての精度が改善できる可能性を示した．

今後の課題としては，ファジィクラスタリングにおける文書の重複したクラスタへの所属が，より広範囲に行われるためのアルゴリズムの改良，評価に適したデータセットの作成と手法全体の評価などを考えている．

参考文献

[1] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York:Plemium Press, 1981.

[2] Ricaldo Campos, Gael Dias, and Celia Nunes. Wise: Hierarchical soft clustering of web page search results based on web content mining techniques. In *Proc. IEEE/WIC/ACM International Conference on Web Intelligence (WI '06)*, pp. 301–304, 2006.

[3] Guillaume Cleuziou, Lionel Martin, and Christel Vrain. Disjunctive learning ith a soft-clustering method. In *Proc. of Inductive Logic Programming: 13th International Conference (ILP'03)*, pp. 75–92, 2003.

[4] Geoge Karypis. *CLUTO A Clustering Toolkit Release2.1.1*. University of Minnesota, Department of Computer Sciense, 2003.

[5] M.E.S. Mendes and L. Sacks. Evaluating fuzzy clustering for relevance-based information access. In *Proc. The IEEE International Conference on Fuzzy Systems*, pp. 648–653, 2003.

[6] Ying Zhao and Geoge Karypis. Criterion function for document clustering. Technical report, Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, 2003.

[7] Ying Zhao and Geoge Karypis. Hierachical clustering algorithms for document datasets. Technical report, Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, 2003.

[8] Ying Zhao and Geoge Karypis. Soft clustering criterion function for partitional document clustering. In *Proc. Conference on Knowledge Manageent*, pp. 246–247, 2004.

[9] 遠藤靖典, 山口真吾. クラスタ数推定機能をもつ階層的ファジィクラスタリング. 電子通信情報学会論文誌, Vol. J79-A, pp. 1276–1288, 1996.

[10] 宮本定明. クラスター分析入門. 森北出版, 1999.

[11] 大熊和彦. 新しい大学研究「ソリューション研究」の意義と課題. 研究・技術計画学会 第 21 回年次学術大会予稿集, pp. 88–91, 2006.