ARG WI2 No.11, 2017
WI2-2017-31

Word2Vecによる次元圧縮と重回帰分析型協調フィルタリングへの応用

藤井 流華[†,a]　　岡本 一志[††,b]

†電気通信大学 情報理工学部 総合情報学科　††電気通信大学 大学院情報理工学研究科 情報学専攻

a) *f1410112@edu.cc.uec.ac.jp*　b) *kazushi@uec.ac.jp*

概要　モデルベース協調フィルタリングを重回帰分析により実現することを目指し，ユーザ-アイテム行列をWord2Vecにより次元圧縮し圧縮空間で回帰係数を推定する手法を提案する．協調フィルタリング用ベンチマークデータセットを用いた実験により，提案手法の予測性能と適切なWord2Vecのハイパーパラメータを明らかにする．具体的には，重回帰分析ではL1正則化またはL2正則化を適用し，圧縮次元数10通り，ウィンドウサイズ5通りの計50通りのWord2Vecのハイパーパラメータにおける予測精度を検証する．実験結果から，提案手法にはL2正則化が適しており，Word2Vecの次元数は小さく，ウィンドウサイズは大きくするほうが予測精度の向上に寄与することを確認している．これらの結果より，推薦理由の説明能力を有したモデルベース協調フィルタリングを重回帰分析により実現できる可能性を得ている．

キーワード　協調フィルタリング，重回帰分析，次元圧縮，Word2Vec，ブートストラップ法

1　はじめに

情報推薦システムは，アイテム（商品や店舗，記事など）の内容や特徴，好みの度合いなどからユーザが好みそうなアイテムを予測し提示するシステムである．協調フィルタリング[1]はその実現手法のひとつであり，購入履歴やアイテムに付与されたスコアなどから，類似ユーザの発見やアイテムに付与されるスコアを予測する．

本研究では，協調フィルタリングへの重回帰分析の応用により，目的とするアイテムのスコアの予測だけでなく，予測に影響を与える他のアイテムも推定する技術を開発する．重回帰分析には変数の数に応じたデータ数が必要であり，協調フィルタリングで扱うユーザ-アイテム行列のように次元が高く疎性の強いデータに対しては回帰係数を推定できない可能性がある．この課題の解決のため，本研究では，単語のベクトル表現を計算するWord2Vec[2]を用いてユーザ-アイテム行列の次元を圧縮し圧縮空間で回帰係数を推定する手法を提案する．

協調フィルタリング用ベンチマークデータセットを用いた実験により，提案手法の予測性能や適切なハイパーパラメータの設定法を明らかにする．具体的には，学習データとテストデータをブートストラップ法により作成し，適用する正則化法とWord2Vecの適切なハイパーパラメータの選択について，アイテムのスコアの予測精度の観点で検証する．

2　関連研究

情報推薦システムにおける推薦理由の説明は，推薦の受け入れられやすさやシステムへの信頼性[3]，迅速な意思決定やシステムの利用満足度[4]などに寄与することが知られている．協調フィルタリングのアプローチとして，メモリベース法とモデルベース法の2つがある．メモリベース法は，ユーザやアイテムの類似関係からスコアを予測する手法であり，逐次全てのユーザやアイテムについて近傍探索を行うため推薦処理の計算コストが高い課題がある．協調フィルタリングにおける推薦理由の説明性の研究は，メモリベース法が主流である．また，モデルベース法は，学習アルゴリズムに基づいてスコアの予測モデルを構築する手法であり，学習処理が必要なものの，推薦処理自体の計算コストは低い特徴がある．モデルベース協調フィルタリングで用いられる手法として，行列因子分解[5]やアソシエーションルール分析[6]，ベイジアンネットワーク[7]などがある．しかし，行列因子分解には推薦理由の説明性がなく，アソシエーションルール分析はアイテムのスコア予測ができない．ベイジアンネットワークにはアイテムのスコア予測も推薦理由の説明性もあるが，学習処理の計算コストが極めて高い課題がある．その一方で，多変量解析のひとつである重回帰分析はモデルベースのアプローチであり，目的変数の値の予測だけでなく，回帰係数から目的変数に影響を与えている説明変数を調べることができるため，予測に対する説明性がある手法といえる．

3　Word2Vecによる次元圧縮と重回帰分析

本研究では，協調フィルタリングに重回帰分析を応用することを試みる．重回帰分析の応用により，学習データからアイテムのスコアの予測モデルを回帰式により構築することができ，得られた回帰係数から他のアイテムが目的のアイテムのスコアに与える影響を調べることが

表 1 ユーザ-アイテム行列 $\boldsymbol{X}$ の例

	アイテム 1	アイテム 2	アイテム 3	⋯
ユーザ A	10	–	5	
ユーザ B	7	–	6	
ユーザ C	–	2	–	
⋮	⋮	⋮	⋮	

表 2 表 1 を 2 値化した例（$\boldsymbol{X}_b$）

	アイテム 1	アイテム 2	アイテム 3	⋯
ユーザ A	1	0	1	
ユーザ B	1	0	1	
ユーザ C	0	1	0	
⋮	⋮	⋮	⋮	

できる．しかしながら，重回帰分析では変数（アイテム）の数に対して学習に用いるデータの数が 10 倍程度必要とされている [8]．協調フィルタリングでは，ユーザがアイテムに付与したスコアを行列で表現したユーザ-アイテム行列（表 1）が扱われる．情報推薦システムが扱うユーザやアイテムの数は増大傾向にあり，ユーザが付与できるスコアは全体のごく一部のアイテムに限られるため，一般に，ユーザ-アイテム行列は巨大な疎行列となる．このことは，行列の次元が高く学習に使えるデータが少なくなる傾向を示唆しており，ユーザ-アイテム行列を直接重回帰分析すると回帰係数が推定できない可能性がある．この課題の解決法として，正則化と次元圧縮のアプローチがある．巨大なユーザ-アイテム行列への正則化の適用は計算コストが高くなることが想定されるため，本研究では次元圧縮の適用を検討する．

3.1 Word2Vec による次元圧縮

次元圧縮のアプローチとして，自然言語処理技術のひとつである Word2Vec[2] を活用する．Word2Vec は，評価されているデータのみを用いて低次元への線形写像を学習できるため計算コストが低く，巨大なユーザ-アイテム行列の低計算コストでの次元圧縮を期待できる．

Woed2Vec による次元圧縮では，まず，学習用ユーザ-アイテム行列 $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ から Word2Vec 学習用ユーザーアイテム行列 $\boldsymbol{X}_b \in \{0,1\}^{m \times n}$ を作成する．$\boldsymbol{X}_b$ は $\boldsymbol{X}$ をアイテムが評価されているかどうかで 2 値化（1: 評価した，0: 評価していない）したものであり，m はユーザ数，n はアイテム数とする．表 2 に表 1 のユーザ-アイテム行列 $\boldsymbol{X}$ から $\boldsymbol{X}_b$ を作成する例を示す．次に，$\boldsymbol{X}_b$ の中から値が 1 のデータを集め，

ユーザ 1: { アイテム 1，アイテム 3，⋯}
ユーザ 2: { アイテム 1，アイテム 3，⋯}
ユーザ 3: { アイテム 2，⋯}

のようなトランザクションデータを生成し，各ユーザのアイテムの順番をランダムシャッフルする．これらのデータを Word2Vec に入力し，$\boldsymbol{X}_b$ から線形写像 $\boldsymbol{W} \in \mathbb{R}^{n \times k}$ を算出する．ここで，k は圧縮次元数（$k < n$）とする．

3.2 重回帰分析による予測モデルの構築

$\boldsymbol{X}$ と $\boldsymbol{W}$ のそれぞれから目的変数に該当する列または行（i 番目のアイテム）を除いた $\boldsymbol{X}_i \in \mathbb{R}^{m \times (n-1)}$ と $\boldsymbol{W}_i \in \mathbb{R}^{(n-1) \times k}$ を作成し，$\boldsymbol{X}'_i = \boldsymbol{X}_i \boldsymbol{W}_i$ により $\boldsymbol{X}_i$ の k 次元表現を得る．重回帰分析による予測モデルは

$$\boldsymbol{y}_i = \boldsymbol{\alpha}_0 + \boldsymbol{X}_i \boldsymbol{\alpha}$$

であり，$\boldsymbol{y}_i$ と $\boldsymbol{X}_i$ から $\boldsymbol{\alpha}_0$ と $\boldsymbol{\alpha}$ を推定する．このとき，$\boldsymbol{y}_i$ は m ユーザ分の目的変数のアイテムのスコアとする．また，圧縮空間での重回帰分析の予測モデルは

$$\boldsymbol{y}_i = \boldsymbol{\beta}_0 + \boldsymbol{X}'_i \boldsymbol{\beta}$$

であり，$\boldsymbol{y}_i$ と $\boldsymbol{X}'_i$ から $\boldsymbol{\beta}_0$ と $\boldsymbol{\beta}$ を推定する．ここで，

$$\boldsymbol{y}_i = \boldsymbol{\beta}_0 + \boldsymbol{X}'_i \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \boldsymbol{X}_i \boldsymbol{W}_i \boldsymbol{\beta}$$

であるため，$\boldsymbol{\alpha}_0 = \boldsymbol{\beta}_0$，$\boldsymbol{\alpha} = \boldsymbol{W}_i \boldsymbol{\beta}$ とすることで圧縮空間で重回帰分析をして得られた回帰係数 $\boldsymbol{\beta}$ から直接重回帰分析をして得られる回帰係数 $\boldsymbol{\alpha}$ を計算できる．そして，構築したスコア予測モデル

$$\hat{y}_i = \alpha_0 + \boldsymbol{x}\boldsymbol{\alpha}$$

に，予測したいユーザのスコアベクトル $\boldsymbol{x} \in \mathbb{R}^{n-1}$ を入力することにより，予測値 $\hat{y}_i$ を求めることができる．

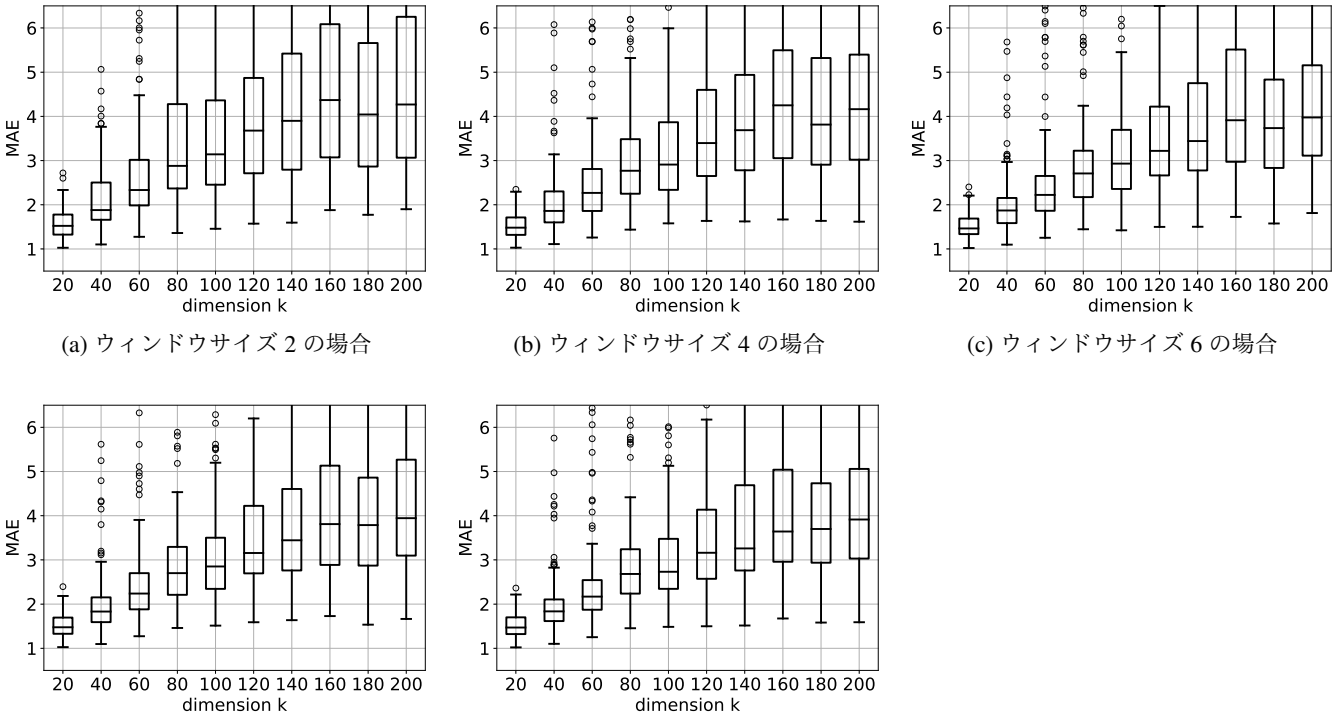
4 次元圧縮が予測精度に与える影響の評価

圧縮空間での重回帰分析を経て元の空間の回帰係数を推定する提案法の予測性能を明らかにするため，協調フィルタリング用ベンチマークデータセットを用いた評価実験を行う．Word2Vec を用いた次元圧縮では，定義より圧縮空間での変数間に相関が出ることが予想されるため，重回帰分析に L1 正則化および L2 正則化を適用する．併せて，Word2Vec のハイパーパラメータが与える影響についても検証する．なお，重回帰分析の計算には R の glmnet パッケージを使用する．

本研究では，協調フィルタリング用ベンチマークデータセットとして，Book-Crossing データセットを用いる．このデータセットはユーザが書籍に付与した 1〜10 までの 10 段階のスコアを集計したものであり，ユーザ数 278,858，書籍数 271,379，総スコア数 383,852 である．

4.1 予測精度の評価法

評価しているユーザ数が多い書籍上位 100 件を目的変数とし，残りの全書籍を説明変数として選択する．目的変数として選択した各変数について，それぞれ回帰式を立てることとする．なお，学習処理では目的変数の書籍を評価しているユーザのデータのみを用い，説明変数に



(a) ウィンドウサイズ 2 の場合　(b) ウィンドウサイズ 4 の場合　(c) ウィンドウサイズ 6 の場合

(d) ウィンドウサイズ 8 の場合　(e) ウィンドウサイズ 10 の場合

図 1 L1 正則化を適用した予測精度

含まれる欠損値はスコアの中央値である 5.5 で補間する．

本実験では，予測精度として平均絶対誤差（MAE: Mean Absolute Error）

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

を用いる．ここで，N はテストデータ数，$\hat{y}_i$ は予測値，y_i は真値を意味する．

4.2 学習・テストデータの構築法

一般に，モデルベース協調フィルタリングの予測モデルの汎化性能の検証には交差検証法が用いられる．本研究においては，評価しているユーザ数の多い書籍上位 100 件を目的変数としているものの，そのユーザ数自体少なく，交差検証法の適用の際には，目的変数を評価しているユーザが学習データとテストデータの両方に適切な数含まれることを保証できない．そのため，本研究では交差検証法の適用が困難と考え，ブートストラップ法を適用する．ブートストラップ法では，ユーザ-アイテム行列をランダムに学習データとテストデータに分割し，重複を許してそれぞれから 200,000 回ランダムサンプリングを行う．これらの操作を 20 回繰り返し，学習データとテストデータのペアを 20 組作成する．

4.3 Word2Vec のハイパーパラメータ

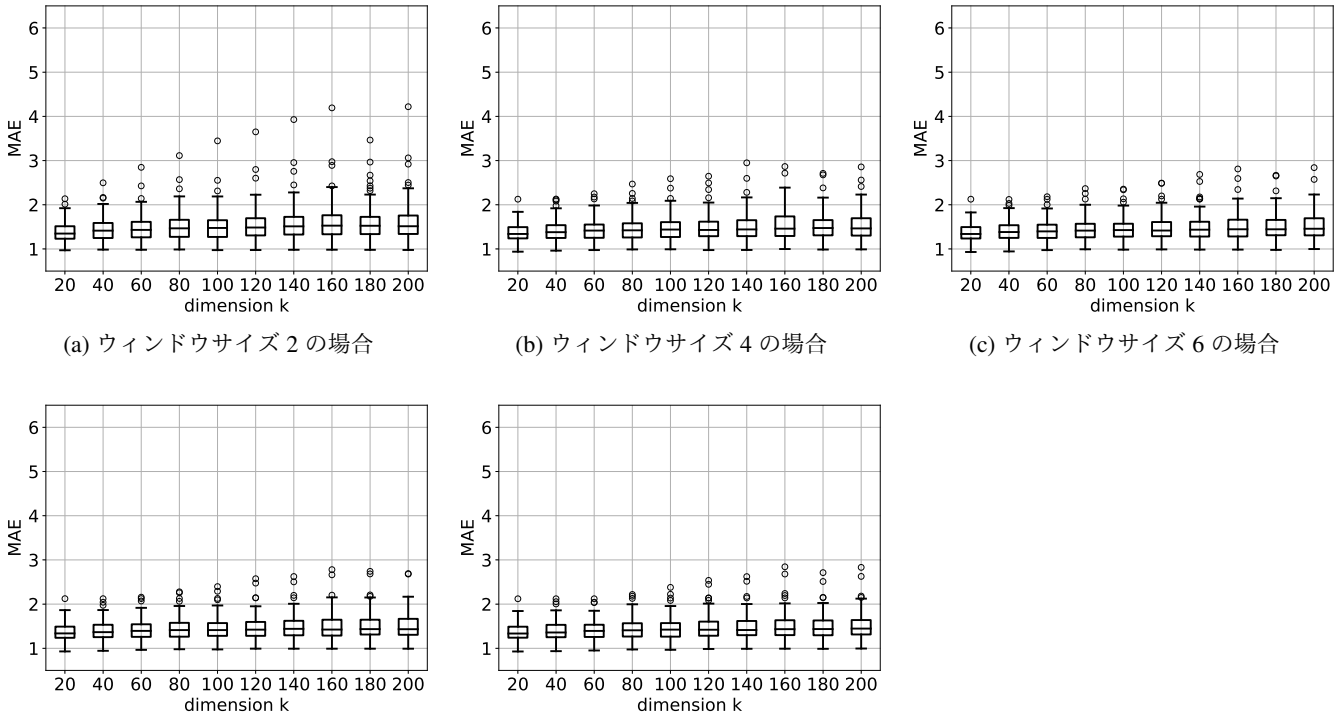
線形写像 $\boldsymbol{W}$ を得るための Word2Vec のハイパーパラメータが予測精度に与える影響を明らかにするため，圧縮次元数 k を 20 から 200 まで 20 ずつ増加させた 10 種類，学習の最大単語数を 2 から 10 まで 2 ずつ増加させた 5 種類の計 50 通りのハイパーパラメータについて検証する．なお，Word2Vec の他のハイパーパラメータは，最小出現数 0，Skip-gram 学習モデルとし，ネガティブサンプリング数と反復回数は使用プログラム [9] のデフォルトの値を用いる．

5 次元圧縮が予測精度に与える影響の考察

図 1 と図 2 に，L1 正則化と L2 正則化を適用した場合のウィンドウサイズ毎の予測精度の箱ひげ図を示す．縦軸は MAE の値，横軸は圧縮次元数 k を表している．

図 1 では，どのウィンドウサイズの場合でも，圧縮次元数 k が増えるにつれ MAE の中央値も大きくなっており，予測精度の悪化が確認できる．これは，圧縮次元数が大きくなるにつれて学習データ数に対する変数の数が増え，重回帰分析の予測モデルが適切に構築できなくなったためと考える．また，図 1 より，ウィンドウサイズが大きくなるにつれ MAE のばらつきが小さくなっていることがわかる．ウィンドウサイズを大きくすると，予測精度のばらつきを抑えられる傾向にあると考える．

図 2 より，L1 正則化を適用した場合は圧縮次元数が大きくなるにつれ MAE の値も大きくなっていたのに対し，L2 正則化を適用した場合は圧縮次元数が増えても MAE の値は変わらず，安定していることがわかる．また，L1 正則化の適用時と同様にウィンドウサイズが大きくなるに従って MAE のばらつきが小さくなっている．



(a) ウィンドウサイズ 2 の場合　(b) ウィンドウサイズ 4 の場合　(c) ウィンドウサイズ 6 の場合

(d) ウィンドウサイズ 8 の場合　(e) ウィンドウサイズ 10 の場合

図 2 L2 正則化を適用した際の予測精度

図 1 と図 2 を比較すると，L1 正則化を適用した場合よりも L2 正則化を適用した場合のほうが予測精度が向上している．Word2Vec による次元圧縮にあたっては，L2 正則化が適していると考える．また，Word2Vec のハイパーパラメータは，圧縮次元数 k はできるだけ小さく，ウィンドウサイズはできるだけ大きくするほうが予測精度の向上に寄与するといえる．

6 おわりに

本研究では，モデルベース協調フィルタリングを重回帰分析により実現することを目指し，ユーザ-アイテム行列を Word2Vec により次元圧縮し，圧縮空間で回帰係数を推定する手法を提案している．Book-Crossing データセットを用いた実験結果から，提案手法には L2 正則化の適用が望ましく，Word2Vec の圧縮次元数は小さく，ウィンドウサイズは大きくするほうが予測精度の向上に寄与することを確認している．これらの結果より，推薦理由の説明能力を有したモデルベース協調フィルタリングを重回帰分析により実現できる可能性を得ている．

今後は，まずは，L1 正則化に提案手法が適しておらず，L2 正則化が適していた理由を解明する．そして，推薦理由の説明能力について明らかにするため，重回帰分析で得られた回帰係数をどのように解釈すべきかを検討する．併せて，回帰係数の有意性も考慮したスコア予測モデルの構築も考えている．

参考文献

[1] Su, X. and Khoshgoftaar, T. M.: A Survey of Collaborative Filtering Techniques, Advances in Artificial Intelligence, 2009.

[2] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781, 2013.

[3] Sinha, R. and Swearingen, K.: The Role of Transparency in Recommender Systems, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 830-831, 2002.

[4] Gedikli, F., Jannach, D., and MouzhiGe: How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems, International Journal of Human-Computer Studies, Vol. 72, No. 4, pp. 367–382, 2014

[5] Koren, Y., Bell, R., and Volinsky, C.: Matrix Factorization Techniques for Recommender Systems, Journal Computer, Vol. 42, No. 8, pp. 30-37, 2009.

[6] Lin, W. and Alvarez, S. A.: Efficient Adaptive-Support Association Rule Mining for Recommender Systems, Data Mining and Knowledge Discovery, Vol.6, No.1, pp. 83-105, 2002

[7] Pearl, J: Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning, Proceedings Cognitive Science Society, pp. 329-334, 1985.

[8] Peduzzi, P., Concato, J., Feinstein, A. R., and Holford, T. R.: Importance of Events Per Independent Variable in Proportional Hazards Regression Analysis II. Accuracy and precision of regression estimates, Journal of Clinical Epidemiology, Vol. 48, No. 12, pp. 1503-1510, 1995.

[9] Word2Vec, https://github.com/svn2github/word2vec.git（2017 年 11 月 13 日アクセス）