2D-4

Sequential Co-clustering を用いた行列近似手法とその定量的分析

稲垣 隆史 † ダヌシカ ボレガラ ‡ 石塚 満 ††
† 東京大学工学部電子情報工学科 ‡ 東京大学大学院情報理工学系研究科 †† 東京大学大学院情報理工学系研究科

1 はじめに

1.1 背景

行列の低 rank 行列による近似は，自然言語処理においてノイズの削減や同義語・多義語問題の解消に役立っている．一般に知られている行列近似手法として特異値分解 (Singular Value Decomposition:SVD) を用いたものがある．この手法は行列を表す寄与率の高い基底を求め，元行列との二乗誤差を最小とする低 rank 近似行列を得ることができるが，計算量は $O(N^2)$〜$O(N^3)$ となる．一方クラスタリングは本来，データの分類に用いられるもので，行列近似への利用は研究されてこなかった．本研究は，クラスタリングで得られたクラスタの特徴を行列近似に利用し，その近似精度を検証するものである．

1.2 特異値分解を用いた行列近似手法

任意の実 $m \times n$ 行列 D の特異値分解は，次のように定義される．

$$D = U\Sigma V^T$$

U は $m \times m$ 直交行列，V は $n \times n$ 直交行列である．σ は $m \times n$ 行列であり，$r = rank(D)$ としたとき，対角線上に $\sigma_1, \sigma_2, ..., \sigma_r (\sigma_1 \geq ... \geq \sigma_r > 0)$ が並び，その他の成分は 0 となる．ここで $\sigma_i (i = 1, 2, ..., r)$ を行列 D の特異値といい，U の列ベクトル $u_1 ... u_m$ および V の列ベクトル $v_1 ... v_n$ の，行列 D を表現する基底としての重要度を表す．また，$DD^T = U\Sigma^2 U^T$ および $D^T D = V\Sigma^2 V^T$ で示されるように，$DD^T, D^T D$ の固有値の平方根となる．特異値分解を用いた D の近似行列 $D_k (rank(D_k) = k)$ は，U，V，Σ それぞれの k 列目までを残し，$k+1$ 列目以降を 0 とした行列 U_k，V_k，Σ_k を用いて，

$$D_k = U_k \Sigma_k {V_k}^T$$

により求められる．また，U の列ベクトルを基底として，次のように求めることもできる．

$$D_k = U_k {U_k}^T D$$

1.3 クラスタリングを用いた行列近似

自然言語処理における一般的なクラスタリングは，1 文書がベクトルの形で表された行列データを分類し，結果として得られたクラスタがそれぞれ異なる特徴を持つようにするものである．例えば，ニュース記事の文書群をクラスタリングしたとき，あるクラスタにはスポーツ関連の文書，別のクラスタには芸術関連の文書が属するようにする．各クラスタの特徴となるベクトルは，行列を表現する基底として有効であると考えられる．行列近似の際には，クラスタの特徴となるベクトルをどのように定義するかということと，多数のクラスタの特徴ベクトルのうち，どのベクトルを基底として採用するかが重要となる．

2 手法

2.1 手法の概要

クラスタリングによって得られた各クラスタにおいて，クラスタ内の文書ベクトルの重心ベクトルを，クラスタの特徴ベクトルとする．これらの重心ベクトルからどのベクトルを基底として採用するかを選ぶ際には，SVD における特異値のような値が必要である．本研究では，その値 (擬似特異値) として，3 つの尺度を考案した．その尺度を元に，目的 rank の個数のベクトルを選ぶ．また，基底ベクトルは互いに直交していることが望ましいため，採用したベクトルに対して直交化を行う．直交化された基底を用いて，行列近似を行い，低 rank 行列を得る．

本研究では，classic3 dataset[1] 内の文書 j における単語 i の出現頻度を (i, j) 成分とする単語文書行列をデータとして用いた．行列の大きさは 7399 × 999 であった．

クラスタリングには，D.Bollegala らの Sequential Co-clustering[1] を用いた．このクラスタリングアルゴリズムは，行と列を交互にクラスタリングしていくもので，閾値を変化させることで結果クラスタ数を調整することができ、平均計算量は $O(NlogN)$ となる。

2.2 手法の手順

1. 入力された行列に対し，クラスタリングを実行する．
2. 結果として得られた文書クラスタそれぞれについて，重心ベクトルを求める．
3. 各クラスタの重心ベクトルc_iに対して，次のように擬似特異値を設定する．この際，特異値の定義である，DD^T の固有値に近いものに高い擬似特異値を設定するようにした．
   (a) $|DD^T c_i| = |\lambda c_i|$ となる λ
   (b) DD^T と c_i の内積
   (c) DD^T と c_i のコサイン尺度
4. 擬似特異値の高い重心ベクトルから順に，入力された目的の rank$(= k)$ 個のベクトル $b_1, ..., b_k$ を選び，グラムシュミットの正規直交化法を適用する．
5. $B = (b_1 ... b_k)$ を用いて，近似行列 $D_k = BB^T D$ を得る．

3 実験

まず，クラスタリング結果を行列近似手法に利用できるか判断するために，先行実験を行った．先行実験では，SVD による手法と Sequential Co-clustering を用いた手法を，ベースラインと比較した．擬似特異値としてどの尺度を用いれば良いかの判定，および提案手法の有用性を評価するために，本実験を行った．本実験では，SVD を用いた行列近似手法，および提案手法について，元行列との誤差を測定した．

3.1 先行実験の手順

1. 目的 rank を 3 とし，次の基底を用いてそれぞれ近似行列を作成した．
   (a) SVD を用いた手法により得られた特異値が大きい上位 3 つのベクトル

The quantitative analysis of matrix approximation method using Sequential Co-clustering

†Takashi INAGAKI, Faculty of Engineering, University of Tokyo
‡Danushka Bollegala, Graduate School of Information Science and Technology, University of Tokyo
††Mitsuru ISHIZUKA, Graduate School of Information Science and Technology, University of Tokyo

[1] http://www.dataminingresearch.com/index.php/2010/09/classic3-classic4-datasets/

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.

(b) 得られるクラスタ数が 3 になるように調整して Sequential Co-clustering を実行し，得られた 3 つのクラスタの重心ベクトルにグラムシュミットの正規直交化法を適用したもの

(c) 元行列からランダムに 3 つ選んだ列ベクトル

2. それぞれの近似行列について，次の式で表される元行列との誤差 (Root Mean Square Error:RMSE) を計算した．ただし d_i，d'_i はそれぞれ元行列，近似行列の i 番目の列ベクトルであり，n は元行列の列の個数である．

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} |d_i - d'_i|^2}{n}}$$

(c) については，ランダムに 3 つ選び，近似行列を作成して RMSE を計算するという手順を 30 回繰り返したものの平均を，誤差とした．

3.2 先行実験の結果

各手法の RMSE を表 1 に示す．この結果から，Sequential Co-clustering をグラムシュミットの正規直交化法と共に用いることで，高い精度で行列近似が行えることが分かった．

表 1: 各手法の近似誤差

手法	RMSE
SVD	14.29
Sequential Co-clustering	14.4
ランダム	15.09

3.3 本実験の手順

1. SVD，および擬似特異値の尺度の異なる 3 つの提案手法を用いて，それぞれ近似行列を作成した．
2. 近似行列それぞれに対して，元行列との RMSE を計算した．
3. 1,2 の過程を，目的の rank を変化させて繰り返した．SVD における手法については 1 から 999 まで変化させ，提案手法においては 1 から結果クラスタ数まで変化させて近似した．

3.4 本実験の結果

各手法について，目的の rank を横軸に，近似行列の RMSE を縦軸にとったグラフを図 1 に示す．グラフの曲線は下から，

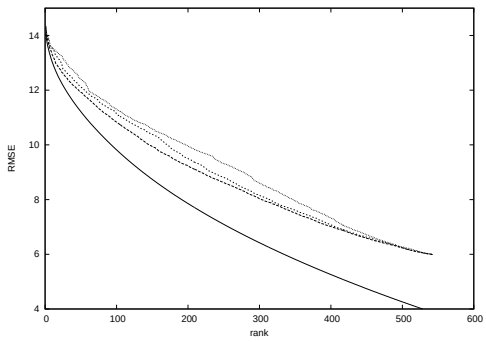


図 1: 各行列近似手法の RMSE

SVD による手法，擬似特異値としてコサイン尺度を用いた手法，内積を用いた手法，ノルム比を用いた手法である．提案手法の中では，擬似特異値として DD^T と c_i のコサイン尺度を用いた手法が最も近似誤差が少ないことが分かる．また，曲線の形状を見ると，ほぼ下に凸である．このことより，行列を表すのに良い基底が優先して選ばれていることが分かる．

4 結論

本稿においては，Sequential Co-clustering を用いた行列近似手法について述べた．また，行列近似の近似誤差について評価実験を行い，高い近似精度が得られることが分かった．実行時間に関しては圧倒的に提案手法が高速であった．なお，今回はクラスタリングアルゴリズムとして Sequential Co-clustering を用いたが，他のクラスタリングアルゴリズムの結果も行列近似に利用できると考えられる．その点の検証，および提案手法による近似行列が情報検索などの応用において有効であるかの検討は，今後の課題としたい．

5 謝辞

本研究を進めるにあたり、指導教官の石塚満先生、および様々なご指導を頂きましたダヌシカ ボレガラ先生に深謝いたします。また、日常の議論を通じて多くの知識や示唆を頂いた研究室の皆様に感謝します。

参考文献

[1] Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web In *Proc. of the 19th International World Wide Web Conference (WWW 2010)*, pages 151-160, 2010.

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.