ユーザの観点に対応した多重解像度 NMF によるマルチラベル文書の分類

丸田要　永井秀利　中村貞吾
(九州工業大学)

1　はじめに

文書集合を効率良く整理・検索する手法の一つとして検索結果をクラスタリングする手法がある．しかし，ユーザの望む分類でなければ見当違いのカテゴリを探してしまうため，効率的に目的の文書を探す事ができない．さらに，ユーザの望む分類を行う際，分類を行うユーザの目的・観点により分類結果が異なってしまう．特にマルチラベル文書と呼ばれるデータは複数のクラスに属し，クラス間の類似度が高い場合が多いため多少の観点の違いにより分類結果が変わり易い．その場合ユーザが望む分類とシステムによる分類に差異ができ，その差異部分に含まれる文書データはユーザの情報検索の阻害や見落としを発生させると考えられる．そこで，ユーザの観点を反映した分類を目指す．

以前の論文 [1] で検証した 3 つの分類手法の中では NMF-I[2] に適用した手法が有効であった．そこで本論文では，論文 [1] の際と同様にユーザの観点情報を文書分類例から算出し，その観点情報を反映する文書分類手法の新たな提案手法として多重解像度 NMF[3] を用いてこれまでの手法と比較する．

2　提案手法

ユーザの観点は人の感覚に依るところが大きいため，ユーザが文書分類を行う際の観点を明示的に表現することは困難である．そこで，ユーザが分類した教師文書から観点の特徴を抽出し，それを分類に反映させる．

本論文では観点の特徴は教師文書中の単語の分布に現れると考えた．そこで，各クラスに対する単語の寄与度を求め，全寄与度の集合を近似的な観点として表現する．

提案手法の流れを以下に示す．まず，観点抽出手法により寄与度を算出する．全寄与度の値を観点行列 $U_m \in \mathbb{R}^{w\times k}$ として扱う．ここで，w は単語数，k はクラス数である．次に，行列分解を利用した分類手法に観点行列 U_m を導入し，ユーザの観点を反映した分類を行う．現在は，観点抽出手法として 4 つ (EM-1，EM-2，EM-3，EM-4) の方法と，観点行列を導入した分類手法として 2 つ (CM-I，CM-M) の方法を考え検証している．

3　観点抽出手法 - Extraction Method(EM)

各観点抽出手法は各クラスに対する単語の寄与度を求める手法である．各抽出手法の説明では任意のクラス A に対する単語 t の寄与度の算出式について述べる．$D = \{d_1, d_2, ..., d_n\}$ を全文書ベクトルの集合，n は文書数とする．その時，$D_A(\subseteq D)$ はクラス A 所属の文書ベクトル集合であり，$\overline{D}_A$ は D_A の補集合である．特徴量である $f(D,d,t)$ は文書 d における単語 t に対する TF・IDF 値を使用する．

3.1　EM-1 - 平均値

EM-1 は，D_A に属する全文書における単語 t に対する特徴量の平均値をクラス A に対する単語 t の寄与度とする．EM-1 では，単語 t がクラス A に属する多くの文書に沢山出現する時，寄与度は高くなる．

$$\underset{d\in D_A}{\text{mean}}\ f(D,d,t) \tag{1}$$

3.2　EM-2 - 平均値+比率

EM-2 は，EM-1 で求めた寄与度に対してクラス間の比率を使用する．つまり，EM-2 における寄与度は式 (2) で算出する．EM-2 では，単語 t がクラス A 以外のクラスに属する文書に多く出現する時，寄与度は低くなる．

$$\underset{d\in D_A}{\text{mean}}\ f(D,d,t)/\underset{d\in \overline{D}_A}{\text{mean}}\ f(D,d,t) \tag{2}$$

3.3　EM-3 - 最大値

EM-3 は，D_A の全文書における単語 t に対する特徴量の最大値をクラス A に対する単語 t の寄与度とする．単語 t に対して，文書 $d \in D_A$ における特徴量が高く，それ以外の D_A に属する文書における特徴量が全て低い場合を考える．その場合，EM-1 の寄与度は平均化されるため低くなり，EM-3 の寄与度は最大値を取るため高くなる．

$$\max_{d\in D_A}\ f(D,d,t) \tag{3}$$

3.4　EM-4 - 最大値+比率

EM-4 は，EM-2 の様に EM-3 で求めた寄与度に対してクラス間の比率を使用する．

$$\max_{d\in D_A}\ f(D,d,t)/\max_{d\in \overline{D}_A}\ f(D,d,t) \tag{4}$$

4　関連研究

4.1　NMF

NMF[4] は n 個の文書データと w 個の索引語から作られる索引語文書行列 $X \in \mathbb{R}^{w\times n}$ を基底行列 $U \in \mathbb{R}^{w\times k}$ と特徴行列 $V^T \in \mathbb{R}^{w\times k}$ の積に分解することで文書分類を行う．

$$X \simeq UV^T \tag{5}$$

ここで，文書行列 X の要素である特徴量は TF・IDF 値であり，V^T の h 行目の要素の値が，各文書と h 番目のクラスタとの関連度の大きさを表している．

U と V は式 (6) の目的関数 J を最小にするような更新式の繰り返しにより求める．

$$J = ||X - UV^T|| \tag{6}$$

U と V を求める更新式は，式 (7) である．

$$V_{ij} \leftarrow V_{ij}\frac{(X^TU)_{ij}}{(VU^TU)_{ij}}\quad ,\qquad U_{ij} \leftarrow U_{ij}\frac{(XV)_{ij}}{(UV^TV)_{ij}} \tag{7}$$

ここで，$(X)_{ij}$ は行列 X の i 行 j 列の要素を表す．NMF には初期値の問題が存在する．その問題とは，NMF が初期値に依存してクラスタリング結果が変化し，通常は初期値として乱数を使用しているため，使用する乱数によってはクラスタリング結果として悪い局所解に収束してしまうことである．

4.2　NMF-I

我々が以前の論文で NMF を教師あり手法へと拡張した手法の一つに NMF-I[2] がある．この NMF-I は既述した初期値の問題に対応している．

NMF-I では，教師文書ベクトルを用いて各クラスタにおける平均ベクトルを求める．それが教師文書における理想的な基底ベクトルに近いものであろうとの期待に基づ

き，その平均ベクトルを基底ベクトルの初期値にする．つまり，式 (8) で求まる教師基底行列 U_s を教師あり NMF における基底行列 U の初期値とした．

$$U_s = X_{\text{train}}(V_{\text{train}}^T)^+ \tag{8}$$

ここで，教師データ数を t とした時，教師文書行列 $X_{\text{train}} \in \mathbb{R}^{n\times t}$ は教師文書のみの索引語行列であり，$V_{\text{train}}^T \in \mathbb{R}^{k\times t}$ は各文書の正解クラスタに対応する要素を 1 とし，それ以外の要素を 0 とした行列である．また，"+" は疑似逆行列である．

4.3 SSNMF

SSNMF[5] は H.Lee 等が提案した半教師あり NMF の一つである．SSNMF では，教師情報を含んだ制約項を目的関数に追加することで収束方向を制御している．SSNMF における目的関数を式 (9) に示す．

$$J_s = ||X - UV^T||^2 + \lambda||L * (Y - WV^T)||^2 \tag{9}$$

$Y \in \mathbb{R}^{k\times n}$ は教師文書の正解クラスタを 1 とし，それ以外を 0 としている．$W \in \mathbb{R}^{k\times k}$ は制約項における基底行列である．そして，$L \in \mathbb{R}^{k\times n}$ は重み行列である．

5 分類手法 - Classification Method(CM)

5.1 NMF-I へ観点行列を導入 (CM-I)

CM-I では 4.2 節で述べた NMF-I を利用する．NMF-I の教師基底行列 U_s の代わりに観点行列 U_m を NMF の初期値にする．それにより，U_m を初期値とした更新式による柔軟な収束ができると期待できる．

5.2 多重解像度 NMF へ観点行列を導入 (CM-M)

多重解像度 NMF は画像処理において異なる解像度を利用することで各解像度で抽出できる 2 つの要素を同時に扱うことができ細胞検出などで有効性が示されている [3]．この多重解像度 NMF を文書分類に適用させた手法を提案する．文書分類における低解像度は単語を素性とした文書ベクトルとし，高解像度は bigram を素性とした文書ベクトルとする．伊東等の論文 [3] とは異なり文書分類を目的とした多重解像度 NMF の目的関数として式 (10) を提案する．

$$\begin{aligned} J_m = &||X - U_1V^T||^2 + \mu||Y - MU_2V^T||^2 \\ &+ \lambda||U_1 - M^TMU_2||^2 \end{aligned} \tag{10}$$

ここで，$Y \in \mathbb{R}^{b\times n}$ は bigram を素性にした高解像度文書行列であり，b は bigram の要素数である．$M \in \mathbb{R}^{b\times w}$ は各 bigram 要素とそれを構成する単語群の関係を表す行列であり，構成する単語に対応する要素は 0.5，それ以外の要素は 0 である．μ と λ は各項に対する重みである．式 (10) の第三項目は各解像度における基底行列の差を小さくする．目的関数 J_m から求めた更新式は式 (11)，(12)，(13) である．

$$V_{ij} \leftarrow V_{ij}\frac{(X^TU_2 + \mu Y^TMU_2)_{ij}}{(VU_1^TU_1 + \mu VU_2^TM^TMU_2)_{ij}} \tag{11}$$

$$U_{1ij} \leftarrow U_{1ij}\frac{(XV + \lambda M^TMU_2)_{ij}}{(UV^TV + \lambda U_1)_{ij}} \tag{12}$$

$$U_{2ij} \leftarrow U_{2ij}\frac{(M^TYV + \lambda U_1)_{ij}}{(M^TMUV^TV + \lambda M^TMU_2)_{ij}} \tag{13}$$

式 (12)，(13) の初期値に観点行列を使用することで多重解像度 NMF へ観点を導入する．

6 実験

実験では，4 つの抽出手法と 2 つの分類手法を組み合わせた 8 通りの提案手法を検証する．そして，既存 NMF，NMF-I，SSNMF，NaiveBayes(NB)，SVM，多項ロジスティック回帰 (MLR) と 8 通りの提案手法の分類性能を調査し比較する．

シングルラベルとマルチラベルを混合させたデータセットを使用する．混合データは Web 朝日[1]で公開されている記事を利用する．

表 1: データセット

Data	docs	terms	class	Data	docs	terms	class
ps	99	3149	2	et	500	7048	2
se	149	3317	2	it	500	7374	2

実験の評価値には Entropy，Purity，Precision，Recall，RandIndex を用いる．最終的な分類性能値は上記の五種類の評価値の調和平均 Hm とする．実験は初期値と教師文書を変化させた 20 セットの結果に対して評価値の平均を求めた．NMF の更新回数は 20 回，教師文書は各クラスに 10 文書とした．

現在，混合データでは観点の違いを簡単に比較するために 2 クラス文書データを対象に実験を行う．実験では，全てのマルチラベル文書の正解ラベルが 2 つの内どちらか一方に偏っている擬似的な 2 セットの観点を使用した．

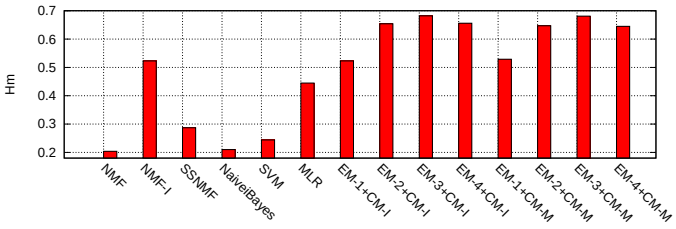

図 1: Macro Average of Hm

EM-3+CM-I と EM-3+CM-M が比較手法の中では良い分類性能を示している．CM-I と CM-M を比較すると，同じ観点抽出手法を使用した場合，CM-I と CM-M の平均化された評価値は同等である．しかし，結果を詳細に分析すると使用するデータにより最も良い評価値を出す手法が変わる．そのため，データによっては bigram を使用した多重解像度 NMF は有効であると考えられる．

7 まとめ

観点を反映させた文書分類手法として NMF-I を利用した手法と多重解像度 NMF を利用した手法を提案し，実験によりシングルラベルデータとマルチラベルデータの混合データでの分類性能を示した．そして，多重解像度 NMF を利用した文書分類への適用方法の更なる調査と 3 クラス以上の混合データに対する実験が今後の課題である．

参考文献

[1] K.Maruta, H.Nagai, T.Nakamura, "Document Classification using Matrix Decomposition with Varied Viewpoints", IEE, AAI, Vol.1, No.4, pp65-74, 2015.

[2] 丸田要，永井秀利，中村貞吾，"文書分類のための教師制約を用いた非負値行列因子分解"，IAS2013，情報アクセスシンポジウム 2013，pp.14-21，(2013).

[3] 伊東翼，太田圭輔，村山正宣，青西享，"多重解像度非負値行列因子分解によるカルシウムイメージングデータ解析"，IEICE Technical Report，Vol.115，No.514，pp131-136，2016.

[4] D.D.Lee , H.S.Seung "Algorithms for Non-negative Matrix Factorization", NIPS , pp.556-562 , (2000).

[5] H.Lee , J.Yoo , S.Choi "Semi-Supervised Nonnegative Matrix Factorization", IEEE SIGNAL PROCESSING LETTERS , Vol.17 No.1 , pp.4-7，(2010).

[1] http://asahi.com