文書クラスタリングのためのCollapsed変分ベイズLDAによる次元圧縮

正田　備也†　　喜安　千弥†　　宮原　末治†

† 長崎大学工学部　〒 852–8521 長崎県長崎市文教町 1–14
E-mail: †{masada,kiyasu,miyahara}@cis.nagasaki-u.ac.jp

あらまし　この論文では，2006 年に Teh らによって提案された，collapsed 変分ベイズ法に基づく潜在的ディリクレ配分法 (CVBLDA) を，文書ベクトルの次元圧縮に応用する．次元圧縮された文書ベクトルをクラスタリングし，得られたクラスタの質によって CVBLDA の性能を評価する．評価に際しては，他の次元圧縮手法として，PLSI および PCA と比較する．また，LDA のような新しい確率論的モデルについて，非英語の文書集合を利用した実験結果の蓄積が多くないことに鑑み，日本語および韓国語の文書集合を，実験対象とする．

キーワード　LDA, collapsed 変分ベイズ法, 次元圧縮, 文書クラスタリング

Applying Latent Dirichlet Allocation with Collapsed Variational Bayesian Inference to Dimensionality Reduction in Document Clustering

Tomonari MASADA†, Senya KIYASU†, and Sueharu MIYAHARA†

† Faculty of Engineering, Nagasaki University　Bunkyo-machi 1–14, Nagasaki-shi, Nagasaki, 852–8521 Japan
E-mail: †{masada,kiyasu,miyahara}@cis.nagasaki-u.ac.jp

Abstract　In this paper, we use LDA with collapsed variational Bayesian inference (CVBLDA), proposed by Teh et al. in 2006, for dimensionalty reduction. We evaluate the efficiency of CVBLDA as a dimentionality reduction method with the quality of clusters obtained by using document vectors of reduced dimension. We compare CVBLDA with PLSI and PCA. There are not so many experimental results where an emerging probabilistic document model such as LDA is applied to non-English documents. Therefore, we use Japanese and Korean documents and contribute to the accumulation of experimental data on the applicability of LDA on non-English documents.

Key words　LDA, collapsed variational Bayesian inference, dimentionality reduction, document clustering

1. はじめに

近年，データ・マイニングの分野では，従来から使われてきた確率モデルに，過学習を防ぐ等の目的でベイズ推定の立場から改良を加え，様々な新しい確率モデルが提案されている [4]．本論文では，そのようなモデルのうち，潜在的ディリクレ配分法 (LDA: latent Dirichlet allocation) [6] に着目する．この確率モデルは，確率的潜在的意味インデキシング (PLSI: probabilistic latent semantic indexing) [10] をベイズ推定の立場から改良版したモデルとして理解できる．特に，本論文では，LDA について新たに提案されたパラメータ推定法である collapsed 変分ベイズ法 (collapsed variational Bayesian inference) [15] に注目する．以下，collapsed 変分ベイズ法によりパラメータ推定を行う LDA を CVBLDA と略記する．そして，CVBLDA を文書ベクトルの次元圧縮に用い，次元圧縮後の文書ベクトルをクラスタリングして得られるクラスタの質を評価することで，CVBLDA の有効性を評価する．比較のために，PLSI，および主成分分析 (PCA: principal component analysis) [14] を次元圧縮に用い，やはり次元圧縮後の文書ベクトルをクラスタリングすることで評価する．文書クラスタリングには，k-means 法，混合ガウス分布に基づく方法 [5]，混合多項分布に基づく方法 [12] を用いる．論文の構成は，次の通りである．2 節で LDA に関する従来研究を紹介する．第 3 節では，本論文で次元圧縮に用いる手法，第 4 節ではクラスタリングに用いる手法を，それぞれ説明する．第 5 節では，用いる文書集合の基礎データを示し，第 6 節で，評価実験の詳細と考察を提示する．第 7 節は，まとめと今後の課題を含む．

2. 従来研究

LDA を次元圧縮に用いて文書分類によって評価した実験はある [6] が，実験対象の文書の量も少なく，また，英語の文書のみが対象となっている．LDA による次元圧縮をクラスタリングに応用した例はあるが，画像データについての実験である [8]．我々は，次元圧縮法としての LDA を文書クラスタリン

グによって評価する実験を，すでに一度おこなっている [11] が，これと本論文は，以下の点で異なる．

- LDA のパラメータ推定に，変分ベイズ法 [6] より近似が良いと言われる collapsed 変分ベイズ法 [15] を全面的に用いた．
- ランダム投影法ではなく PCA をベースラインとし，LDA や PLSI の相対的な性能を分かりやすくした．
- 同じ文書集合を SVM によって分類したときの精度を，併せて示した．もちろん，文書分類は教師あり学習によるカテゴリー分けであり，一方，文書クラスタリングは教師なし学習によるカテゴリー分けである．そのため，直接に結果を比較することはできない．だが，文書分類の精度は，取り組んでいる問題の難易度を知る手がかりになると考えられる．

3. 文書ベクトルの次元圧縮

与えられた文書集合に文書が I 件含まれ，それらに全部で J 種類の語彙が現れているとする．これらの文書の特徴ベクトルとして，各語彙が文書に現れる回数をエントリとする J 次元ベクトルを考える．一般に，個々の文書に現れる語彙の種類は，J よりかなり少ない．つまり，文書と語彙とのユニークなペアの個数を M とすると，通常 $M \ll IJ$ が成り立つ．つまり，文書ベクトルは疎である．ここに，検索やクラスタリング，分類など，ベクトルを文書の特徴量として利用するアプリケーションの性能を落とさずに，小さい次元のベクトルを代わりに用いることができるか，という課題を設定する余地が生じる．以下，次元圧縮後の文書ベクトルの次元を K とする．

本論文では，次元圧縮法として CVBLDA，PLSI，PCA を用いる．そして，次元圧縮後のベクトルをクラスタリングし，各文書に人手で付与されたカテゴリを正解とみなしてクラスタの質を評価することで，次元圧縮の性能評価とする．

3.1 PCA

PCA には，通常，$J \times J$ の共分散行列と $I \times I$ のグラム行列のうち，サイズの小さい行列を入力として用いることができる．だが，I と J がともに大きい場合，さらに空間計算量の少ない方法が必要となる．本論文では，PCA のための EM アルゴリズム [14] を用い，空間計算量を $O(IK + JK + M)$ に抑えた．EM アルゴリズムにおいて，パラメータはランダムに初期化する．だが，そもそも PCA は解が一意に定まる問題であるため，初期値への依存性は小さいことが予想される．

3.2 PLSI

今回の PLSI の実装では，EM アルゴリズムにアニーリングを併用する [10]．反復計算は，対数尤度の変化率が一定値を下回ったら打ち切る．次元圧縮後の文書ベクトルは，文書 d_i が所与のときのトピック t_k の確率 $P(t_k|d_i)$ に，文書 d_i の文書長（文書 d_i を構成するトークン数）を掛けた値によって構成する．これにより，ベクトルのエントリの総和が，元の文書ベクトルのエントリの総和に等しくなる．

3.3 CVBLDA

本論文のための実装では，論文 [15] に示された，正規分布による近似を伴う変分ベイズ推定法を用いる．また，文書 d_i と語彙 w_j が所与のときのトピック t_k の確率を表す変分パラメータ γ_{ijk} の全語彙にわたる和 $\sum_j \gamma_{ijk}$ を，次元圧縮後の文書ベクトルのエントリとする．なお，$\sum_k \gamma_{ijk} \equiv 1$ となるので，次元圧縮後のベクトルのエントリの総和は，やはり，元の文書ベクトルのエントリの総和に等しくなる．Dirichlet パラメータである α, β は，ともに 0.1 と設定している．

4. 次元圧縮の評価のためのクラスタリング

本論文では，次元圧縮後の文書ベクトルをクラスタリングするために，k-means，混合ガウス分布に基づく方法，混合多項分布に基づく方法という，3 種類の方法を用いる．本論文は次元圧縮手法の性能比較を目的とするので，文書のカテゴリ数は既知と仮定し，クラスタ数の推定は考察の対象外とした．

4.1 k-means

今回の実装では，各文書の属するクラスタをランダムに決めた後，クラスタの重心を計算するステップと，各文書を最も近い重心に対応するクラスタへと属し直させるステップとを，全クラスタのメンバーが変化しなくなるまで繰り返す．なお，単語の出現頻度をエントリとする文書ベクトルに，そのまま k-means を適用すると，クラスタの質が良くなかった．そのため，次元圧縮前の文書ベクトルとしては，tf-idf 値をエントリとするベクトルを長さ 1 に正規化して用いる．tf-idf の式は $\sqrt{n_{ij}} \log(I/I_j)$ とする．ここで，n_{ij} は文書 d_i に語彙 w_j が出現する回数，I_j は語彙 j が現れる文書数である．次元圧縮後の文書ベクトルには，正規化のみおこない，k-means を適用する．

4.2 混合ガウス分布 (GM)

本論文では，共分散行列が対角行列のガウス分布の混合分布を，クラスタリングのモデルとして用いる．今回の実装では，上述の k-means を初期化ステップとして利用した後，EM アルゴリズムを実行し，対数尤度の変化率が一定値を下回ったら反復を終える．k-means と同様，次元圧縮前の文書ベクトルとしては，tf-idf によりエントリを計算したものを長さ 1 に正規化して使う．次元圧縮後の文書ベクトルには，正規化だけを施す．

4.3 混合多項分布 (MM)

混合多項分布に基づくクラスタリングでは，モデルのパラメータをランダムに初期化した後，EM アルゴリズムを対数尤度の変化率が一定値を下回るまで反復する．今回は，スムージング [12] とアニーリング [13] を併用する．ただし，論文 [11] において，次元圧縮後のベクトルについては，スムージングの有無でクラスタリングの性能がほぼ変わらないことが確認されている．そのため，次元圧縮前の文書ベクトルにのみ，Dirichlet スムージングを適用する．スムージング・パラメータの値は，予備実験の結果に基づき，0.1 としている．

5. 実験に用いる文書集合

今回の実験では 4 種類の文書集合を用いた．どの文書集合からも，あらかじめ高頻度語と低頻度語を除去している．

JIC Web サイト `japan.internet.com` で公開されている 2001〜2006 年の記事．カテゴリは，Web ビジネス (9,059)，Web マーケティング (6,109)，Web テクノロジー (4,596) の 3 つ（括弧内は記事件数）．データ全体のスペックは $I = 19,764, J =$

表 1　LIBLINEAR [3] による，次元圧縮前の文書ベクトルを用いた場合の 2-fold 文書分類の精度

文書集合	JIC	OK	S5	S6
カテゴリ数	3	8	4	9
文書数	19,764	52,112	22,049	37,017
語彙数	13,517	13,300	14,563	25,584
分類精度	75.12	78.37	85.29	81.40

$13,517, M = 1,476,483$．記事内容の性格上，異なるカテゴリの記事で同じ語彙が共有されることも多い．

OK　Web サイト `okwave.jp` で公開されている質問文．カテゴリは，コンピューター (7,055)，旅行・レジャー・生活 (4,103)，デジタルライフ (11,909)，ビジネス＆キャリア (5,985)，マネー (4,150)，ライフ (7,672)，学問＆教育 (7,149)，社会 (4,589)，趣味 (6,725)，美容＆健康 (3,030) の 8 つ．データ全体のスペックは $I = 52,112, J = 13,300, M = 1,424,907$．質問文のため，文書長は比較的短いが，文書数は多くしてある．

S5　ソウル新聞の Web サイト `www.seoul.co.kr` で公開されている 2005 年の記事．カテゴリは，教育 (6,172)，国際 (3,048)，政治 (3,608)，社会 (9,222) の 4 つ．データ全体のスペックは $I = 22,049, J = 14,563, M = 1,150,512$．

S6　同じソウル新聞の Web サイトで公開されている，2006 年の記事．カテゴリは，行政 (1,504)，文化 (4,870)，経済 (6,745)，芸能 (1,710)，国際 (2,498)，政治 (3,806)，地域 (3,923)，社会 (8,947)，スポーツ (3,016) の 9 つ．データ全体のスペックは $I = 37,017, J = 25,584, M = 2,044,669$．年次が違うだけなので文書として S5 と質が似ているが，カテゴリ数は異なる．

形態素解析は，日本語である JIC と OK は MeCab [2] で，韓国語である S5 と S6 は KLT [1] で，それぞれおこなった．各文書集合について，文書中の単語の出現頻度をエントリとする文書ベクトルを使い，大規模データ用 SVM ライブラリ LIBLINEAR [3] で 2-fold の文書分類をおこなって評価した結果を表 1 に示す．パラメータ値の探索は付属の `grid.py` というスクリプトによる．正規化は，分類精度が下がるため，おこなっていない．本論文では，次元圧縮を，文書分類（教師あり学習）ではなく文書クラスタリング（教師なし学習）によって評価するので，直接比較はできないが，各文書集合のカテゴリ分けの難易度を把握する手がかりにはなると考えられる．

6. 実験の詳細

6.1 評価方法

あるクラスタで，文書数が最多のカテゴリを支配的と呼ぶとする．クラスタの precision とは，クラスタ内での支配的カテゴリの文書数を，クラスタに含まれる文書数で割った値である．クラスタの recall とは，クラスタ内での支配的カテゴリの文書数を，そのカテゴリに含まれる全文書数で割った値である．1 つのクラスタリング結果は，複数のクラスタから成るが，各クラスタの precision と recall を計算するための分母と分子について，それぞれ総和をとってから，割り算した結果を，microaveraged precision および microaveraged recall と呼び，これらを x 座標，y 座標とする点をグラフ上にプロットする．

次元圧縮手法は，いずれも初期値に依存するので，10 通りのランダムに設定された初期値を用いている．さらに，クラスタリング手法のほうも，いずれも初期値に依存するので，やはり 10 通りのランダムな初期値を使う．結局，ある文書集合から得られる文書ベクトルを，特定の次元 K に圧縮した場合について，100 通りの評価結果を得る．つまり，microaveraged precision/recall のペアを 100 個得る．これらをすべて，同じグラフ上にプロットした結果を，図 1，2，3，4 に示す．4 つの図は，順に，4 つの文書集合 JIC，OK，S5，S6 に対応する．

すべての図は 8 つのグラフを含む．どのグラフにおいても，横軸が microaveraged precision，縦軸が microaveraged recall である．大きいマーカ “△” は，元の文書ベクトルをクラスタリングした結果の評価である．この場合は次元圧縮が介在しないので，次元圧縮した場合と同じようにクラスタリングの初期値を 10 通りだけ用意すると，評価データの数が 10 個のみと少なくなる．そこで，クラスタリングにおけるランダムな初期値を 30 通り用意した．つまり，マーカ “△” は 30 個ある．

“×”，“+”，“○” という 3 種類のマーカは，それぞれ，文書ベクトルの次元を 16，32，64 次元に圧縮してからクラスタリングした結果の評価を示す．これらのマーカについては，各々 100 通りの結果がどのグラフにもプロットされている．

各グラフの説明は，グラフ内にキャプションとして書かれている．図 1，2，3，4 で最上段の 2 つのグラフは，CVBLDA（左）または PLSI（右）で次元圧縮をおこなって得られたベクトルを，混合多項分布 (MM) に基づく手法によりクラスタリングした場合の評価結果を示している．これら 2 つのグラフには，元の文書ベクトルをクラスタリングした結果の評価データも，マーカ “△” で示されている．2 段目のグラフは，CVBLDA（左）または PLSI（右）で次元圧縮してから，混合ガウス分布 (GM) をクラスタリング手法として用いた場合に対応している．3 段目には，CVBLDA（左），または PLSI（右）で次元圧縮し，k-means をクラスタリング手法として用いた場合の評価結果を示されている．マーカ “△” は，次元圧縮せずに元の文書ベクトルをクラスタリングした場合に対応する．なお，以上 3 段目までのグラフでは，左右でクラスタリング手法が同一のため，マーカ “△” の位置は，左右のグラフで同じである．

最下段は，PCA により次元圧縮されたベクトルに，GM（左）または k-means（右）を適用して得られたクラスタリングの評価を示している．PCA は，PLSI と LDA に比べ質の低いクラスタを与えているため，その評価結果は最下段にまとめた．ここでも，元の文書ベクトルをクラスタリングした場合については “△” で示したが，最下段のみ，左右のグラフでクラスタリング手法が異なるため，マーカ “△” の位置が異なっている．

6.2 考　　察

図 1，2，3，4 に共通して，以下のことが観察される．

- CVBLDA と PLSI により次元圧縮されたベクトルは，元の文書ベクトルと，ほぼ同じ質のクラスタを与える．
- CVBLDA と PLSI とを比べると，どの次元に文書ベクトルを圧縮した場合も，クラスタの質は互角である．

表 2　LIBLINEAR [3] による次元圧縮後の 2-fold 文書分類の精度

文書集合	JIC	OK	S5	S6
元の文書ベクトル	75.12	78.37	85.29	81.40
CVBLDA16	66.53	68.74	76.69	65.78
PLSI16	67.08	68.68	78.11	66.09
PCA16	62.63	45.10	74.08	61.12
CVBLDA32	69.17	71.93	79.81	71.14
PLSI32	71.32	71.83	80.98	71.91
PCA32	68.69	55.20	77.50	65.37
CVBLDA64	72.17	**75.02**	81.90	73.75
PLSI64	**73.25**	73.22	**82.77**	**74.58**
PCA64	71.02	61.96	78.84	69.66

- PCA は，CVBLDA や PLSI に及ばない．
- クラスタリング手法は MM，GM，k-means の順で優れている．この順序は，どの次元圧縮手法でもほぼ変わらない．

すべての図で，PCA により次元圧縮されたベクトルの評価結果（最下段の 2 つのグラフ）のばらつきが小さいのは，PCA のための EM アルゴリズムでパラメータをランダムに初期化しても，そもそも PCA は解が一意に定まる問題のため，次元圧縮の結果に多様性が乏しいからである．つまり，PCA のグラフにおけるマーカの散らばりは，ほぼ，クラスタリング・アルゴリズムの初期値のランダム性によると考えてよい．

なお，グラフの対角線上にあるマーカは，microaveraged precision と microaveraged recall の一致を意味するが，これは，どのカテゴリについても，対応するクラスタが 1 つ構成されている場合に対応する．言い換えれば，すべてのクラスタで少数派として埋もれてしまっているカテゴリが 1 つもない場合に対応する．つまり，precision や recall の値が高いか低いかにかかわらず，マーカ全体の分布が対角線に近いほど，クラスタリング処理を経ることで見落とされるカテゴリが少ない．この観点からは，S5 と S6 において，CVBLDA が PLSI よりも，特に $K = 64$ の場合にやや優れている．だが，全体としては，両手法は大きな差を示していない．

表 2 に，各次元圧縮手法によって得られた文書ベクトルを，表 1 と同様，大規模データ用 SVM ライブラリである LIBLINEAR によって 2-fold 分類をおこなった結果の，分類精度を示す．いずれの次元圧縮方法についても，10 通りのランダムな初期値からパラメータ推定を始めることで，次元圧縮結果を得ている．そのため，表に示した分類精度は，これら 10 通りの圧縮結果について得られた精度の平均値である．太字は，各文書集合について得られた最も高い平均精度を表す．文書集合 OK では CVBLDA が PLSI をわずかに上回ったが，その差は小さい．したがって，クラスタリングによる評価実験と同様，CVBLDA と PLSI は注目すべき差を示さないと結論できる．なお，計算時間の短さでは，PLSI が CVBLDA よりもやや有利だった．

7. おわりに

本論文における日本語と韓国語の文書を用いた実験は，少なくとも文書クラスタリングのための次元圧縮に関しては，CVBLDA が PLSI と同等の結果を示すに止まったことを示している．そもそも LDA が PLSI の洗練化であること，また，英語の文書を用いた既存の実験では LDA の利点が主張されてきたことに鑑みれば，日本語や韓国語の文書の統計的性質のうちで，LDA の利点を打ち消すものがあるのかもしれない．

ここで，次のような事柄について考えてみたい．まず，英語では多様な意味を同じ単語を使って表せる場合（同じ動詞＋異なる副詞，同じ動詞＋異なる前置詞，等）が多い一方，日本語と韓国語では，単語を変えることで容易に意味を表し分けられる場合が多い．さらに，日本語と韓国語では，固有語（日本語のやまとことば，韓国語のウリマル）と漢字語，さらには外来語が併用されるなど，同じ意味を異なる単語で表せる場合も多い．また，この傾向は，漢字・ひらがな・カタカナと複数の文字を使う日本語のほうが強いだろう．すると，過学習を防ぐことができるという LDA の特徴は，日本語や韓国語では，むしろ，語彙の多様性への適応が難しくなる，という結果につながらないだろうか．ベイズ推定の立場から既存の確率モデルの洗練化として提案された，LDA 以外の様々な確率モデルについても，英語以外の言語の文書集合において，今後，その有効性が検証されるべきだろう．そして，各言語がもつ統計的性質と確率モデルとの相性について，組織的な検証が必要であろう．

文　献

[1] http://nlp.kookmin.ac.kr/HAM/kor/

[2] http://mecab.sourceforge.net/

[3] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

[4] M. J. Beal. Variational Algorithms for Approximate Bayesian Inference. PhD. Thesis, University College London, 2003.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] D. Blei, A. Y. Ng and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, 2003.

[7] A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, Vol. 39, No. 1, pp. 1-38, 1977.

[8] P. K. Elango and K. Jayaraman. Clustering Images Using the Latent Dirichlet Allocation Model. 2005. (available at http://www.cs.wisc.edu/~pradheep/)

[9] T. Griffiths and M. Steyvers. Finding Scientific Topics. *Proc. of the National Academy of Sciences* 101 (suppl. 1), pp. 5228-5235, 2004.

[10] T. Hofmann. Probabilistic Latent Semantic Indexing. in *Proc. of SIGIR'99*, pp. 50-57, 1999.

[11] T. Masada, S. Kiyasu and S. Miyahara. Comparing LDA with pLSI as a Dimensionality Reduction Method in Document Clustering. in *Proc. of LKR'08*, 2008. (to appear)

[12] K. Nigam, A. McCallum, S. Thrun and T. Mitchell. Text Classification from Labeled and Unlabeled Documents Using EM. *Machine Learining*, Vol. 39, No. 2/3, pp. 103-134, 2000.

[13] K. Rose, E. Gurewitz, and G. Fox. A Deterministic Annealing Approach to Clustering. *Pattern Recognition Letters*, Vol. 11, pp. 589-594, 1990.

[14] S. Roweis. EM Algorithms for PCA and SPCA. in *Proc. of NIPS'97*, pp.626-632, 1997.

[15] Y. W. Teh, D. Newman and M. Welling. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. in *Proc. of NIPS'06*, pp. 1353-1360, 2006.

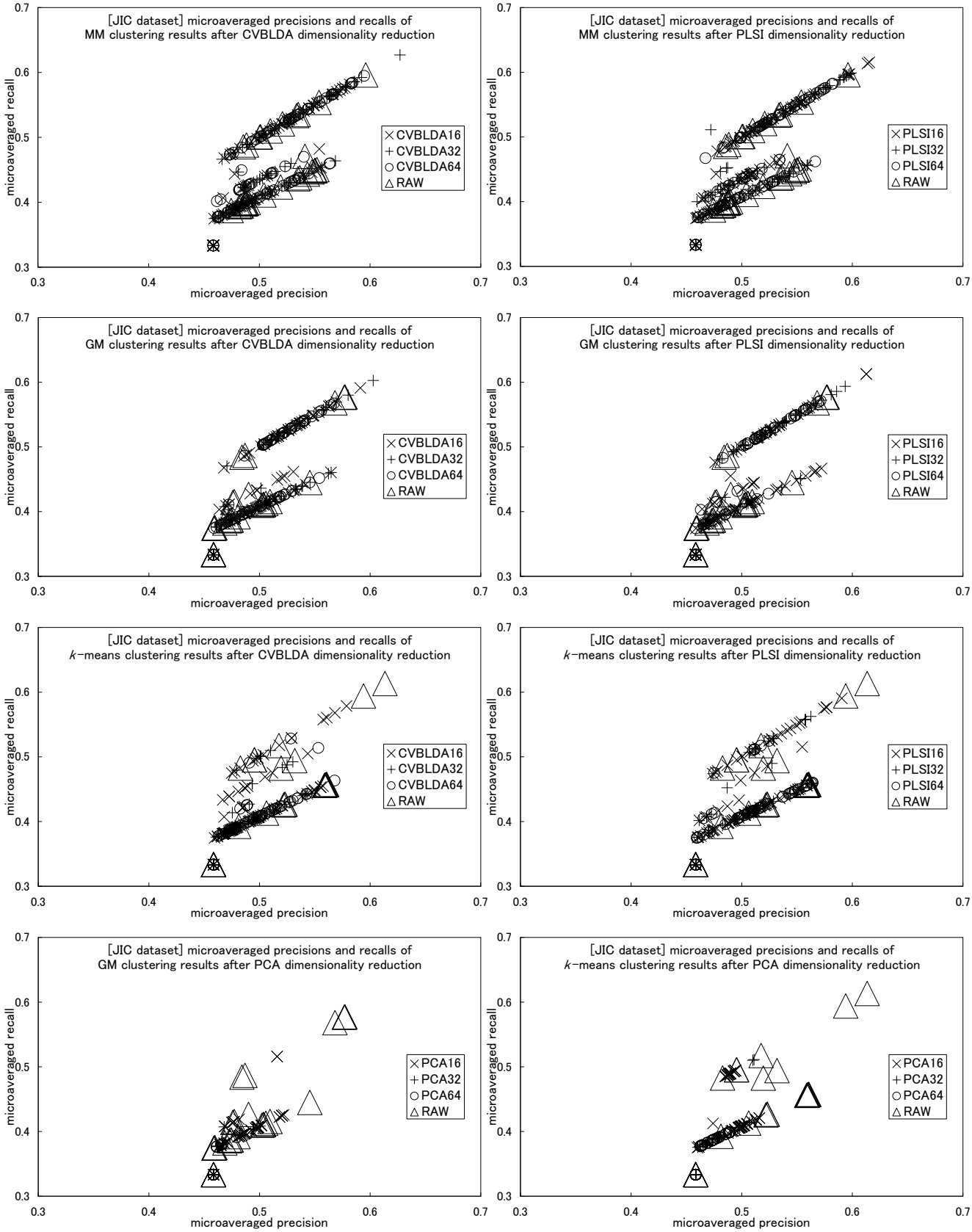

図 1　文書集合 JIC について次元圧縮後の文書ベクトルをクラスタリングした場合の評価結果

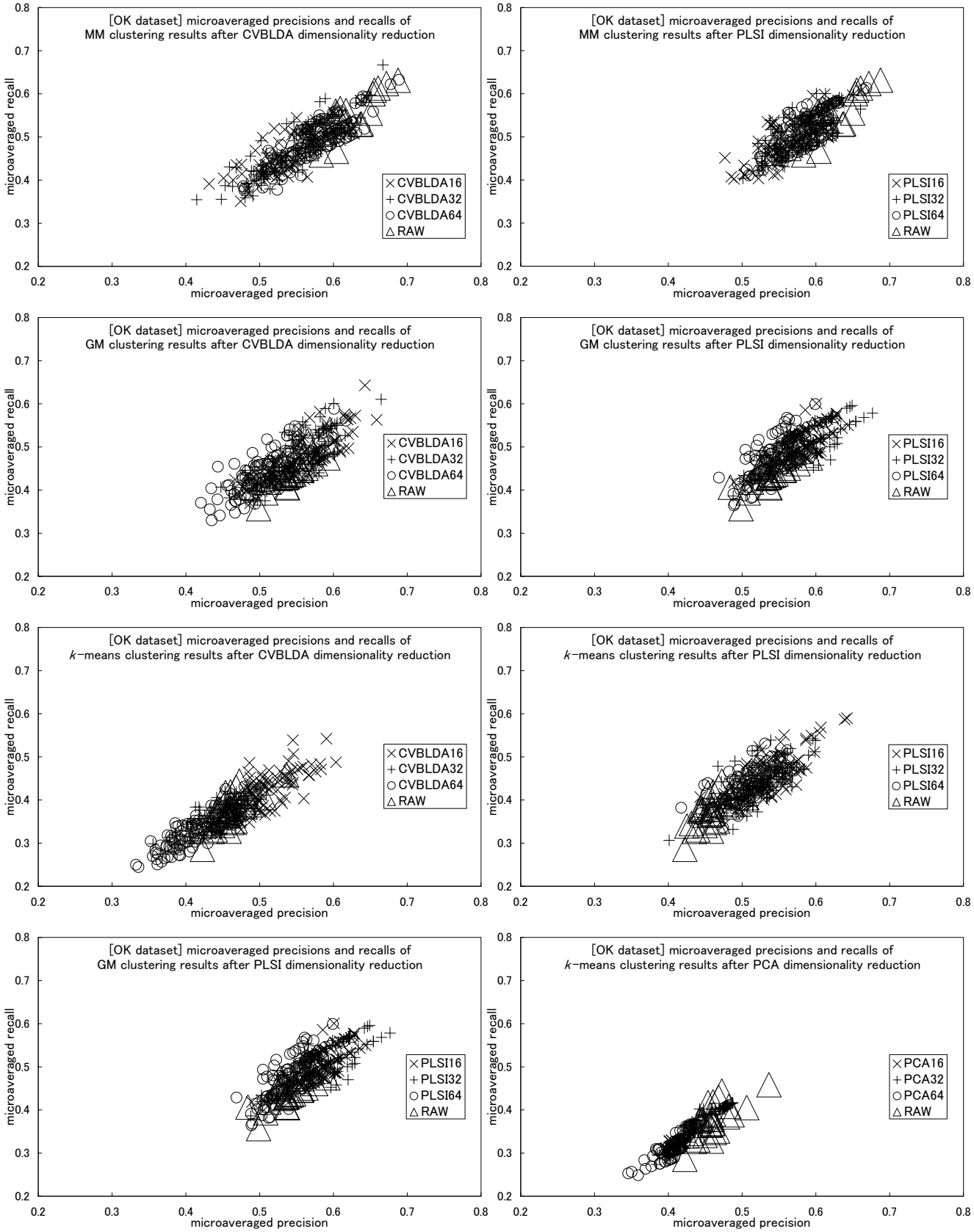

図 2　文書集合 OK について次元圧縮後の文書ベクトルをクラスタリングした場合の評価結果

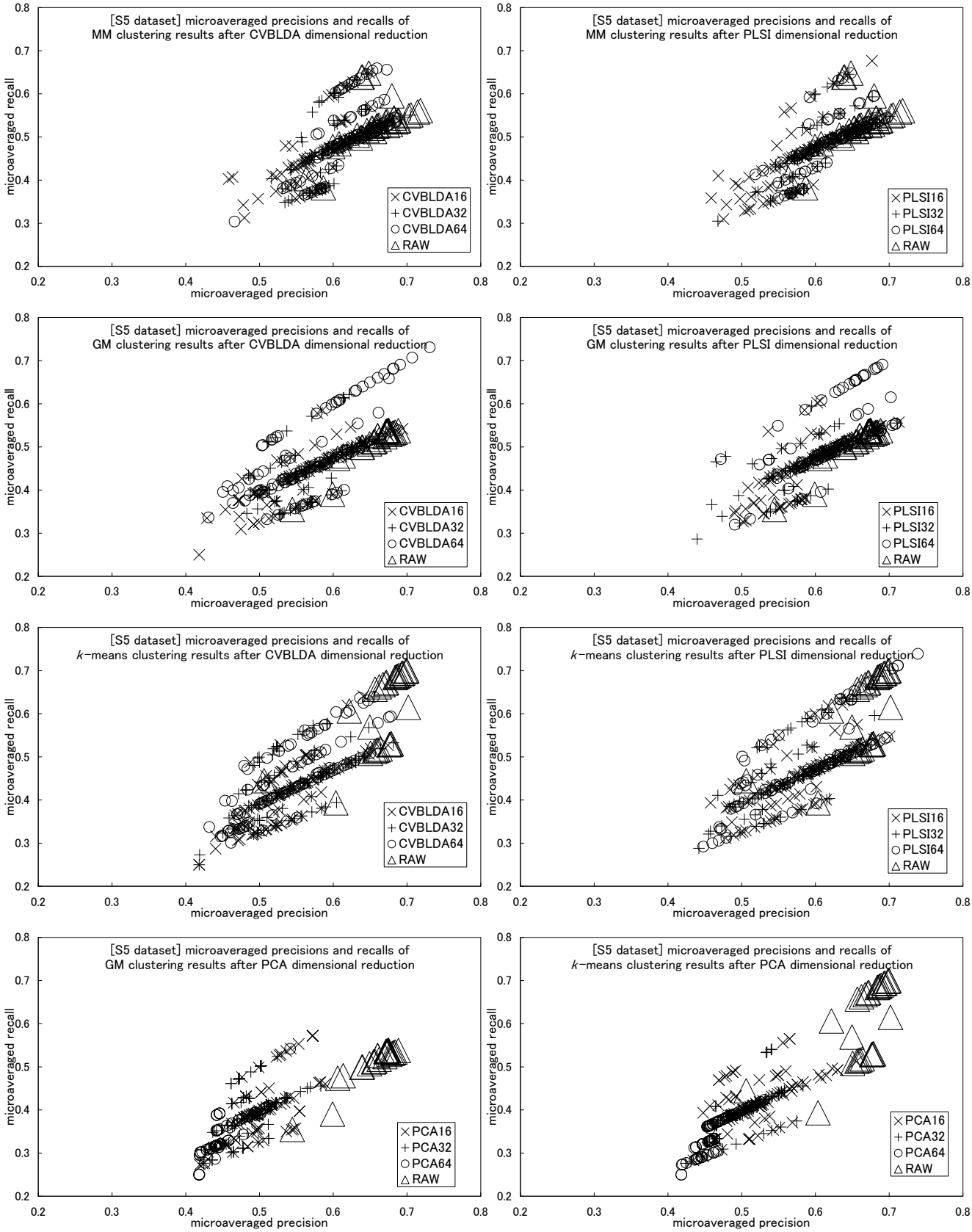

図 3　文書集合 S5 について次元圧縮後の文書ベクトルをクラスタリングした場合の評価結果

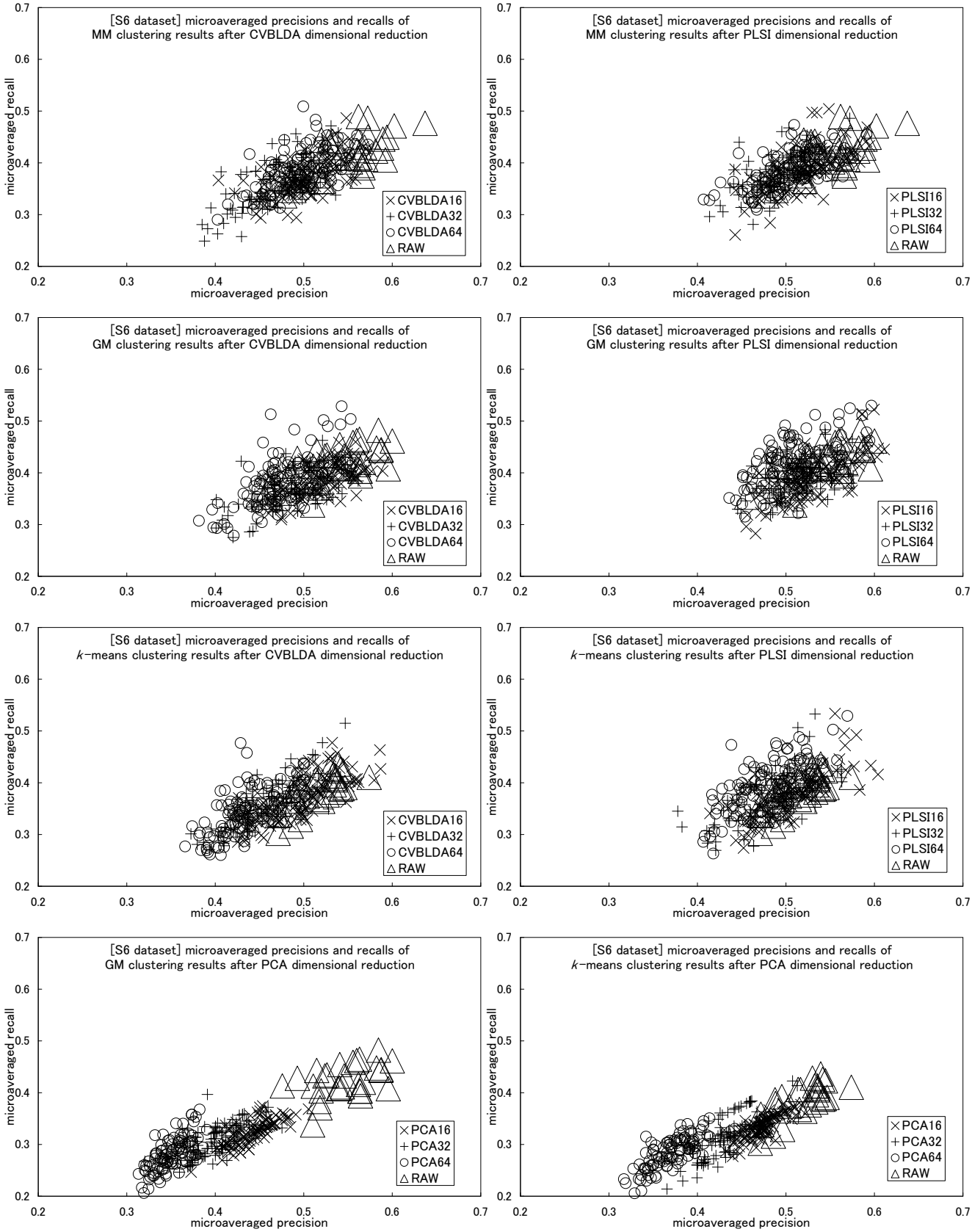

図 4　文書集合 S6 について次元圧縮後の文書ベクトルをクラスタリングした場合の評価結果