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

丸田 要[1] 永井 秀利[1] 中村 貞吾[1]

概要：非負値行列因子分解 (NMF：Non-negative Matrix Factorization)[1] は，索引語文書行列 X を各クラスタに対する基底行列 U と特徴行列 V^T の積の形に分解することで文書分類を行う．従来の NMF における目的関数では，クラスタリングの解に関係なく基底行列と特徴行列の値を更新させることで，それらの積を索引語文書行列に近づける方向に収束させる．その際，従来の NMF では分類結果の悪い方向への収束を防ぐことができない．そこで，本論文では，文書分類に適した初期値の提案と教師データに対する分類の成否を表す制約の提案よりなるべく良質なクラスタリング結果に収束するように制御することを目的とする．提案手法の効果を示すため，提案する改良 NMF と既存 NMF で実際に文書集合を分類し比較する．

1. はじめに

インターネットが普及して以来，多くの人が気軽にインターネットを利用して情報収集することができるようになった．また，文書群である記事もネット上に膨大な量が存在している．そのため，膨大な量の文書集合を効率よく分類する手法が必要である．

さらに，本来文書分類とは分類を行うユーザの観点により結果が異なるものである．そこでユーザの望む分け方を教師データとして与えることでユーザの望む分け方による良質な文書分類を目指す．

本論文では，最近文書分類で注目を浴びている非負値行列因子分解 (NMF)[2] を用いる．NMF は，文書クラスタリング手法の一つであり，高次元でスパースな文書行列をクラスタリングするのに適している．

この NMF は，索引語文書行列である入力行列 X を各クラスタに対する基底行列 U と特徴行列 V^T の積の形に分解することで索引語文書行列の次元圧縮を行うことができる．その次元圧縮結果である特徴行列 V が各文書と各クラスタとの関連度を表している．そのため，関連度が最大であるクラスタをその文書のクラスタと判断することで文書クラスタリングを行うことができる．

従来の NMF における目的関数は，基底行列 U と特徴行列 V の値を更新させることで，それらの積を入力行列 X に近づける方向に基底行列 U と特徴行列 V を収束させる．つまり，その従来の目的関数における収束方向は，ユーザの望む分け方を考慮してしておらず，単に入力行列 X に近づける方向と言える．そのため，その目的関数による収束方向が必ずしもユーザの望む分け方としての良いクラスタリング結果に収束する方向であるとは言えない．さらに，従来の NMF における目的関数から求まる収束値は基底行列 U と特徴行列 V の初期値に依存するという問題も存在している．

そこで，初期値に依らずになるべくユーザの望む分け方としての良い文書分類結果へ収束するような方向に収束方向を制御することが目的となる．その目的のために二つの手法を提案する．

提案手法の一つ目が，NMF における基底行列 U の初期値を教師データから求めて NMF を行う方法である．この方法を NMF-I(NMF with Initial basis value by training data) と呼ぶことにする．二つ目が，教師データに対する分類の成否を考慮し成功する方向へと収束方向を制御するための教師制約を追加した方法である．この方法を NMF-S(NMF with Supervised constraint) と呼ぶことにする．

実験では，実際に文書データを分類することで，この二つの提案手法と既存の NMF との比較を行う．

2. NMF

NMF は式 (1) のように m 個の文書データと n 個の索引語から作られる $n \times m$ の索引語文書行列 X を $n \times k$ の基底行列 U と $k \times m$ の特徴行列 V^T の積の形に分解することにより文書データを次元圧縮することができる．ここで，k はクラスタ数である．

$$X = UV^T \tag{1}$$

つまり，n 次元の文書データである索引語文書行列 X が k 次元の文書データである特徴行列 V^T へと次元圧縮され

[1] 九州工業大学

る．NMF を文書クラスタリングへ適用する際には次元圧縮後行列である特徴行列 V^T を利用する．特徴行列 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．そのため，i 番目の文書データのクラスタは (2) 式で得られる．

$$\arg\max_h \ v_{ih} \tag{2}$$

基底行列 U と特徴行列 V への分解は NMF の目的関数である式 (3) の J を最小にするような基底行列 U と特徴行列 V を推定することで求まる．

$$J = ||X - UV^T||^2 \tag{3}$$

そして，ラグランジュの未定乗数法を用いて式 (3) の J を最小にする基底行列 U と特徴行列 V の乗算型更新式を求める．r を反復更新の更新回数として式 (4)・(5) のように表される．

$$v_{ij}^{(r+1)} \leftarrow v_{ij}^{(r)} \frac{(X^TU)_{ij}}{(VU^TU)_{ij}} \tag{4}$$

$$u_{ij}^{(r+1)} \leftarrow u_{ij}^{(r)} \frac{(XV)_{ij}}{(UV^TV)_{ij}} \tag{5}$$

ここで，$u_{ij}^{(r)}$ と $v_{ij}^{(r)}$ はそれぞれ更新回数 r 回目である U と V の i 行 j 列の要素を表し，$(X)_{ij}$ は行列 X の i 行 j 列の要素を表す．

また，各繰り返し後には発散を防ぐためと各基底を単位ベクトルにするために基底行列 U を以下の式 (6) に従い正規化を行う．

$$u_{ij} \leftarrow \frac{u_{ij}}{\sqrt{\sum_i u_{ij}^2}} \tag{6}$$

通常，基底行列 U と特徴行列 V の初期値 $U^{(0)}$ と $V^{(0)}$ はランダムな値を与えることで作成される．

2.1 NMF の文書分類に関する問題点

2.1.1 初期値に関する問題点

NMF では，$U^{(0)}$ と $V^{(0)}$ の値によって，最終的に得られる $U^{(R)}$ と $V^{(R)}$ は大きく異なる．ここで R は最大更新回数とする．つまり，$V^{(R)}$ はクラスタリング結果を表しているため，クラスタリング結果は初期値 $U^{(0)}$ と $V^{(0)}$ に依存していると言える．

ここで，新納等が提案しているピンポン型クラスタリング [4][5] の結果から，初期値 $U^{(0)}$ により精度の高い初期値を与えることで，より精度の高いクラスタリング結果を得ることができると言える．そこで，本論文ではユーザが望む分け方に文書分類する際の初期値を考える必要がある．

2.1.2 収束方向に関する問題点

NMF では，式 (4)，(5) による繰り返しでは式 (3) の J に対する最適解にしか収束しない．また，NMF による文書クラスタリング [2] はある程度の成果が確認できるが，その繰り返しによる収束方向が必ずしもユーザが望む分け方としての良いクラスタリング結果に収束する方向であるとは言えない．なぜなら，式 (4)，(5) の収束方向が式 (3) における J を最小にするという方向であるのに対し，初期値が異なるクラスタリング結果を比較すると J の小さい方が必ずしも良いクラスタリング結果であると判断できないからである．

そのため，本論文では，既存 NMF の目的関数における収束方向を改良することで，繰り返しによるクラスタリング結果がなるべくユーザの望む分け方としての最適なクラスタリング結果に近づくように収束方向を制御する必要がある．

3. 提案手法

本論文で提案する手法ではユーザの望む分け方を表す教師データを利用する．しかし，実際にユーザの望む分け方で分類する際には得られる教師データは少ないと予想できる．そのため，なるべく少ない教師データの利用による手法でより精度の高い分類を目指す．本論文で提案する二つの新たな手法を以下で説明する．

NMF-I は 2.1.1 節の問題に対する手法であり，教師データから初期値を求めた手法である．

NMF-S は 2.1.2 節の問題に対する手法であり，教師データで収束方向を制御した手法である．

3.1 NMF-I

NMF では 2.1.1 節で挙げたように収束結果が初期値 $U^{(0)}$ と $V^{(0)}$ に依存するという問題が存在する．一般的には NMF での初期値は乱数で与えるが，単純な乱数ではクラスタリング結果が悪い局所解に収束するような初期値となる可能性がある．

そこで，既知である教師文書ベクトルの各クラスタにおける平均ベクトルを求め，それを平均教師ベクトルとする．その平均教師ベクトルを各クラスタ毎に並べ基底とした教師基底行列 U_s を考えた場合，NMF での理想的な文書分類がなされた場合の基底行列の収束値は U_s に近いものであろうとの期待に基づき，この教師基底行列 U_s を教師あり NMF における基底行列 U の初期値とする手法を提案する．教師基底行列 U_s は式 (7) で与える．

$$U_s = X_{train}(V_{train}^T)^+ \tag{7}$$

ここで，教師データ数を t とした時，X_{train}(n 行 t 列) は教師データのみの索引語行列であり，V_{train}^T(k 行 t 列) は各文書の正解クラスタに対応する要素を 1 としてそれ以外の要素を 0 とした行列である．また，“+” は疑似逆行列である．

教師基底行列 U_s を U の初期値 $U^{(0)}$ として，以降は既存

NMF と同様の更新式を実行する手法を NMF-I(NMF with Initial basis value by traing data) と呼ぶことにする．

3.2 NMF-S

2.1.2 節で挙げた収束方向に関する問題に対処するため，本論文では収束方向を既知である教師データに対する分類が成功する方向へと制御する手法を提案する．

NMF では特徴行列 V がクラスタリング結果を表している．そのため，教師データが全て正しいクラスタへと収束しているクラスタリング結果を表す教師特徴行列 V_s を考えた場合，クラスタリング結果である V を V_s に近づくような方向が，教師データに対する分類が成功する方向であると考えられる．そこで，クラスタリング結果を表す V を教師特徴行列 V_s に近づけるような教師制約を追加する．

この手法を NMF-S(NMF with Supervised constraint) と呼ぶことにする．NMF-S では式 (8) を目的関数とする．

$$J_s = ||X - UV^T||^2 + \mu||L * (V_s - V)||^2 \tag{8}$$

ここで，$(V_s)_{ij}$ は既知データであり正解クラスタならば 1，既知データであり正解でないクラスタか未知データならば 0 とする．そして行列 $L(m$ 行 k 列) は既知データならば 1，未知データならば 0 とする．また “$*$” は要素毎の乗算とし，μ は教師制約項に対する重みである．

そして，ラグランジュの未定乗数法を用いて式 (8) の J_s を最小にする基底行列 U と特徴行列 V の新たな乗算型更新式を求める．U の更新式は既存の式 (5) と同じで，V の更新式は式 (9) となる．

$$v_{ij}^{(r+1)} \leftarrow v_{ij}^{(r)} \frac{\{X^T U + \mu(L * V_s)\}_{ij}}{\{VU^T U + \mu(L * V)\}_{ij}} \tag{9}$$

さらに，NMF-S では更新毎に特徴行列 V を以下の式 (10) に従い正規化を行う．これは，特徴行列 V の各要素である文書と各クラスタの関連度に上限を付け効率的に教師データを正しいクラスタへ収束させるためである．また，特徴行列 V を正規化するため発散防止としても働くと考えられる．そのため発散防止として行っていた基底行列 U の正規化を NMF-S では行わない．

$$v_{ij} \leftarrow \frac{v_{ij}}{\sqrt{\sum_j v_{ij}^2}} \tag{10}$$

ただし，基底行列 U の正規化には各基底を単位ベクトルにする効果もあるため NMF-S において基底行列 U を式 (6) に従い正規化した手法を正規化 NMF-S と呼ぶことにする．

NMF-S の教師制約により既知である教師データが正しいクラスタへと収束する方向へ収束方向を制御されると，教師データが正しいクラスタへ収束するのに影響されて間接的に未知データも正しいクラスタへ収束するのではないかと期待される．この期待は以下の理由に基づくものである．NMF において基底行列 U と特徴行列 V の更新式はお互いに影響している．そのため，V の教師部分などの一部分が理想的な値に近づけば各クラスタの基底ベクトルである基底行列 U も理想的な値に近づくと考えられる．さらに各クラスタの基底ベクトルが理想的な値に近づくと未知データに関してもよりよいクラスタリング結果に収束すると考えられる．以上が先ほどの期待に対する理由である．

3.3 NMF-IS

さらに，3.1 節と 3.2 節で提案した NMF-I と NMF-S を同時に適用した手法を NMF-IS とする．同様に NMF-I と正規化 NMF-S を同時に適用した手法を正規化 NMF-IS とする．

4. 関連研究

本論文で提案する手法の有効性を示すために類似手法と比較する．その類似手法の一つである SSNMF(Semi-Supervised Nonnegative Matrix Factorization)[3] を以下で簡単に説明する．

4.1 SSNMF

SSNMF は H.Lee 等が提案した半教師あり NMF の一つである．この SSNMF の目的は 2.1.2 節で挙げた収束方向に関する問題の解決である．そのため，既存 NMF の目的関数に対して，教師情報を含んだ我々とは異なる制約項を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 (11) に示す．

$$J_s = ||X - UV^T||^2 + \lambda||L * (Y - WV^T)||^2 \tag{11}$$

式 (11) において Y は教師データの正解クラスタがラベル付けされた $k \times m$ のラベル行列であり，教師データの正解クラスタを 1 としそれ以外を 0 としている．W は $k \times k$ の第二項目における基底行列でり，クラスタ間の関係を表している．そして，L は式 (12) のように教師データのみを制御するための $k \times m$ の重み行列である．また，λ は第二項目に対する重みである．

$$L_{ij} = \begin{cases} 0.001 & \text{if } \mathrm{Y_{ij}} = 1 \\ 1 & \text{if } \mathrm{Y_{ij}} = 0 \\ 0 & \text{if } \mathrm{Y_{ij}} \text{ is unknown.} \end{cases} \tag{12}$$

式 (11) の第二項目が SSNMF における制約である．この制約により，基底行列 W と特徴行列 V^T の積がラベル行列 Y に近づく方向へ特徴行列 V の収束方向は制御される．つまり，特徴行列 V における文書とクラスタの関連度に対して似ているクラスタ同士の関連度を上げる効果があると考えられる．

さらに，SSNMF では最終更新後の $V^{(R)}$ に対して K-means を適用している．その結果をクラスタリング結果と

している，

提案手法である NMF-S と SSNMF の相違点は二つ存在する．一つ目は SSNMF の制約では V は教師データにおいて正解クラスタと正解クラスタに類似するクラスタの両方の関連度を上げるが，我々の提案手法である NMF-S では正解クラスタのみの関連度を上げる．二つ目は SSNMF における基底行列 W は既存の NMF における基底行列 U や特徴行列 V と同じように乗算型更新式で求める必要があり，初期値 $W^{(0)}$ はランダムな値を与える．そのため 2.1.1 節で挙げた初期値に関する問題点に関して更に初期値に対する依存度を高めてしまう恐れがある．

5. 実験

実際に文書集合を既存 NMF，NMF-I，NMF-S，正規化 NMF-S，NMF-IS，正規化 NMF-IS と SSNMF で分類し比較することで，提案手法の有効性を示す．さらに，NMF ではない他の分類手法である K-means，SVM とナイーブベイズ (NB) による分類結果との比較も行う．SVM は Gauss カーネル，多項式カーネル，シグモイドカーネル，線形カーネルによる事前実験を行った．その結果，線形カーネルが良好であったため比較する SVM には線形カーネルを用いる．

5.1 実験用文書データセット

実験には CLUTO のサイト[*1] で公開されている文書集合を利用する．各文書データセットの詳細は表 1 に示す．データセットの k1a,k1b と wap は Yahoo!内の様々な web ページから構成され，re0 はロイターのニュースワイヤーから取得したニュース記事で構成されている．

表 1 文書データセット

データ名	文書数	索引語数	クラス数
k1a	2340	21839	20
k1b	2340	21839	6
re0	1504	2886	13
wap	1560	6460	20

5.2 評価方法

分類結果の評価値には Entropy，Purity，RandIndex，Precision，Recall 及び F 値 (Precision と Recall の調和平均）を用いる．さらに最終的な分類性能値はこれらの F 値を除く五種類の評価値の調和平均 Hm により評価する．

Entropy は式 (13) より求める．Entropy は各クラスタにおける正解集合の分布度合を表しており，小さな値ほどクラスタリング結果が良好であることを意味している．ここで N は総文書数を示す．また，調和平均 Hm を算出する際には $(1 - Entropy)$ として計算する．

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

$$Entropy = \sum_{i=1}^{k} \frac{|C_i|}{N} \times (-\sum_{h=1}^{k} P(A_h|C_i) \log P(A_h|C_i)) \quad (13)$$

Purity は結果クラスタに一番多く含まれている正解クラスタを用いて，結果クラスタに正解データが含まれている割合を示す指標である．クラスタリング結果の Purity は，各クラスタのデータ数による重み付き平均をとるように定義し，式 (14) に示す．

$$Purity = \frac{1}{N} \sum_{i=1}^{k} \max_{h} |C_i \cap A_h| \quad (14)$$

式 (13)，式 (14) において C_i はクラスタリング結果に対する i 番目の文書のクラスタであり，A_h は正解データに対する h 番目の文書のクラスタである．$A_h \cap C_i$ は正解データであるクラスタ A_h とクラスタリング結果のクラスタ C_i が共通している文書数である．

RandIndex はデータの各ペア同士の正解が同じクラスタならば同じクラスタになるかどうかの判定の正解率を表し式 (15) で求める．

$$RandIndex = \frac{TP + TN}{TP + FP + FN + TN} \quad (15)$$

ここで TP は同じ正解クラスタであるデータのペアが結果クラスタで同じクラスタである対の数，TN は異なる正解クラスタであるデータのペアが結果クラスタで異なるクラスタである対の数，FP は異なる正解クラスタであるデータのペアが結果クラスタで同じクラスタである対の数，FN は同じ正解クラスタであるデータのペアが結果クラスタで異なるクラスタである対の数を表す．

Precision はクラスタリング結果の中にどの程度正解が含まれているを表す．実際には以下の式 (16) で求める．

$$Precision = \frac{TP}{TP + FP} \quad (16)$$

Recall は正解データがどの程度結果クラスタで正しくクラスタリングされているかを表す．実際には以下の式 (17) で求める．

$$Recall = \frac{TP}{TP + FN} \quad (17)$$

5.3 実験方法及び結果

実験では，20 種類の異なる初期値 $U^{(0)}$ と $V^{(0)}$ を準備する．それらの初期値に対して各手法で文書分類を行い平均評価値を調査した．実験における μ は 1 とする．

NMF における更新回数は 50 回とする．各文書データセットの文書数に対して 1 %と 10 %を教師データとして使用する．教師データの内訳は各クラスタの 1 %と 10 %とし，各クラスタのデータを最低一つは含むこととした．教師データ使用率 1 %における分類結果を評価したものを表 2〜8 に示し，教師データ使用率 10 %の結果を表 9〜15 に示す．SSNMF の結果に関しては論文 [3] を参考に自作したプログラムによる結果である．

表 2 教師データ 1 %-Entropy

分類手法	k1a	k1b	re0	wap	平均
NMF	0.402	0.254	0.407	0.416	0.370
NMF-I	0.352	0.151	0.393	0.350	0.312
NMF-S	0.379	0.199	0.415	0.372	0.341
正規化 NMF-S	0.382	0.236	0.405	0.373	0.349
NMF-IS	0.355	**0.127**	0.389	0.344	0.304
正規化 NMF-IS	**0.324**	0.181	**0.380**	**0.320**	**0.301**
SSNMF	0.384	0.248	0.405	0.371	0.352
K-means	0.429	0.250	0.407	0.417	0.376
SVM	0.848	0.673	0.694	0.833	0.762
NB	0.771	0.624	0.690	0.776	0.715

表 3 教師データ 1 %-Purity

分類手法	k1a	k1b	re0	wap	平均
NMF	0.614	0.819	0.639	0.605	0.669
NMF-I	0.661	0.892	0.636	0.665	0.714
NMF-S	0.636	0.846	0.623	0.643	0.687
正規化 NMF-S	0.636	0.823	0.634	0.645	0.685
NMF-IS	0.653	**0.909**	0.632	0.668	0.716
正規化 NMF-IS	**0.686**	0.851	**0.651**	**0.691**	**0.720**
SSNMF	0.631	0.835	0.641	0.644	0.688
K-means	0.579	0.832	0.611	0.587	0.652
SVM	0.229	0.600	0.420	0.246	0.374
NB	0.327	0.609	0.430	0.330	0.424

表 4 教師データ 1 %-RandIndex

分類手法	k1a	k1b	re0	wap	平均
NMF	0.907	0.712	0.763	0.900	0.821
NMF-I	0.937	0.863	0.773	**0.935**	**0.877**
NMF-S	0.890	0.759	0.756	0.890	0.824
正規化 NMF-S	0.916	0.713	0.763	0.903	0.824
NMF-IS	0.920	**0.895**	0.766	0.916	0.874
正規化 NMF-IS	**0.942**	0.802	**0.775**	0.934	0.863
SSNMF	0.893	0.738	0.764	0.888	0.821
K-means	0.847	0.755	0.749	0.872	0.806
SVM	0.131	0.415	0.266	0.152	0.241
NB	0.668	0.555	0.533	0.632	0.597

表 5 教師データ 1 %-Precision

分類手法	k1a	k1b	re0	wap	平均
NMF	0.530	0.799	0.501	0.503	0.583
NMF-I	0.702	0.949	0.549	0.712	0.728
NMF-S	0.429	0.874	0.468	0.443	0.554
正規化 NMF-S	0.584	0.813	0.498	0.525	0.605
NMF-IS	0.597	**0.962**	0.515	0.598	0.668
正規化 NMF-IS	**0.748**	0.917	**0.558**	**0.721**	**0.736**
SSNMF	0.445	0.771	0.504	0.438	0.540
K-means	0.305	0.804	0.441	0.364	0.479
SVM	0.099	0.410	0.241	0.103	0.213
NB	0.178	0.462	0.252	0.173	0.266

表 6 教師データ 1 %-Recall

分類手法	k1a	k1b	re0	wap	平均
NMF	0.400	0.388	0.216	0.336	0.335
NMF-I	0.620	0.701	0.235	0.590	0.537
NMF-S	0.369	0.475	0.202	0.334	0.345
正規化 NMF-S	0.445	0.382	0.185	0.343	0.339
NMF-IS	0.563	0.770	0.246	0.498	0.519
正規化 NMF-IS	0.617	0.564	0.235	0.564	0.495
SSNMF	0.418	0.509	0.218	0.371	0.379
K-means	0.375	0.526	0.209	0.351	0.365
SVM	**0.974**	**0.995**	**0.979**	**0.962**	**0.978**
NB	0.656	0.588	0.493	0.690	0.607

表 7 教師データ 1 %-F 値

分類手法	k1a	k1b	re0	wap	平均
NMF	0.455	0.521	0.302	0.402	0.420
NMF-I	0.658	**0.806**	0.329	**0.645**	**0.610**
NMF-S	0.396	0.612	0.282	0.380	0.418
正規化 NMF-S	0.503	0.520	0.270	0.414	0.427
NMF-IS	0.579	0.853	0.332	0.543	0.577
正規化 NMF-IS	**0.676**	0.698	0.330	0.633	0.584
SSNMF	0.430	0.609	0.304	0.400	0.436
K-means	0.331	0.629	0.283	0.354	0.399
SVM	0.179	0.581	**0.387**	0.186	0.333
NB	0.279	0.517	0.333	0.276	0.351

表 8 教師データ 1 %-調和平均 Hm

分類手法	k1a	k1b	re0	wap	平均
NMF	0.569	0.641	0.447	0.529	0.547
NMF-I	0.698	0.842	0.472	0.693	**0.676**
NMF-S	0.535	0.714	0.427	0.525	0.550
正規化 NMF-S	0.607	0.643	0.418	0.551	0.555
NMF-IS	0.656	**0.877**	0.474	0.642	0.662
正規化 NMF-IS	**0.719**	0.768	**0.477**	**0.700**	0.666
SSNMF	0.557	0.700	0.450	0.540	0.562
K-means	0.472	0.714	0.427	0.492	0.526
SVM	0.168	0.473	0.343	0.183	0.292
NB	0.311	0.501	0.373	0.307	0.373

6. 考察

6.1 教師データ 1 %についての考察

まずは既存 NMF と各提案手法を比較する．表 2〜8 から Entropy,Purity,RandIndex,Recall と Hm においては既存 NMF より各提案手法の方がそれぞれ良い結果であることが分かる．しかし，Precision と F 値では NMF-I，正規化 NMF-S，NMF-IS と正規化 NMF-IS が既存 NMF よりも良い評価値であることが分かる．

次に SSNMF と各提案手法を比較する．NMF-I, NMF-IS と正規化 NMF-IS はどの評価値においても SSNMF より良い評価値であることが分かる．そして，NMF-S と正規化 NMF-S は Entropy と RandIndex において SSNMF より良い評価値であるが，その他の評価値では SSNMF の方

表 9 教師データ 10 %-Entropy

分類手法	k1a	k1b	re0	wap	平均
NMF	0.402	0.254	0.407	0.416	0.370
NMF-I	0.313	**0.150**	0.374	0.307	0.286
NMF-S	0.344	0.152	0.385	0.341	0.306
正規化 NMF-S	0.301	0.177	0.384	0.306	0.292
NMF-IS	0.314	0.098	0.372	0.304	0.272
正規化 NMF-IS	**0.272**	0.168	**0.360**	**0.281**	**0.270**
SSNMF	0.374	0.246	0.391	0.367	0.345
K-means	0.429	0.250	0.407	0.417	0.376
SVM	0.484	0.336	0.448	0.521	0.447
NB	0.565	0.447	0.654	0.592	0.565

表 10 教師データ 10 %-Purity

分類手法	k1a	k1b	re0	wap	平均
NMF	0.614	0.819	0.639	0.605	0.669
NMF-I	0.710	0.893	0.623	0.717	0.736
NMF-S	0.674	0.894	0.662	0.681	0.728
正規化 NMF-S	0.725	0.858	0.657	0.716	0.739
NMF-IS	0.693	**0.935**	0.642	0.709	0.745
正規化 NMF-IS	**0.744**	0.861	0.655	**0.737**	**0.749**
SSNMF	0.644	0.837	0.659	0.649	0.697
K-means	0.579	0.832	0.611	0.587	0.652
SVM	0.623	0.853	**0.710**	0.587	0.693
NB	0.496	0.719	0.440	0.477	0.533

表 11 教師データ 10 %-RandIndex

分類手法	k1a	k1b	re0	wap	平均
NMF	0.907	0.712	0.763	0.900	0.821
NMF-I	0.943	0.861	0.769	0.943	**0.879**
NMF-S	0.915	0.830	0.774	0.912	0.858
正規化 NMF-S	0.950	0.785	0.775	0.942	0.863
NMF-IS	0.932	**0.937**	0.772	0.933	0.894
正規化 NMF-IS	**0.952**	0.827	**0.780**	**0.945**	0.876
SSNMF	0.894	0.746	0.769	0.887	0.824
K-means	0.847	0.755	0.749	0.872	0.806
SVM	0.853	0.803	0.720	0.806	0.796
NB	0.895	0.630	0.640	0.879	0.761

が良い評価値である．

また，他の分類手法 (K-means，SVM，NB) と各提案手法を比較すると Recall 以外の評価値において各提案手法の方がよい結果であることが分かる．SVM を始めとする Recall の評価値が高い他の分類手法はクラスタリング結果が一つのクラスタに集中してしまったため Recall が高い値となった．そのため，一つのクラスタに集中した手法は Precision がとても低い値となり F 値の値も低い値である．

そして，教師データ使用率 1 %の調和平均 Hm において NMF を用いた手法の中では NMF-I が最も良い分類性能を示した．

6.2 教師データ 10 %についての考察

まずは既存 NMF と各提案手法を比較する．表 9〜15 から全ての評価値において既存 NMF より各提案手法の方がそれぞれ良い評価値であることが分かる．このことから，教師データを 10 %使用した場合は各提案手法は既存 NMF

表 12 教師データ 10 %-Precision

分類手法	k1a	k1b	re0	wap	平均
NMF	0.530	0.799	0.501	0.503	0.583
NMF-I	0.738	0.954	0.526	0.757	0.744
NMF-S	0.578	0.929	0.557	0.585	0.662
正規化 NMF-S	0.764	0.906	0.565	0.755	0.748
NMF-IS	0.666	**0.980**	0.539	0.698	0.721
正規化 NMF-IS	**0.806**	0.937	**0.579**	**0.783**	**0.776**
SSNMF	0.449	0.778	0.527	0.435	0.547
K-means	0.305	0.804	0.441	0.364	0.479
SVM	0.347	0.681	0.451	0.316	0.449
NB	0.459	0.596	0.265	0.414	0.434

表 13 教師データ 10 %-Recall

分類手法	k1a	k1b	re0	wap	平均
NMF	0.400	0.388	0.216	0.336	0.335
NMF-I	0.653	0.692	0.252	0.642	0.560
NMF-S	0.471	0.626	0.225	0.435	0.439
正規化 NMF-S	0.700	0.523	0.213	0.634	0.518
NMF-IS	0.606	0.864	0.283	0.584	0.584
正規化 NMF-IS	0.668	0.615	0.270	0.631	0.546
SSNMF	0.398	0.524	0.226	0.365	0.378
K-means	0.375	0.526	0.209	0.351	0.365
SVM	**0.777**	**0.972**	**0.837**	**0.778**	**0.841**
NB	0.445	0.283	0.293	0.464	0.371

表 14 教師データ 10 %-F 値

分類手法	k1a	k1b	re0	wap	平均
NMF	0.455	0.521	0.302	0.402	0.420
NMF-I	0.693	0.802	0.341	0.694	0.633
NMF-S	0.518	0.744	0.321	0.498	0.520
正規化 NMF-S	**0.730**	0.661	0.309	0.688	0.597
NMF-IS	0.635	**0.918**	0.370	0.635	**0.640**
正規化 NMF-IS	**0.730**	0.742	0.368	**0.699**	0.635
SSNMF	0.421	0.620	0.316	0.396	0.436
K-means	0.331	0.629	0.283	0.354	0.399
SVM	0.480	0.801	**0.586**	0.449	0.579
NB	0.451	0.384	0.276	0.437	0.387

表 15 教師データ 10 %-調和平均 Hm

分類手法	k1a	k1b	re0	wap	平均
NMF	0.569	0.641	0.447	0.529	0.547
NMF-I	0.734	0.840	0.482	0.738	0.699
NMF-S	0.629	0.810	0.469	0.618	0.631
正規化 NMF-S	0.757	0.749	0.459	0.735	0.675
NMF-IS	0.702	**0.922**	0.509	0.708	**0.710**
正規化 NMF-IS	**0.769**	0.798	0.510	**0.750**	0.707
SSNMF	0.555	0.709	0.463	0.539	0.567
K-means	0.472	0.714	0.427	0.492	0.526
SVM	0.561	0.779	**0.624**	0.527	0.623
NB	0.507	0.500	0.360	0.487	0.464

よりも良い分類性能を示した．

次に SNMF と各提案手法を比較する．全ての評価値において SSNMF より各提案手法の方がそれぞれ良い評価値であることが分かる．このことから，教師データを 10 %使用した場合は各提案手法は既存 NMF よりも良い分類性能を示した．特に教師データ 1 %の時は SSNMF は NMF-S と正規化 NMF-S より良い分類性能を示したが，10 %の時は NMF-S と正規化 NMF-S の方が良い分類性能を示した．

また，他の分類手法 (K-means，SVM，NB) と各提案手法を比較すると Entropy，Puirty，RandIndex，Precision と Hm においては他の分類手法より各提案手法の方が平均評価値が良いことが分かる．しかし，Puirty，F 値と Hm においてデータセット re0 は SVM が良い評価値であることが分かる．これは，データセット re0 が他のデータセットと比べると特徴量が少ないため現在の特徴空間では他のデータセットより分類し辛い可能性が考えられる．そのため，カーネルを利用した SVM の方が良い分類性能を示したと考えられる．

そして，教師データ使用率 10 %の調和平均 Hm において NMF を用いた手法の中では NMF-IS が最も良い分類性能を示した．

6.3 教師データ 1 %と 10 %の結果を比較

教師データ使用率を 1 %とした実験の結果は NMF-I が良い分類性能を示したが，教師データ使用率を 10 %とした実験の結果は NMF-IS が良い分類性能を示した．つまり，教師データ使用率 1 %では，各クラスタの特徴をうまく捉えきれず期待したほど NMF-I より NMF-IS は向上しなかったが，教師データ使用率を 10 %ほどにすると，各クラスタの特徴をうまく捉えることができ教師データ以外の未知データの正解率を上げることができたと考えられる．

6.4 更新回数に対する評価値の推移

更新回数に対する評価値の推移に関して教師データ使用率を 10 %とした実験における文書データセット k1a の結果に着目して考察する．図 1〜6 が更新回数に対するそれぞれの評価値の推移図である．

図 1〜6 を見るとどの評価値においても NMF，NMM-S は更新回数 35 回目位で収束している．正規化 NMF-S は RandIndex と Precision において更新回数 15 回目まで急激に上がり一度収束の兆しを見せるがその後再度上昇してから収束している．さらに，NMF-I の教師基底行列 U_s を使用している NMF-I，NMF-IS と正規化 NMF-IS は早い段階で高い評価値に推移している．しかし，NMF-IS はその後少し降下し収束している．

6.5 NMF-S の正規化問題

NMF-IS と正規化 NMF-IS を調和平均 Hm において比較する．文書データセットの k1a と wap では教師データ使用率 1 %と教師データ使用率 10 %の両方で正規化 NMF-IS の方が高い値であることが確認できる．しかし，k1b ではその両方で NMF-IS の方が高い値であることが確認できる．re0 は NMF-IS と正規化 NMF-IS の値はほぼ同値である．以上のように正規化 NMF-IS の方が良い分類性能を示した文書データセットは多いが，調和平均 Hm において NMF-IS は NMF-I より高く正規化 NMF-IS は NMF-I より低いという場合があるため一概に正規化 NMF-IS の方が NMF-IS より良い分類性能であるとは言えない．このように正規化を行うか無視するかにより教師制約の有効性が変動するという問題が確認された．

6.6 SVM との教師データ使用率比較

教師データ使用率 1 %における各提案手法 (特に NMF-IS や正規化 NMF-IS) と同等の分類性能を得るには SVM がどの程度教師データを必要とするか比較することで提案手法の特徴を分析する．

表 16 各教師データ使用率による SVM の Hm

Dataset	1 %	10 %	20 %	30 %	40 %
k1a	0.168	0.561	0.749	0.816	0.862
k1b	0.473	0.779	0.867	0.908	0.934
re0	0.343	0.624	0.740	0.799	0.839
wap	0.183	0.527	0.696	0.783	0.835

表 8 と表 16 を比較すると教師データ使用率 1 %の提案手法と同等の分類性能を SVM で得るには約 20 %の教師データが必要であると考えられる．

7. おわりに

本論文では，ユーザの望む分け方による良質な文書分類を目指し，そのためのいくつかの改良 NMF を示した．

まず，NMF-I の教師基底行列 U_s を初期値として使用することを提案した．それにより，教師あり NMF における文書分類に対して少ない更新回数で高い分類性能が得られることを示した．

次に，NMF-S の教師制約を NMF の目的関数に追加し収束方向をユーザの望む分け方による最適なクラスタリング結果への方向へ制御することを提案した．教師データ使用率を 10 %程度に上げると NMF-I 単体よりもこの教師制約を併用した方が高い分類性能が得られることを示した．しかし，NMF-S における基底行列 U の正規化を行う場合と，行わない場合のそれぞれの手法が得意とする文書データがあるためさらに多くの文書データセットに対する調査が必要である．

さらに，教師有り NMF の一つである SSNMF と提案手法との比較を行った．その結果，NMF-I，NMF-IS 及び正規化 NMF-IS の方が SSNMF よりも教師あり NMF での

文書分類に対して効果があるとことを示した．また教師データ 1 %では SSNMF の方が NMF-S と正規化 NMF-S より良い分類性能を示したが，10 %では NMF-S と正規化 NMF-S の方が良い分類性能を示した．

また，教師データ使用率 10 %におけるデータセット re0 の結果のように現在の制約では分類し辛い文書データセットが確認された．そのため，別の制約による制御も検討する必要がある．

参考文献

[1] D.D.Lee , H.S.Seung :“Algorithms for Non-negative Matrix Factorization”, NIPS , pp.556-562 , (2000).

[2] W.Xu, X.Liu, and Y.Gong, “Document clustering based on non-negative matrix factorization”, in Proc.ACM SIGIR Conf.Research and Development in Information Retrieval (SIGIR), Toronto, ON, Candada, 2003.

[3] H.Lee , J.Yoo , S.Choi “Semi-Supervised Nonnegative Matrix Factorization”, IEEE SIGNAL PROCESSING LETTERS , Vol.17 No.1 , pp.4-7 , JANUARY 2010.

[4] 新納浩幸，佐々木稔，“NMF とリンクベースの修正法によるピンポン型文書クラスタリング”，情報処理学会，自然言語処理研究会報告，Vol.2007,no.47,p.7-12.

[5] 新納浩幸，佐々木稔，“Mcut + NMF による文書クラスタリング”，言語処理学会年次大会発表論文集，Vol.13，pp,558-561,(2007).

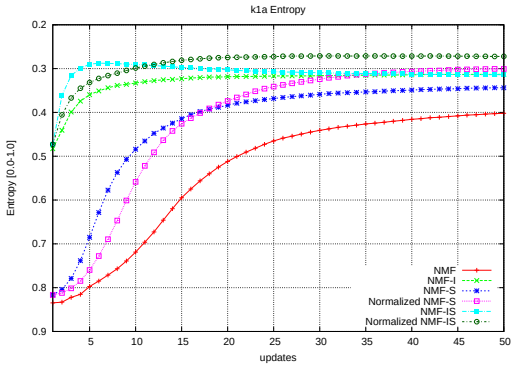

図 1 Entropy の推移

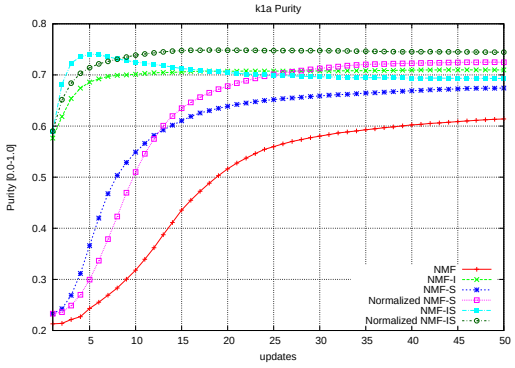

図 2 Purity の推移

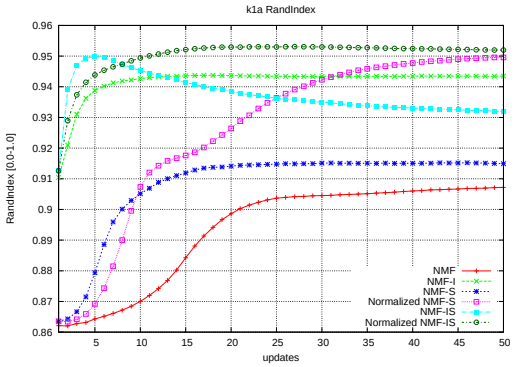

図 3 RandIndex の推移

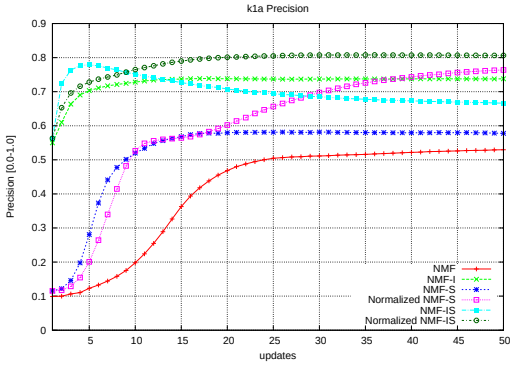

図 4 Precision の推移

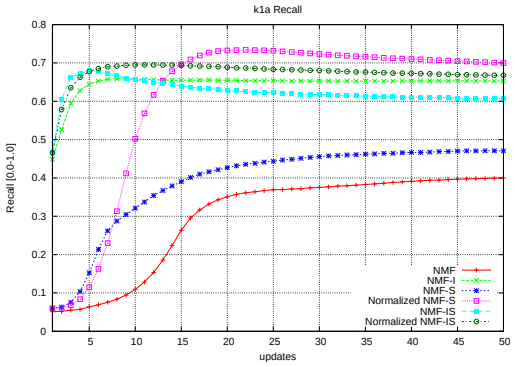

図 5 Recall の推移

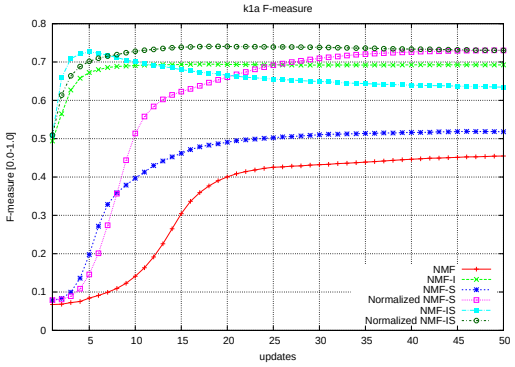

図 6 F 値の推移