次元圧縮を用いた重回帰分析による協調フィルタリング

発表者: 総合情報学科 経営情報学 コース 学籍番号 1410112 藤井 流華
指導教員: 岡本 一志 助教

1 はじめに

協調フィルタリング [1] は，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つけられるスコアの予測をする技術である．重回帰分析の協調フィルタリングへの応用により，目的とするアイテムのスコアの予測モデルを構築でき，予測に他のアイテムが与える影響を調べることができる．その一方で，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，ユーザ-アイテム行列のように次元が高くスパース性の強いデータに対してはパラメータを推定できない可能性がある．

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，圧縮空間で重回帰分析する手法を提案する．また，提案法によるスコアの予測精度と推定したパラメータについて，直接重回帰分析した結果と比較検証する．

2 word2vec による次元圧縮と重回帰分析

重回帰分析では，変数の数に対して観測データの数が 10 倍程度必要とされている [3]．また，一般にユーザ-アイテム行列は巨大な疎行列であり，次元が高く解析に利用できるデータは少なくなる傾向にある．そのため，ユーザ-アイテム行列を直接重回帰分析するとパラメータ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-アイテム行列に正則化を適用すると計算コストが高くなることが想定されるため，次元圧縮のアプローチとして，本研究では自然言語処理技術のひとつである word2vec[2] の活用を試みる．word2vec は，評価されているデータのみを用いて学習するため計算コストが低く，巨大なユーザ-アイテム行列を低い計算コストで次元圧縮することが期待できる．

提案法では，まず word2vec により学習用ユーザ-アイテム行列 $X \in \mathbb{R}^{m \times n}$ から線形写像 $W \in \mathbb{R}^{n \times k}$ を算出する．ここで，m はユーザ数，n はアイテム数，k は次元数（$k < n$）である．次に，$X' = XW$ により X の k 次元表現を得る．重回帰分析による予測モデルは $\boldsymbol{y} = \boldsymbol{\alpha_0} + X\boldsymbol{\alpha}$ であり，$\boldsymbol{y}$ と X から $\boldsymbol{\alpha_0}$ と $\boldsymbol{\alpha}$ を推定する．このとき，$\boldsymbol{y}$ は m ユーザ分のアイテムのスコアとする．一方で，圧縮空間での重回帰分析の予測モデルは $\boldsymbol{y} = \boldsymbol{\beta_0} + X'\boldsymbol{\beta}$ であり，$\boldsymbol{y}$ と X' から $\boldsymbol{\beta_0}$ と $\boldsymbol{\beta}$ を推定する．ここで，$\boldsymbol{y} = \boldsymbol{\beta_0} + X'\boldsymbol{\beta} = \boldsymbol{\beta_0} + XW\boldsymbol{\beta}$ であり，$\boldsymbol{\alpha_0} = \boldsymbol{\beta_0}$，$\boldsymbol{\alpha} = W\boldsymbol{\beta}$ とすることで圧縮空間で重回帰分析をして得られたパラメータ $\boldsymbol{\beta}$ から直接重回帰分析をして得られるパラメータ $\boldsymbol{\alpha}$ を計算することができる．

3 次元圧縮が与える影響の評価実験

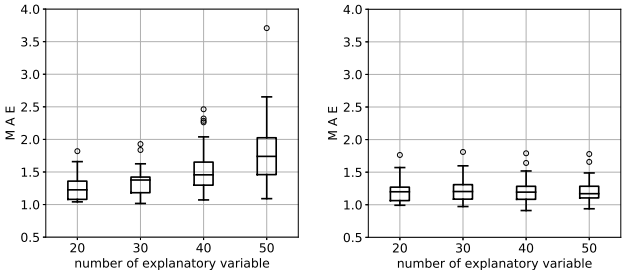
本研究では，ユーザ-アイテム行列を直接重回帰分析する場合（直接法）と，圧縮空間で重回帰分析を経て元の次元のパラメータを推定する場合（提案法）の予測精度を求める．また，直接法と提案法により得られたパラメータについて，その差異を検証する．併せて，ユーザ-アイテム行列の重回帰分析に L1 正則化（lasso）と L2 正則化（ridge）を適用した場合についても比較する．

3.1 実験環境

本実験では，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 から 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．

予測精度の検証には，ユーザ-アイテム行列を直接重回帰分析する必要がある．データの全書籍を説明変数とすると，観測データ数が足りず重回帰分析のパラメータ推定ができない可能性があるため，評価しているユーザの多い書籍上位 $n+1$ 件を変数として選択し，選択した変数のスコアデータのみを用いる．$n+1$ 件の変数から目的変数をひとつ選択し，残りの n 件を説明変数とする．本実験では，説明変数の数 n は 20 から 100 まで 10 ずつ増やしていく．選択したすべての変数をそれぞれ目的変数として設定し，目的変数毎に回帰式を立てる．なお，説明変数に含まれる欠損値は，スコアの中央値である 5.5 で補完することとする．

線形写像 $\boldsymbol{W}$ を得るための word2vec のハイパーパラメータは，次元数 $k = 10$，学習の最大単語数 100，ネガティブサンプリング数 5，トレーニング反復回数 500，最小出現数 0，Skip-gram 学習モデルとしている．



（a）直接法の場合　（b）提案法の場合

図 1: 重回帰分析の予測精度

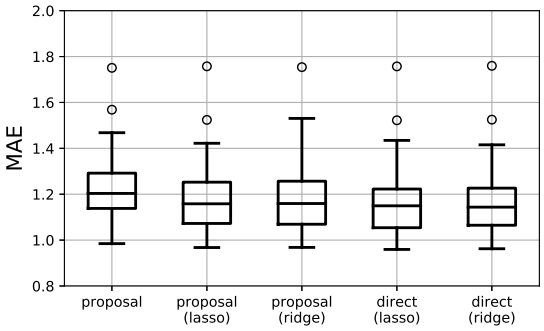


図 2: $n = 20$ で正則化を適用したときの予測精度

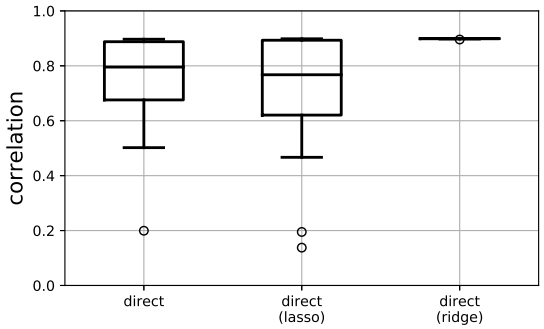


図 3: $n = 20$ の提案法のパラメータとの相関係数

3.2 予測精度の比較と考察

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．N はテスト用データ数，$\hat{y}_i$ は予測値，y_i は真値とする．MAE は値が小さいほど予測精度が良いという指標である．

図 1 に，$n = 50$ までの直接法と提案法の予測精度の箱ひげ図を示す．縦軸は MAE の値，横軸は説明変数の数を表している．$n = 20$ においては，直接法と提案法の予測精度に違いがないことがわかる．また，図 1（a）より，直接法は目的変数の数が増えるに従って MAE も大きくなり，予測精度が悪くなっている．目的変数の数に対して観測データが足りなかったためと考える．一方で，図 1（b）より，提案法の予測精度は目的変数の数に影響されないことがわかる．そのため，直接法では重回帰分析できない場合でも提案法は適用できると考える．

図 2 に，直接法（direct）と提案法（proposal）それぞれに L1 正則化（lasso）および L2 正則化（ridge）を適用した際の予測精度の箱ひげ図を示す．図 2 より，提案法に正則化を適用した場合と正則化を適用しない場合の予測精度は，MAE の中央値で 0.05 程度の差しかないことがわかる．また，提案法に正則化を適用しない場合と直接法に正則化を適用した場合の予測精度も，同様に MAE の中央値で 0.05 程度の差しかない．そのため，提案法は直接法に正則化を適用したときと同様の結果を得られると考える．

3.3 推定したパラメータの比較と考察

図 3 に，提案法で得られたパラメータについて直接法で得られたパラメータ，直接法に L1 正則化（lasso）および L2 正則化（ridge）を適用したときのパラメータとの相関係数の箱ひげ図を示す．重回帰分析では，推定したパラメータから目的のアイテムに影響を与える他のアイテムを調べることができる．相関係数が高いほど推定したパラメータの値が似ている傾向にある．図 3 より，提案法と直接法に L2 正則化（ridge）を適用した場合のパラメータの相関係数の中央値が高く，ばらつきが非常に小さいことがわかる．これより，提案法は直接法に L2 正則化（ridge）を適用したときと同様の効果があると考える．

4 おわりに

本研究では，word2vec で次元圧縮したユーザ-アイテム行列を重回帰分析し，得られたパラメータから元の次元のパラメータ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，直接法と提案法の予測精度と推定したパラメータには違いがないことを確認している．また，推定したパラメータの比較より，提案法は直接法に L2 正則化（ridge）を適用した際と同様の効果を示していると考える．

今後は，説明変数の数を増やし，提案法の予測精度や直接法に正則化を適用した場合との計算コストの違いを明らかにすることを検討している．また，適切な word2vec のハイパーパラメータについても検証していきたい．

参考文献

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

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

[3] P. Peduzzi, J. Concato, A.R. Feinstein, T.R. Holford: Importance of Events Per Independent Variable in Proportional Hazards Regression Analysis II. Accuracy and precision of regression estimates, J. of Clinical Epidemiology, 48(12), 1503-1510, 1995.