Cluster-Preserving Dimension Reduction Methods for Efficient Classification of Text Data

by
Peg Howland and Haesun Park

新納浩幸

Overview

文書クラスタリングを目的に term-document 行列の
次元削減の手法を提案

Centroid QR　　一般的な手法、理論的に整理

Disc GSVD　　判別分析を応用した手法

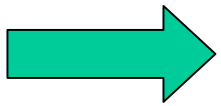
メモ

クラスタリングの手法ではない。
行列の（行＝単語の）次元を減らすだけ。
Nonsingular な行列（単語数＞文書数）にも適用できることがウリ

Dimension Reduction

情報検索で使われる term-document matrix は巨大

（行が term、列が document）



次元を減らす

効率的な処理に繋がる

注意）

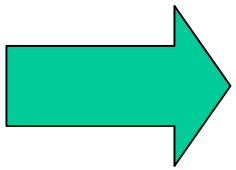
＊term の方の次元を減らした場合、残っている次元の表す物は単純ではない。射影。

＊列の次元を減らすことは、文書クラスタリングそのもの

Centroid QR

行列の分解

$A : m \times n$行列



$A \approx BY$

$$\begin{cases} B : m \times l \\ Y : l \times n \end{cases}$$

Rank の低い行列の積に近似させる

注意）分解は一意ではない

こっちが重要

$$BY = (BZ)(Z^{-1}Y)$$

行列分解の２つのフェイズ

（１） matrix rank reduction formula

階数の低い行列の積に分解する

$$A \approx BY$$

（2） minimization problem

分解された行列の積（変換された行列）の
近似を良くする

$$\min_{B,Y} \| A - BY \|_P$$

← P ノルム

Matrix Rank Reduction Theorem

$A : m \times n$行列　　$rank(A) = r$

$$E = A - (AP_2)(P_1AP_2)^{-1}(P_1A)$$

$$P_1 : l \times m \qquad P_2 : n \times l \qquad l < r$$

$$rank(E) = rank(A) - rank((AP_2)(P_1AP_2)^{-1}(P_1A))$$

P_1AP_2 : nonsingular

事前知識の利用

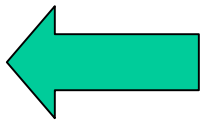
$$P_1 : l \times m \qquad P_2 : n \times l \qquad l < r \qquad P_1 A P_2 : nonsigular$$

$$BY = (AP_2)(P_1 AP_2)^{-1}(P_1 A)$$

となるようにとれば階数を落とした分解が可能

$$\min_{B,Y} \| A - BY \|_P$$

これを実現するようにとる



難しい問題、事前知識を利用

Centroid QR（準備）

$A = [A_1, A_2, \cdots, A_k]$　文書が k クラスに分解されると仮定

$A_i : m \times n_i \quad \sum_{i=1}^{k} n_i = n$　クラス i の要素数は n_i

$c^{(i)} = \frac{1}{n_i} \sum_{j \in N_i} a_j$　クラス i の centroid

$c = \frac{1}{n} \sum_{j=1}^{n} a_j$　全体の centroid

Centroid の性質

$$\sum_{j=1}^{n} \| a_j - c \|_2^2 = \min_x \sum_{j=1}^{n} \| a_j - x \|_2^2 = \min_x \sum_{j=1}^{n} \| A - xe^T \|_F$$

x : n次元列ベクトル

$e = (1,1,\cdots,1)^T$

$\| \cdots \|_F$: フロベニウスノルム

Centroid を使った分解

$$A \longrightarrow BY$$

$$B = C = [c^{(1)}, c^{(2)}, \cdots, c^{(k)}]$$

と取るのが常識的、事前知識

残った問題

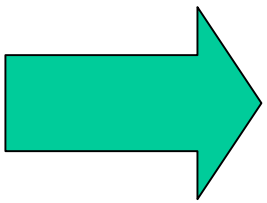
$$\min_{Y} \| A - CY \|_F$$

$$Y : k \times n$$

問題の解

$$H = F \cdot (F^T F)^{-1}$$

$$F : n \times k \qquad F(i, j) = \begin{cases} 1 & \text{文書 i がクラス j に属す} \\ 0 & \text{other} \end{cases}$$



$$C = AH$$

先の最小化問題は解けて

$$Y = (C^T C)^{-1} C^T A$$

B の列が直交していたら

$$B^T B = I$$

$$\Rightarrow \quad A^T A \approx Y^T B^T B Y = Y^T Y$$

$$\Rightarrow \quad A \approx Y$$

B は QR 分解で直交するように変換できる

⇒ Centroid QR
アルゴリズム 1.1

Centroid QR

A の近似行列
行数が削減されている

入力 $A : m \times n$ ⇒ 出力 $Y : k \times n$

(step.1) $c^{(i)}$ の作成

(step. 2) $C = [c^{(1)}, c^{(2)}, \cdots, c^{(k)}]$ $C : m \times k$

(step. 3) C を QR分解 $C = Q_k R$ $Q_k : m \times k$, $R : k \times k$

(step. 4) $\min_Y \| Q_k Y - A \|_F$ を解く ⇒ $Y = Q_k^T A$

DiscGSVD

作用素としての次元削減

文書 m 次元ベクトル → 文書 l 次元ベクトル

A の近似行列
行数が削減されている

$$A : m \times n \rightarrow Y : l \times n$$

$$Y = G^T A$$

$$G : m \times l$$

G を求める

Cluster Quality（1）

すべて
$m \times m$
の行列

within-cluster matrix（クラス内分散行列）

$$S_w = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)})(a_j - c^{(i)})^T$$

between-cluster matrix（クラス間分散行列）

$$S_b = \sum_{i=1}^{k} \sum_{j \in N_i} (c^{(i)} - c)(c^{(i)} - c)^T = \sum_{i=1}^{k} n_i (c^{(i)} - c)(c^{(i)} - c)^T$$

mixture scatter matrix（全体の分散行列）

$$S_m = \sum_{i=1}^{n} (a_j - c)(a_j - c)^T$$

$$S_m = S_w + S_b$$

Cluster Quality（２）

$$trace(S_w) = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)})^T (a_j - c^{(i)}) = \sum_{i=1}^{k} \sum_{j \in N_i} \| a_j - c^{(i)} \|_2^2$$

⇒ クラス内の要素内の近さを表す。小さい方がよい。

$$trace(S_b) = \sum_{i=1}^{k} \sum_{j \in N_i} (c^{(i)} - c)^T (c^{(i)} - c) = \sum_{i=1}^{k} \sum_{j \in N_i} \| c^{(i)} - c \|_2^2$$

⇒ クラス間の分離の度合いを表す。大きい方がよい。

Cluster Quality（３）

$$S_w^Y = G^T S_w G$$
$$S_b^Y = G^T S_b G$$
$$S_m^Y = G^T S_m G$$

$l \times l$

$trace((S_w^Y)^{-1} S_b^Y)$ ← 最大化する G を求める

S_w が singular だと簡単ではない

$m > n$　単語数が文書数よりも大きいと singular

従来手法の問題

S_w が singular → ２段階処理

（１）LSI 等により削減後の次元数を決める

（２）Cluster Quality で G を決める

（問題点）LSI が決める次元数には理論的な裏付けがない

（提案手法）削減後の次元数までも同時に求める。
理論的。

DiscGSVD

GSVD Theorem（1）

$$K = \begin{pmatrix} K_A \\ K_B \end{pmatrix} \qquad K_A : n \times m \quad K_B : p \times m$$

$$K : (n+p) \times m \qquad t = rank(K)$$

以下の4つの直交行列がある

$$U : n \times n \quad V : p \times p \quad W : t \times t \quad Q : m \times m$$

ただし

$$U^T K_A Q = \Sigma_A (W^T R, 0) \qquad n \times m$$

$$V^T K_B Q = \Sigma_B (W^T R, 0) \qquad p \times m$$

GSVD Theorem（２）

$R : t \times t$　　K の固有値からなる対角行列

$$\Sigma_A = \begin{pmatrix} I_A & & \\ & D_A & \\ & & O_A \end{pmatrix} \qquad \Sigma_B = \begin{pmatrix} O_B & & \\ & D_B & \\ & & I_B \end{pmatrix}$$

$$\Sigma_A : n \times t \qquad \Sigma_B : p \times t$$

$$I_A : r \times r \qquad I_B : (t-r-s) \times (t-r-s)$$

$$D_A = diag(\alpha_{r+1}, \alpha_{r+2}, \cdots, \alpha_{r+s}) \qquad D_B = diag(\beta_{r+1}, \beta_{r+2}, \cdots, \beta_{r+s})$$

$$O_A : (n-r-s) \times (n-r-s) \qquad O_B : (p-t+r) \times r$$

GSVD Theorem（3）

$$r = rank\begin{pmatrix} K_A \\ K_B \end{pmatrix} - rank(K_B)$$

$$s = rank(K_A) + rank(K_B) - rank\begin{pmatrix} K_A \\ K_B \end{pmatrix}$$

$$1 > \alpha_{r+1} \geq \alpha_{r+2} \geq \cdots \geq \alpha_{r+s} > 0$$

$$0 < \beta_{r+1} \leq \beta_{r+2} \leq \cdots \leq \beta_{r+s} < 1$$

$$\alpha_i^2 + \beta_i^2 = 1 \quad (i = r+1, r+2, \cdots, r+s)$$

G を求める

$$J_1(G) = trace((G^T S_2 G)^{-1}(G^T S_1 G))$$

$J_1(G)$ を最大にする G を求める

$G : m \times l$

メモ）

$$S_1 = S_b \qquad S_2 = S_w$$

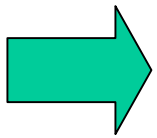
と想定していて良い、論文では特異なケースに
対処するために一般的に扱っている

S2 が nonsingular の場合（1）

$S_2^{-1}S_1$ の固有値と固有ベクトルを λ_i x_i

$$X = [x_1, x_2, \cdots, x_m]$$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q > \lambda_{q+1} = \cdots = \lambda_m = 0$$



$$X^T S_1 X = diag(\lambda_1, \lambda_2, \cdots, \lambda_m)$$

$$X^T S_2 X = I_m$$

S2 が nonsingular の場合（2）

$$J_1(G) = trace((G^T S_2 G)^{-1}(G^T S_1 G))$$
$$= trace((G^T X^{-T} X^{-1} G)^{-1}(G^T X^{-T} \Lambda X^{-1} G))$$
$$= trace((\tilde{G}^T \tilde{G})^{-1}(\tilde{G}^T \Lambda \tilde{G}))$$

$$\tilde{G} = X^{-1} G \qquad \tilde{G} : m \times l$$

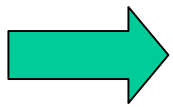
$\tilde{G} = QR$　QR分解　$Q : m \times l$

$$J_1(G) = trace(Q^T \Lambda Q)$$

S2 が nonsingular の場合（3）

$$J_1(G) = trace(Q^T \Lambda Q)$$

$$\leq \lambda_1 + \lambda_2 + \cdots + \lambda_q = trace(S_2^{-1} S_1)$$



$G = X \begin{pmatrix} I_l \\ 0 \end{pmatrix}$ のとき最大

$X : m \times m$

$G : m \times l$

実際、このとき $J_1(G) = trace(S_2^{-1} S_1)$

$S_2^{-1} S_1$ の固有ベクトルを l 個並べた行列

$$G = [x_1, x_2, \cdots, x_l]$$

Sw が singular の場合（1）

$$S_w = H_w H_w^T$$
$$S_b = H_b H_b^T$$
$$S_m = H_m H_m^T$$

$$A = [A_1, A_2, \cdots, A_k]$$

$$H_w = [A_1 - c^{(1)} e^{(1)^T}, A_2 - c^{(2)} e^{(2)^T}, \cdots, A_k - c^{(k)} e^{(k)^T}] \qquad m \times n$$

列ベクトルからクラスの centorid をひいたものを並べた行列

$$H_b = [(c^{(1)} - c) e^{(1)^T}, (c^{(2)} - c) e^{(2)^T}, \cdots, (c^{(k)} - c) e^{(k)^T}] \qquad m \times n$$

クラスの centorid から全体の centorid ひいたものを並べた行列

$$H_m = [a_1 - c, a_2 - c, \cdots, a_n - c] = A - ce^T \qquad m \times n$$

列ベクトルから全体の centorid をひいたものを並べた行列

Sw が singular の場合（2）

$(S_1, S_2) = (S_b, S_w)$ で考える

$rank(S_b) \le k - 1$ ⟸ $S_b = H_b H_b^T$

$J_1(G) = trace((G^T S_w G)^{-1}(G^T S_b G))$ の最大化

⟹

$$\max_G trace(G^T S_b G)$$

$$\min_G trace(G^T S_w G)$$

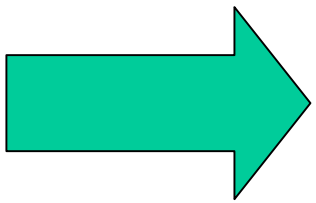
Sw が singular の場合（3）

間違っているかも。私なりの理解を示します。

(S_b, S_w)　ではなく (H_b, H_w)で考える

こっちの方が次元が小さい

$S_b = H_b H_b^T$　　$S_w = H_w H_w^T$　の関係



$$\begin{cases} \max_G trace(G^T H_b G) \\ \min_G trace(G^T H_w G) \end{cases}$$

Sw が singular の場合（４）

$$trace(G^T H_b G) = \alpha_1 + \alpha_2 + \cdots + \alpha_r + \alpha_{r+1} + \cdots + \alpha_{r+s} + \alpha_{r+s+1} + \cdots + \alpha_t$$
$$= \alpha_1 + \alpha_2 + \cdots + \alpha_r + \alpha_{r+1} + \cdots + \alpha_{r+s}$$

$$trace(G^T H_w G) = \beta_1 + \beta_2 + \cdots + \beta_r + \beta_{r+1} + \cdots + \beta_{r+s} + \beta_{r+s+1} + \cdots + \beta_t$$
$$\geq \beta_1 + \beta_2 + \cdots + \beta_r + \beta_{r+1} + \cdots + \beta_{r+s}$$
$$= \beta_{r+1} + \cdots + \beta_{r+s}$$

$$1 > \alpha_{r+1} \geq \alpha_{r+2} \geq \cdots \geq \alpha_{r+s} > 0$$

$$0 < \beta_{r+1} \leq \beta_{r+2} \leq \cdots \leq \beta_{r+s} < 1$$

$$\alpha_i^2 + \beta_i^2 = 1 \qquad (i = r+1, r+2, \cdots, r+s)$$

Sw が singular の場合（５）

$$\max_{G} trace(G^T H_b G)$$
$$\min_{G} trace(G^T H_w G)$$

を実現するには、

$\lambda_i = \frac{\alpha_i^2}{\beta_i^2}$　の大きな i を選び

対応する　x_i　からなる行列Gを作る

いくつえらぶか？ 大きくするにはできるだけたくさん。

$r + s = rank(H_b) \leq k - 1$

Sw が singular の場合（６）

α_i, β_i に対応する x_i とは？

$$U^T K_A X = (\Sigma_A, 0) \qquad V^T K_B X = (\Sigma_B, 0)$$

$$X = Q \begin{pmatrix} R^{-1}W & 0 \\ 0 & I \end{pmatrix}$$

GSVD の定理から示せるが、難しい

X の作り方は DiscGSVD のアルゴリズムで示される

DiscGSVD（1）

行数を m から k-1 に削減する

入力 $A : m \times n$ ⇒ 出力 $G : m \times (k-1)$

(step.1)

$$H_w = [A_1 - c^{(1)} e^{(1)^T}, A_2 - c^{(2)} e^{(2)^T}, \cdots, A_k - c^{(k)} e^{(k)^T}] \qquad H_w : m \times n$$

$$H_b = [\sqrt{n_1}(c^{(1)} - c), \sqrt{n_2}(c^{(2)} - c), \cdots, \sqrt{n_k}(c^{(k)} - c)] \qquad H_b : m \times k$$

次元が違う！計算の効率化

(step.2)

$$K = \begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix} \qquad P^T K Q = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}$$

（GSVD の定理より）

$$K : (k+n) \times m$$

(step.3) $t = rank(K)$

DiscGSVD（2）

$P(1:k,1:t)$ の SVD

(step.4)

$$U^T P(1:k,1:t) W = \Sigma_A$$

Pの1列目から k 列目まで並べ、
更に1列目から t 列目まで並べた行列

(step.5)

$$X = Q \begin{pmatrix} R^{-1}W & 0 \\ 0 & I \end{pmatrix}$$

の最初の k-1 個の列ベクトルから
G を作成する

(step.6)

$$Y = G^T A$$

Y は A の近似

実験

実験（nonsingular のケース）

データ

単語数＜文書数

人工的に作られた文書データ

2000文書、次元数（単語の種類数）150、クラスの数 7

利用した手法

DiscGSVD　次元を 6 or 7 に削減

得られた行列に対して

- Centorid 手法
- K-NN手法

クラスタリングの手法

$k=5$　$k=15$

実験結果（1）

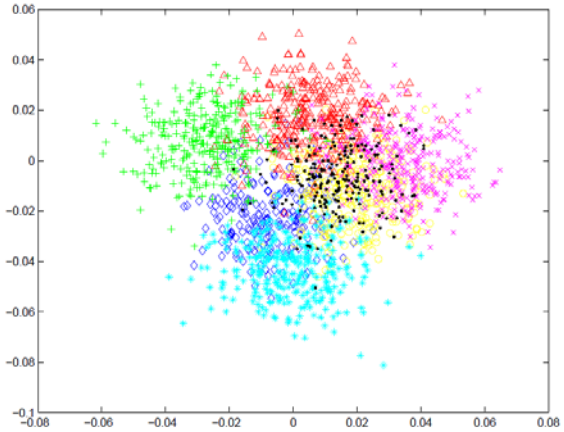
Method	Full	trace($S_w^{-1}S_b$)	trace($S_w^{-1}S_m$)	
Dim	150 × 2000	6 × 2000	6 × 2000	7 × 2000
trace(S_w)	299700	1.97	1.48	1.98
trace(S_b)	22925	4.03	3.04	3.04
trace(S_m)	322630	6.00	4.52	5.02
trace($S_w^{-1}S_b$)	12.6	12.6	12.6	12.6
trace($S_w^{-1}S_m$)	162.6	18.6	18.6	19.6
centroid	2.6 %	2.2 %	2.0 %	2.0 %
5nn	18.7 %	2.2 %	2.2 %	2.4 %
15nn	10.1 %	1.8 %	1.9 %	2.1 %

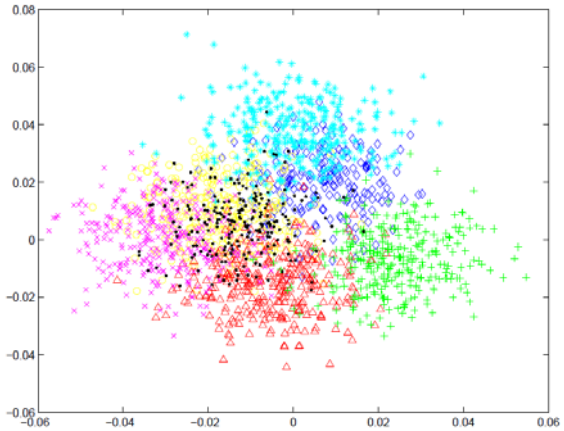
クラスタリング手法

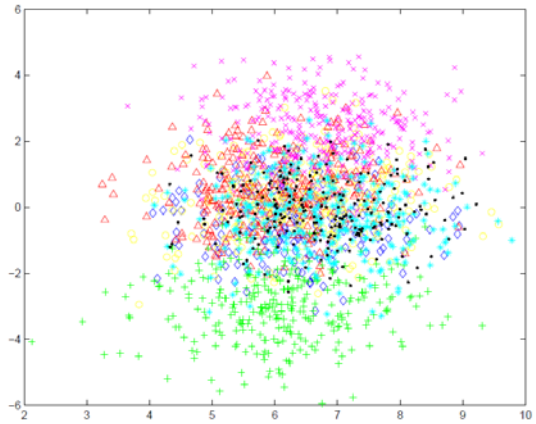
誤り率

たいした差はない

2次元まで落としてみると、、、







$S_w^{-1}S_b$

$S_w^{-1}S_m$

SVD

○

×

実験（singular のケース）

データ

MEDLINE 内の 200 文書

5カテゴリ、各カテゴリで 40文書、計200文書
次元数（単語の種類数）7519

単語数＞文書数

利用した手法

CentroidQR　次元を 5 に削減

DiscGSVD　次元を 4 に削減

得られた行列に対して

- Centorid 手法
- K-NN手法　$k=1$

クラスタリングの手法

実験結果（2）

Method		Full	CentroidQR	DiscGSVD
Dim		7519×200	5×200	4×200
trace values	$\mathrm{trace}(S_w)$	73048	4210	0.05
	$\mathrm{trace}(S_b)$	6229	6229	3.95
	$\frac{\mathrm{trace}(S_b)}{\mathrm{trace}(S_w)}$	0.09	1.5	79
misclassification rate in %	centroid	5	5	1
	1nn	40	3	1

誤り率

クラスタリング手法

改善されている！

まとめ（メモ）

単語数＞文書数　となっているケースで威力を発揮する
文書クラスタリングを目的とした次元削減手法

ただし、GSVD のアルゴリズムが不明なので実装できない

$K = \begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix}$ に対して $P^T K Q = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}$ となる

P, Q, R を計算する具体的な手順が不明

クラスの centorid の与え方については言及なし

（現在、私が研究している内容、、、初期値の与え方）