NMF による重み付きハイパーグラフを用いたアンサンブル文書クラスタリング

新納浩幸[†]・佐々木稔[††]

本論文では Non-negative Matrix Factorization (NMF) を利用したアンサンブル文書クラスタリングを提案する．
NMF は次元縮約を利用したクラスタリング手法であり，文書クラスタリングのようにデータが高次元かつスパースとなる場合に効果を発揮する．ただし NMF は初期値によって得られるクラスタリング結果が異なるという問題がある．そのために通常は初期値を様々に変えて，複数個得られたクラスタリング結果から，NMF の分解の精度の最もよい結果を選択する．しかし NMF の分解の精度はクラスタリング結果の精度を直接表しているわけではないので，最適な選択が行える保証はない．
ここでは NMF によるクラスタリングの精度を高めるために，複数個得られたクラスタリング結果をアンサンブルすることを試みる．アンサンブルは，複数個のクラスタリング結果からハイパーグラフを作成し，そのハイパーグラフで表現されたデータをクラスタリングすることで行える．従来，そのハイパーグラフは 0 か 1 のバイナリ値が用いられていたが，ここでは NMF の結果を用いて，適切な実数値の重みを与えることで改良する．実験では k-means，NMF，通常のハイパーグラフを用いたアンサンブル手法および重み付きハイパーグラフを用いたアンサンブル手法（本手法）のクラスタリング結果を比較し，本手法の有効性を示す．

キーワード：アンサンブルクラスタリング，NMF，ハイパーグラフ，局所解，アンサンブル学習

Ensemble document clustering using weighted hypergraph generated by NMF

Hiroyuki Shinnou[†] and Minoru Sasaki[††]

In this paper, we propose a new ensemble clustering method using Non-negative Matrix Factorization (NMF).
NMF is a kind of the dimensional reduction method which is effective for high dimensional and sparse data like document data. NMF has the problem that the result depends on the initial value of the iteration. The standard countermeasure for this problem is that we generate multiple clustering results by changing the initial value, and then select the best clustering result estimated by the NMF decomposition error. However, this selection does not work well because the NMF decomposition error does not always measure the accuracy of the clustering.

[†]茨城大学工学部情報工学科, Department of Computer and Information Sciences, Ibaraki University
[††]茨城大学工学部情報工学科, Department of Computer and Information Sciences, Ibaraki University

To improve the clustering result of NMF, we propose a new ensemble clustering method. Our method generates multiple clustering results by using the random initialization of NMF. And they are integrated through the weighted hypergraph, which can directly be constructed through the result of NMF, instead of the traditional binary hypergraph.
In the experiment, we compared the k-means, NMF, the ensemble method using the standard hypergraph and the ensemble method using the weighted hypergraph (our method). Our method achieved best.

Key Words: ensemble clustering, NMF, hypergraph, local optimum solution, ensemble learning

1 はじめに

本論文では，ランダムな初期値を使って Non-negative Matrix Factorization (NMF) による文書クラスタリングを複数回行い，それらの結果をアンサンブルすることで，より精度[1]の高い文書クラスタリングの実現を目指す．複数のクラスタリング結果を統合する部分で，従来のハイパーグラフの代わりに重み付きハイパーグラフを用いることが特徴である．

文書クラスタリングは，文書の集合に対して，知的な処理を行う基本的な処理であり，その重要性は明らかである．例えばテキストマイニングの分野では，文書クラスタリングは基本的な構成要素であるし (Michael W. Berry 2003)，情報検索の分野では，検索結果の概観を視覚化するために検索された文書の集合をクラスタリングする研究が盛んに行われている (Hearst and Pedersen 1996)(Leuski 2001)(Zeng, He, Chen, Ma, and Ma 2001)(Kummamuru, Lotlikar, Roy, Singal, and Krishnapuram 2004)．

文書クラスタリングでは，まずデータとなる文書をベクトルで表現する．通常，bag of words のモデルを用い，次に TF-IDF などによって次元の重みを調整する．このようにして作成されたベクトルは高次元かつスパースになるために，文書クラスタリングではクラスタリング処理を行う前に主成分分析や特異値分解などの次元縮約の手法を用いることが行われる (Boley, Gini, Gross, Han, Hastings, Karypis, Kumar, Mobasher, and Moore 1999)(Deerwester, Dumais, Landauer, Furnas, and Harshman 1990)．次元縮約により高次元のベクトルが構造を保った状態で低次元で表現されるため，クラスタリング処理の速度や精度が向上する．

NMF は次元縮約の手法を応用したクラスタリング手法である (Xu, Liu, and Gong 2003)．今，クラスタリング対象の m 次元で表現された n 個の文書を m 行 n 列の索引語文書行列 X で表す．

[1]本論文において用いる「(クラスタリングの) 精度」とは，クラスタリングの正解率 (accuracy) と同義である．つまり，ここでは暗にクラスタリングの正解があることを想定しており，得られた結果がどの程度正解に近いかという尺度の意味で「(クラスタリングの) 精度」という用語を用いる．

目的とするクラスタの数が k である場合，NMF では X を以下のような行列 U と V^T に分解する．そして行列 V がクラスタリング結果に対応する．

$$X = UV^T$$

ここで U は m 行 k 列，V は n 行 k 列である．V^T は V の転置を表す．また U と V の要素は非負である．

与えられた X と k から，ある繰り返し処理により U と V を得ることができる (Lee and Seung 2000)．しかしこの繰り返し処理は局所最適解にしか収束しない．つまり NMF では，与える初期値によって得られるクラスタリング結果が異なるという問題がある．通常は適当な初期値を与える実験を複数回行い，それらから得た複数個のクラスタリング結果の中で X と UV^T の差[2]が最小のもの，つまり X の分解の精度が最も高いものを選ぶ．しかし分解の精度は，直接的にはクラスタリングの精度を意味してはいないため，最も精度の高いクラスタリング結果を選択できる保証がない．

ここでは NMF の分解の精度を用いて，複数個のクラスタリング結果から最終的なクラスタリング結果を選ぶのではなく，複数個のクラスタリング結果をアンサンブルさせて，より精度の高いクラスタリング結果を導くアンサンブルクラスタリングを試みる．

一般にアンサンブルクラスタリングの処理は 2 段階に分けられる．まず第 1 段で複数個のクラスタリング結果を生成し，次の第 2 段でそれらを組み合わせ，最終的なクラスタリング結果を導く．複数個のクラスタリング結果を生成する手法としては，k-means の初期値を変化させたり (Fred and Jain 2002)，ランダムプロジェクションにより利用する特徴を変化させたり (Fern and Brodley 2003)，“weak partition” を生成する研究などがある (Topchy, Jain, and Punch 2003)．また複数個のクラスタリング結果を組み合わせる手法としては，データ間の類似度を新たに構築する手法 (Fred and Jain 2002) や，データの表すベクトルを新たに構築する手法 (Strehl and Ghosh 2002) などがある．ここでは後者の手法を改良して用いる．

論文 (Strehl and Ghosh 2002) では，データの表すベクトルを新たに構築するために，複数個のクラスタリング結果から，データセットに対するハイパーグラフを作成する．このハイパーグラフは，データセットが表す行列に相当する．このハイパーグラフで表現されたデータに対してクラスタリングを行い，最終的なクラスタリング結果を得る．

ただしこのハイパーグラフではエッジの重みが 0 か 1 のバイナリ値である．ハイパーグラフが行列に相当すると考えると，エッジの重みの意味は同じクラスタに属する度合いとなり，バイナリ値で表すよりも非負の実数で表す方がより適切と考えられる．そこで本論文ではハイパーグラフのエッジの重みに非負の実数値を与える．具体的には，NMF のクラスタリング結果が

[2]差は $||X - UV^T||_F$ により測定する．

行列 V で得られ，同じクラスタに属する度合いが V から直接求められることを利用する．またここでは，この実数値の重みを付けたハイパーグラフを重み付きハイパーグラフと呼ぶことにする．

実験では k-means，NMF，通常のハイパーグラフを用いたアンサンブル手法および重み付きハイパーグラフを用いたアンサンブル手法（本手法）の各クラスタリング結果を比較し，本手法の有効性を示す．

2 NMF と初期値の問題

2.1 NMF とその特徴

NMF は $m \times n$ の索引語文書行列 X を，$m \times k$ の行列 U と $n \times k$ の行列 V の転置行列 V^T の積に分解する (Xu et al. 2003)．ただし k はクラスタ数である．

$$X = UV^T$$

NMF はクラスタに対応したトピックの次元を k 個想定し，その基底ベクトルの線形和によって，文書ベクトル及び索引語ベクトルを表現することに対応する．つまり基底ベクトルの係数が，そのトピックとの関連度を表しているので，行列 V 自体がクラスタリング結果と見なせる．具体的には，i 番目の文書 d_i は，行列 X の第 i 列のベクトルで表現され，その次元圧縮された結果が，行列 V の第 i 行のベクトルとなる．このとき，V の第 i 行のベクトルは

$$(v_{i1}, v_{i2}, \cdots, v_{ik})$$

と表せ，文書 d_i のクラスタの番号は

$$\arg\max_{j \in 1:k} v_{ij}$$

となる．

2.2 NMF のアルゴリズム

与えられた索引語文書行列 X から，U と V は以下の繰り返しで得ることができる (Lee and Seung 2000)．

$$u'_{ij} \leftarrow u_{ij} \frac{(XV)_{ij}}{(UV^TV)_{ij}} \tag{1}$$

$$v'_{ij} \leftarrow v_{ij} \frac{(X^T U)_{ij}}{(V U^T U)_{ij}} \tag{2}$$

ここで u_{ij} と v_{ij} はそれぞれ U と V の i 行 j 列の要素を表す．また $(X)_{ij}$ により行列 X の i 行 j 列の要素を表す．上記の式により，現在の U と V から，u'_{ij} と v'_{ij} が得られる，つまり新たな U' と V' が得られるので，それを U と V と見なして，上記の式を繰り返し適用する．

また各繰り返しの後に U を以下のように正規化する．

$$u'_{ij} \leftarrow \frac{u_{ij}}{\sqrt{\sum_i u_{ij}^2}} \tag{3}$$

繰り返しの終了は，繰り返しの最大回数を決めておくか，UV^T と X との距離 J の変化量から判定する．

$$J = ||X - UV^T||_F \tag{4}$$

J の値は NMF の分解の精度を表現している．NMF ではこの分解の精度がクラスタリングの目的関数となっており，この分解の精度が高い，つまり J の値が小さいほど，良好なクラスタリングであると推定する．

また $||\cdot||_F$ は Frobenius ノルムを表し，$m \times n$ の行列 A の Frobenius ノルムは以下で定義される．

$$||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} {a_{ij}}^2}$$

2.3 NMF の解の多様性

通常，行列 V と U の初期値にはランダムな値を与える．しかし式 1 と 2 による繰り返しは局所最適解にしか収束しないために，V と U の初期値の与え方によって，最終的に得られる V と U は大きく異なり，結果としてクラスタリングの精度も大きく異なる．

例えば，図 1 は本論文の実験で用いた文書データセット tr45 に対して，NMF によるクラスタリングの実験を 20 回行った結果である．ただし各実験での NMF の初期値にはランダムな値を与えており，各実験の初期値は異なる．図 1 の横軸は実験の番号を示し，縦軸はクラスタリングの精度を表している．図 1 から初期値によって得られる精度が大きく異なることが確認できる．

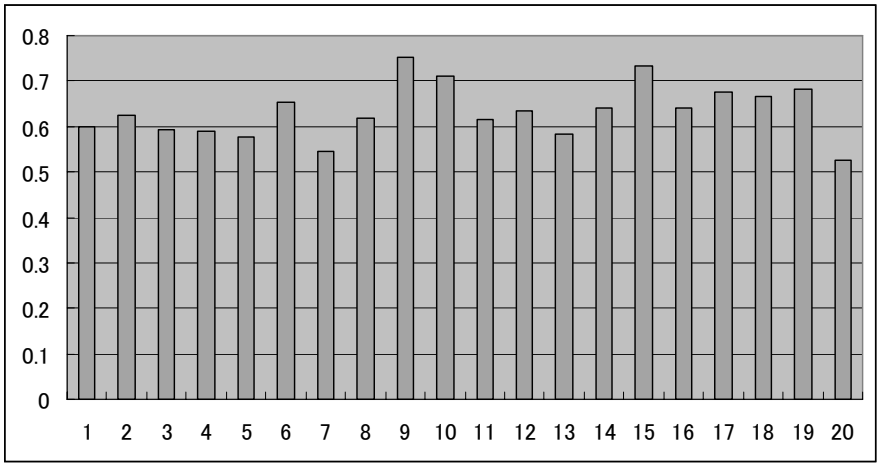

図 1　初期値とクラスタリングの精度

つまり，NMF は初期値によって得られるクラスタリング結果が異なる．通常は適当な初期値を与える実験を複数回行い，それらから得た複数個の解の中で X の分解の精度が最も高いものを選ぶ．しかし分解の精度は，直接的にはクラスタリングの精度を意味していないため，最も精度の高いクラスタリング結果を選択できる保証がない．

ここでは複数個のクラスタリング結果から１つを選択するのではなく，それらをアンサンブルするアンサンブルクラスタリングを試みる．

3　アンサンブルクラスタリング

3.1　ハイパーグラフによるデータの再表現

本手法のアンサンブルクラスタリングでは，NMF の初期値を様々に変化させて，複数個のクラスタリング結果を生成する．次に複数個得られたクラスタリング結果から各データに対するベクトル表現を新たに作成し，その新たにベクトル表現されたデータに対してクラスタリングを行うことで，アンサンブルクラスタリングを実現する．

ここでは複数個得られたクラスタリング結果からデータに対する新たなベクトル表現を作る方法を説明する．基本的には論文 (Strehl and Ghosh 2002) で提案されたハイパーグラフを用いる．

クラスタの数が k 個であり，得られているクラスタリング結果が m 種類の場合，各データは km 次元のベクトルで表現される．データ d の $k(i-1)+c$ 次元の値は，i 番目のクラスタリング結果として，データ d がクラスタ番号 c のクラスタに属していれば 1 を，属していなければ

0 を与える．この結果，データ d の km 次元のベクトル表現が得られる．

例を示す．$k = 3$，$m = 4$ とする．またデータは $\{d_1, d_2, \cdots, d_7\}$ の 7 つとする．4 種類のクラスタリング結果が以下のようになっていたとする．

第 1 のクラスタリング結果：

$$\{d_1, d_2, d_5\}, \{d_3, d_4\}, \{d_6, d_7\}$$

この結果から目的の行列の 1 列目から 3 列目が得られる．

$$\begin{array}{c} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

第 2 のクラスタリング結果：

$$\{d_1, d_5\}, \{d_2, d_3\}, \{d_4, d_6, d_7\}$$

この結果から目的の行列の 4 列目から 6 列目が得られる．

$$\begin{array}{c} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

第 3 のクラスタリング結果：

$$\{d_2, d_5\}, \{d_1, d_4\}, \{d_3, d_6, d_7\}$$

この結果から目的の行列の 7 列目から 9 列目が得られる．

$$\begin{array}{c} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \end{array} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

第 4 のクラスタリング結果：

$$\{d_1, d_5, d_7\}, \{d_3, d_4\}, \{d_2, d_6\}$$

この結果から目的の行列の 10 列目から 12 列目が得られる．

$$\begin{array}{c} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

以上の 4 つの行列を結合させ，以下の 7×12 の行列を得る．これがハイパーグラフである．このハイパーグラフにおける行ベクトルが，各データ（本論文の場合，文書）の新たなベクトル表現に対応している．このベクトルの類似度に基づいて，データをクラスタリングする．

$$\begin{array}{c} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \end{array} \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

3.2　重み付きハイパーグラフ

ハイパーグラフが表す行列の各要素の値は 0 か 1 のバイナリ値である．しかし値の意味を考えれば，その次元に対応するあるクラスタリング結果のあるクラスタに属する度合いと捉えられる．そのため 0 か 1 のバイナリ値ではなく，非負の実数値を与える方が適切である．

しかも NMF の場合，各クラスタリング結果では各クラスタに属する度合いに対応する値が行列 V に記載されている．そこでここではハイパーグラフの要素が 1 である部分を，行列 V の値から得ることで，非負の実数値を与えることにした．このようにして作成したハイパーグラフを，ここでは重み付きハイパーグラフと呼ぶ．

図 2 に重み付きハイパーグラフの作成例を示す．これは先の第 1 のクラスタリング結果に対応する部分である．d_1 から d_7 の 7 個の文書データセットを NMF により 3 グループにクラスタリングする．結果は行列 V で表される．次に行列 V を正規化する．V の各行に注目し，最大値の部分を 1 に，それ以外を 0 に変換したものが通常のハイパーグラフである．V の各行に注目し，最大値の部分はそのままに，それ以外を 0 に変換したものが本論文で提案する重み付きハイパーグラフである．

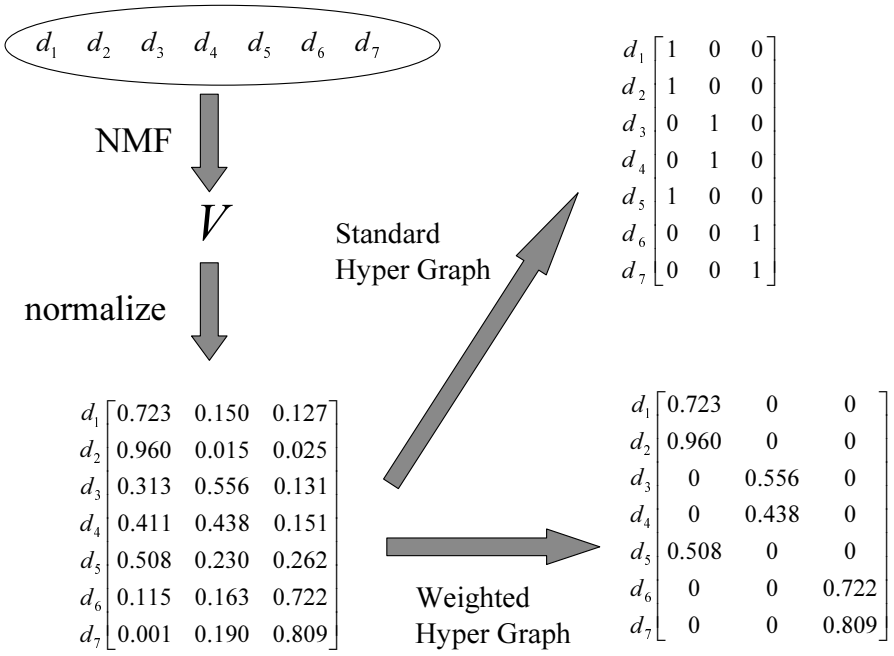

図 2　行列 V から作られる重み付きハイパーグラフ

4 実験

本手法の有効性を示すために，k-means，NMF，通常のハイパーグラフを使うアンサンブル手法および重み付きハイパーグラフを使うアンサンブル手法（本手法）の 4 種のクラスタリング結果を比較する．

利用するデータセットは以下のサイトで提供されている 18 種類である（表 1）．

`http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download`

データセットは通常の索引語文書行列で表現されており，正規化されていない．ここでは TF-IDF によって正規化を行った．

表 1 文書データセット

Data	文書数	語彙数	クラス数
cacmcisi	4663	41681	2
cranmed	2431	41681	2
fbis	2463	2000	17
hitech	2301	126373	6
k1a	2340	21839	20
k1b	2340	21839	6
la1	3204	31472	6
la2	3075	31472	6
re0	1504	2886	13
re1	1657	3758	25
reviews	4069	126373	5
tr11	414	6429	9
tr12	313	5804	8
tr23	204	5832	6
tr31	927	10128	7
tr41	878	7454	10
tr45	690	8261	10
wap	1560	6460	20

実験結果を表 2 に示す．表の値はクラスタリング結果のエントロピーを表し，低い値ほどクラスタリングが良好であることを意味する．

なお，ハイパーグラフのデータからのクラスタリングには，簡単のために，クラスタリング

表 2　他手法との比較実験結果

Data	k-means	NMF	NMF mean	standard hypergraph	weighted hypergraph
cacmcisi	0.750	0.817	0.693	0.691	**0.690**
cranmed	**0.113**	0.963	0.792	0.750	0.450
fbis	0.610	0.393	0.406	0.408	**0.381**
hitech	**0.585**	0.679	0.705	0.683	0.684
k1a	0.374	0.393	0.377	0.386	**0.351**
k1b	0.221	0.259	0.238	0.456	**0.216**
la1	0.641	0.464	0.515	**0.458**	0.459
la2	0.620	0.576	0.551	0.548	**0.468**
re0	**0.368**	0.419	0.401	0.383	0.379
re1	0.374	0.364	0.346	0.334	**0.325**
reviews	**0.364**	0.398	0.538	0.416	0.408
tr11	0.349	0.338	0.311	**0.300**	0.304
tr12	0.493	0.332	0.375	0.308	**0.307**
tr23	0.527	**0.485**	0.489	0.493	0.521
tr31	0.385	0.402	0.383	0.343	**0.334**
tr41	0.277	0.358	0.299	**0.245**	0.270
tr45	0.397	0.345	0.328	0.277	**0.274**
wap	0.408	0.371	0.374	0.336	**0.327**
平均	0.436	0.464	0.451	0.434	**0.397**

toolkit の CLUTO [3]を利用した．CLUTO はクラスタリング手法や類似度関数を様々に設定できるが，ここでは default の設定である k-way clustering と呼ばれる手法と cosine の類似度を用いた．またハイパーグラフのデータからのクラスタリング手法には任意のものが利用可能であり，高機能なクラスタリング手法を用いて，更に高い精度を得ることも可能である．ただしここではアンサンブルすることの効果と，ハイパーグラフに重みを付ける効果を明確に確認するために，簡易なものを用いた．

また，エントロピーについても注記しておく．エントロピーはクラスタリング結果を評価するための１つの尺度である．データセットのクラスタリングの正解が $\{K_h\}_{h=1}^{k}$ であり，得られたクラスタリングが $\{C_j\}_{j=1}^{k}$ となっているとき，クラスタ C_i に対するエントロピー E_i は以下で定義される．

$$E_i = -\sum_{h=1}^{k} P(K_h|C_i) \log P(K_h|C_i)$$

各クラスタに対して E_i を求め，クラスタのデータ数による重み付き平均をとることで全体のエントロピーが定義される．すなわち以下の式となる．

$$\sum_{i=1}^{k} \frac{|C_i|}{N} E_i$$

ここで N は全データ数を表す．また定義中に確率 $P(K_h|C_i)$ が出ているが，これは K_h と C_i に共通に存在するデータの数を n_{hi} と置き，$n_{hi}/|C_i|$ によって推定する．またクラスタリングの精度は，クラスタリング結果の各クラスタを正解のクラスタに対応つけ，n_{hi} の合計を N で割った値により求まる．つまりエントロピーの値の低さとクラスタリングの精度はほぼ対応していると見なせる．

本実験の場合，クラスタリングの精度を求めて，評価を行うことも可能ではあるが，クラスタリングの精度を求めるには，クラスタリング結果の各クラスタを正解のクラスタに対応させなくてはならない．この処理は組み合わせ最適化問題になっているために，単純には最適解が求まらない．そのために，ここではエントロピーによる評価を行っている．

NMF の実験では初期値を 20 個用意し，得られた 20 個のクラスタリング結果において，NMF の分解の精度（式 4 の値）が最も高いものを選び，それを NMF のクラスタリング結果とした．NMF mean とあるのは，20 個のクラスタリング結果の平均のエントロピーである．表の standard hypergraph が通常のハイパーグラフを使うアンサンブル手法，weighted hypergraph が重み付きハイパーグラフを使うアンサンブル手法（本手法）を意味する．

NMF と NMF mean を比較すると，NMF の方が若干エントロピーが大きい．つまりクラス

[3] http://glaros.dtc.umn.edu/gkhome/views/cluto

タリング結果を評価するのに，式 4 を使うのは最良ではないことがわかる．また NMF mean と weighted hypergraph を比較すると，18 個のデータセット中 17 個で本手法の方がエントロピーが小さい．つまりこの点からアンサンブルすることの効果が確認できる．また standard hypergraph と weighted hypergraph を比較すると，18 個のデータセット中 13 個で本手法の方がエントロピーが小さく，ハイパーグラフに重みを与える効果も確認できる．

なお 18 個中 13 個の改善は，統計的には以下のような観点から有意とみなした．standard hypergraph と weighted hypergraph のパフォーマンスが同程度である場合, standard hypergraph のエントロピーから weighted hypergraph のエントロピーを引いた値（値が大きいほど改善の度合いが高い）は平均 0 の正規分布と考えられる．そこで有意水準 0.05 として t-検定の片側検定を用いると，棄却域は自由度が 17 であることに注意すると 1.74 以上となる．実際の値は standard hypergraph のエントロピーから weighted hypergraph のエントロピーを引いた値の標本平均が 0.03706，標本分散が 0.007389 なので，

$$\frac{0.03706 - 0}{\sqrt{0.007389/17}} = 1.78 > 1.74$$

となり，パフォーマンスが同程度という仮説が棄却できる．

5　考察と関連研究

一般に複数の解をアンサンブルすると，複数の解の平均よりも良い値が得られると考えられる．本実験でも 18 個のデータセット中 17 個でアンサンブルの効果が得られているが，データセット tr23 に関しては，本手法のエントロピーの値の方が高い．これは解の分散の影響と考えられる．

実験で得られた各データセットに対する NMF による 20 個のクラスタリング結果のエントロピーの分散と，表 2 における NMF mean と weighted hypergarph との差（つまりアンサンブルによる改善の度合い）をプロットした図を図 3 に示す．図の横軸が分散を示し，縦軸が weighted hypergarph と NMF mean との差（改善の度合い）を示している．

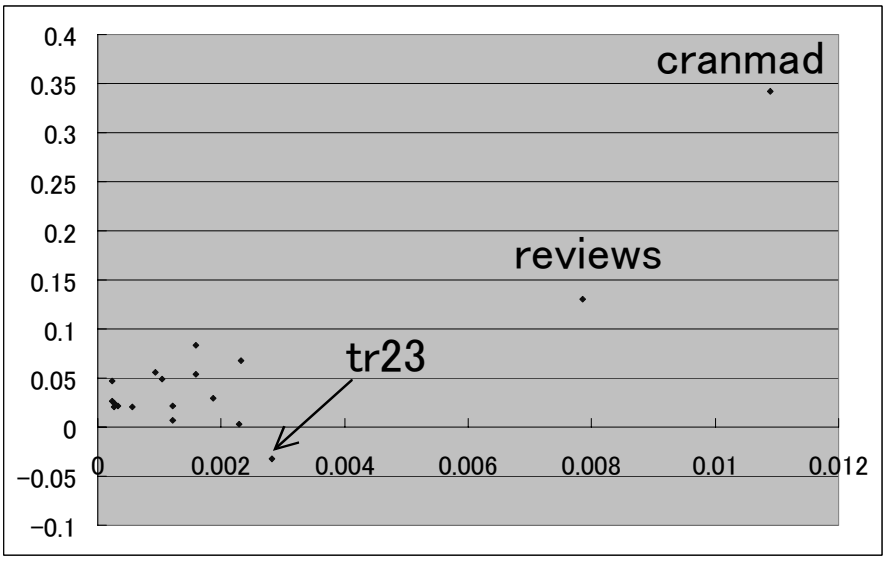

図 3　解の分散とアンサンブルによる改善

図 3 をみると，分散が大きい２つ（cranmad と reviews）は，アンサンブルによる改善の度合いも大きいことが分かる．そして３番目に分散が大きなデータセットが tr23 である．つまり分散の大きな解をアンサンブルすると，非常に良い結果を得ることもあるが，逆に悪い結果を得ることもあり得ると考えられる．

データセット tr23 に対する NMF の結果を見ると，１つだけ非常にエントロピーの低いクラスタリング結果が得られていた．この解を取り除いて，19 個のクラスタリング結果で本手法によるアンサンブルを試したところ，NMF mean のエントロピーは 0.493，weighted hypergarph のエントロピーは 0.492 となり，アンサンブルの効果が現れた．

また，ここでは NMF で複数個のクラスタリング結果を生成する際に，個々のクラスタリング結果のクラスタ数は，最終的なクラスタ数と一致させている．しかしハイパーグラフの考え方を用いれば，生成される個々のクラスタリング結果のクラスタ数は任意でかまわない．実際に k-means では少ないクラスタ数に直接クラスタリングするよりも，多数のクラスタに分割してから，目的のクラスタ数にまとめた方が効果があることが経験的にわかっている．論文 (Fred and Jain 2002) ではこのヒューリスティクスを利用して，多数のクラスタに分割してから，アンサンブルを行っている．本手法においても，そのような工夫を取り入れることも可能である．

本手法ではハイパーグラフの値として，1 に当たる部分を行列 V の値を用いることで，実数値に変換した．この効果は実験で確認できている．この工夫を更に進めると，0 に当たる部分にも行列 V の値を用いることで，実数値に変換することが考えられる．この場合，ハイパーグラ

フは単純に各クラスタリング結果に対応する行列 V を結合させたものになる．実際にこのようにして作ったハイパーグラフに対して，クラスタリングを行ってみた．結果を表 3 に示す．ここで hypergraph V が行列 V を結合させてハイパーグラフを作成する手法を示す．

表 3　V を結合したハイパーグラフによるクラスタリング

Data	hypergraph V	weighted hypergraph
cacmcisi	0.778	**0.690**
cranmed	0.525	**0.450**
fbis	0.402	**0.381**
hitech	0.688	**0.684**
k1a	0.366	**0.351**
k1b	**0.205**	0.216
la1	0.491	**0.459**
la2	0.486	**0.468**
re0	**0.378**	0.379
re1	0.337	**0.325**
reviews	**0.391**	0.408
tr11	**0.280**	0.304
tr12	0.316	**0.307**
tr23	**0.474**	0.521
tr31	**0.310**	0.334
tr41	0.340	**0.270**
tr45	0.380	**0.274**
wap	0.344	**0.327**
平均	0.416	**0.397**

通常のハイパーグラフを使うよりも結果は良好であるが，1 に当たる部分だけを精密化する方が効果があることがわかる．また 0 の値はそのままにしている方が，ハイパーグラフがスパースになり，データ間の類似度が 0 であるケースが生じやすくなる．そのためグラフスペクトル理論を用いたクラスタリング手法 (Ding, He, Zha, Gu, and Simon 2001) なども使えるようになるために好ましい．

最後にアンサンブル学習 (Breiman 1996) との関連について述べる．アンサンブル学習とアンサンブルクラスタリングの違いは，クラスタにラベルがつくかどうかである．アンサンブル学習ではデータにラベルが付くので，そのラベルをもつデータがラベル付きのクラスタと見なせる．アンサンブルクラスタリングの場合は，クラスタにラベルがついていない．もしもクラス

タにラベルをつけることができれば，アンサンブル学習の手法を直接利用できるために，さらなる改良や発展が可能である．クラスタにラベルをつける処理は，クラスタ数が 2 や 3 などの小さい場合はそれほど大きな問題ではないので，今後はクラスタにラベルをつけるという戦略で，アンサンブルを行う手法を開発したい．

6 おわりに

本論文では，NMF を用いたアンサンブルクラスタリングの手法を提案した．NMF の初期値を変化させて，複数個のクラスタリング結果を得る．次に得られた複数個のクラスタリング結果をハイパーグラフで表現し，それをクラスタリングすることで最終的なクラスタリング結果を得る．ハイパーグラフを作成する際に，NMF より得られた行列 V を利用して，1 の部分に実数値の重み付けする工夫を取り入れた．

実験では 18 個のデータセットを用いて，k-means，NMF，通常のハイパーグラフを使うアンサンブル手法および重み付きハイパーグラフを使うアンサンブル手法（本手法）の比較を行った．エントロピーで評価を行い，本手法の有効性を確認できた．

個々のクラスタリングで生成させるクラスタ数を変化させること，クラスタ数が小さい場合は，クラスタにラベルを与えて，アンサンブル学習の手法を利用することなどを今後の課題とする．

参考文献

Boley, D., Gini, M. L., Gross, R., Han, E.-H., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., and Moore, J. (1999). “Document Categorization and Query Generation on the World Wide Web Using WebACE.” *Artificial Intelligence Review*, **13** (5-6), pp. 365–391.

Breiman, L. (1996). “Bagging Predictors.” *Machine Learning*, **24** (2), pp. 123–140.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). “Indexing by Latent Semantic Analysis.” *Journal of the American Society of Information Science*, **41** (6), pp. 391–407.

Ding, C., He, X., Zha, H., Gu, M., and Simon, H. (2001). “Spectral Min-max Cut for Graph Partitioning and Data Clustering.” In *Lawrence Berkeley National Lab. Tech. report 47848*.

Fern, X. Z. and Brodley, C. E. (2003). “Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach.” In *the 20th International Conference of Machine Learning (ICML-03)*.

Fred, A. L. and Jain, A. K. (2002). "Data Clustering Using Evidence Accumulation." In *the 16th international conference on pattern recognition*, pp. 276–280.

Hearst, M. A. and Pedersen, J. O. (1996). "Reexamining the cluster hypothesis: Scatter/gather on retrieval results." In *Proceedings of SIGIR-96*, pp. 76–84.

Kummamuru, K., Lotlikar, R., Roy, S., Singal, K., and Krishnapuram, R. (2004). "A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results." In *Proceedings of WWW-04*, pp. 658–665.

Lee, D. D. and Seung, H. S. (2000). "Algorithms for Non-negative Matrix Factorization." In *NIPS*, pp. 556–562.

Leuski, A. (2001). "Evaluating Document Clustering for Interactive Information Retrieval." In *Proceedings of CIKM-01*, pp. 33–40.

Michael W. Berry (Ed.) (2003). *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer.

Strehl, A. and Ghosh, J. (2002). "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions." In *Conference on Artificial Intelligence (AAAI-2002)*, pp. 93–98.

Topchy, A., Jain, A. K., and Punch, W. (2003). "Combining Multiple Weak Clusterings." In *In The Third IEEE International Conference on Data Mining (ICDM'03)*.

Xu, W., Liu, X., and Gong, Y. (2003). "Document clustering based on non-negative matrix factorization." In *Proceedings of SIGIR-03*, pp. 267–273.

Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., and Ma, J. (2001). "Learning to Cluster Web Search Results." In *Proceedings of SIGIR-04*, pp. 33–40.

略歴

新納浩幸：昭和 60 年東京工業大学理学部情報科学科卒業．昭和 62 年同大学大学院理工学研究科情報科学専攻修士課程修了．同年富士ゼロックス，翌年松下電器を経て，平成 5 年 4 月茨城大学工学部システム工学科助手．平成 9 年 10 月同学科講師，平成 13 年 4 月同学科助教授，現在，茨城大学工学部情報工学科准教授．博士 (工学)．機械学習や統計的手法による自然言語処理の研究に従事．言語処理学会，情報処理学会，人工知能学会，ACL 各会員．

佐々木稔：平成 8 年徳島大学工学部知能情報工学科卒業．平成 13 年同大学大学院博士後期課程修了．博士 (工学)．平成 13 年 12 月茨城大学工学部情報工学科助手．現在，茨城大学工学部情報工学科講師．機械学習や統計的手法による情報検索，自然言語処理等に関する研究に従事．言語処理学会，情報処理学会 各会員．