重回帰分析による推薦の透明性を有した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は推薦システムの手法のひとつで, 購入履歴やレーティングなどから類似ユーザの発見やスコア予測を行う. また, 推薦理由を説明する情報をユーザに提供することを, 推薦の透明性という. 協調フィルタリングにおける推薦の透明性に関する研究 [1][2] はメモリベース法が主流であるが, 推薦のたびに近傍探索を行う必要があるため計算コストが高くなる傾向にある.

本研究では，低計算コストでの推薦の透明性の実現を目指し, モデルベース協調フィルタリングの手法のひとつである重回帰分析の適用を試みる. 目的変数を評価しているユーザのデータのみの使用を想定するため，学習データ数が少なくなり過学習になる可能性がある．そこで, 解決法である正則化と次元圧縮について，スコア予測精度を検証する実験を行う．正則化法には L1・L2 正則化を適用し，次元圧縮法には自然言語処理技術のひとつである Word2Vec[3] を用いた手法を提案する．

2 重回帰分析と Word2Vec による次元圧縮

重回帰分析の協調フィルタリングへの応用では, 偏回帰係数から目的アイテムの予測に影響を与える他のアイテムの推定ができる. 本研究では, 過学習の解決法である正則化と次元圧縮についてスコア予測精度を検証する. 正則化法には L1・L2 正則化を適用し, 次元圧縮法には Word2Vec を適用した手法を提案する. Word2Vec[3] は自然言語処理技術のひとつで, スコアがあるところのみを学習に用いるため低計算コストでの次元圧縮が期待できる.

提案する次元圧縮法では，Word2Vec により m 人のユーザが n 個のアイテムを評価したユーザ-アイテム行列 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ 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．$\boldsymbol{X}$ と $\boldsymbol{W}$ から目的変数を除いた $\boldsymbol{X}_i \in \mathbb{R}^{m\times(n-1)}$ 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 $\boldsymbol{X}'_i$ を得る．回帰式は $\boldsymbol{y}_i = \boldsymbol{\alpha}_0 + \boldsymbol{X}_i\boldsymbol{\alpha}$ であり， $\boldsymbol{\alpha}_0$ と $\boldsymbol{\alpha}$ を推定する．このとき，$\boldsymbol{y}_i$ は目的変数を評価している m' ユーザ分のスコアである．また，圧縮空間での回帰式は $\boldsymbol{y}_i = \boldsymbol{\beta}_0 + \boldsymbol{X}'_i\boldsymbol{\beta}$ であり， $\boldsymbol{\beta}_0$ 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，$\boldsymbol{\alpha}_0 = \boldsymbol{\beta}_0$, $\boldsymbol{\alpha} = \boldsymbol{W}_i\boldsymbol{\beta}$ とすることで偏回帰係数 $\boldsymbol{\beta}$ から偏回帰係数 $\boldsymbol{\alpha}$ を計算できる.

3 正則化と次元圧縮がスコア予測に与える影響の評価実験

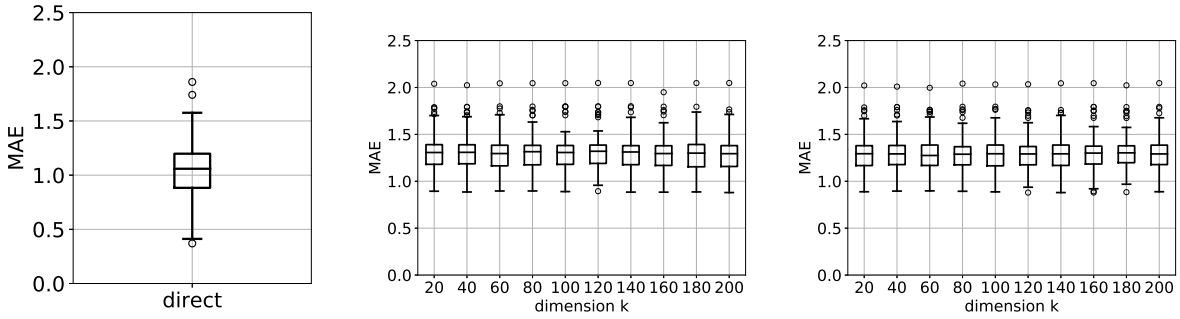
正則化法と次元圧縮法のスコア予測精度を検証し，適切な Word2Vec のハイパーパラメータを明らかにする実験を行う．また，正則化法と次元圧縮法の学習時間を計測した後， 2 つの手法の偏回帰係数の相関を検証する．

本研究では，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 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．また，重回帰分析には R の glmnet パッケージを使用する．適切な Word2Vec のハイパーパラメータの検証では, 10 種類の圧縮次元数と 5 種類のウィンドウサイズの組み合わせ計 50 種類について実験を行う．なお，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，その他は使用プログラム [4] のデフォルトを用いる．評価しているユーザ数が多い書籍上位 100 件を目的変数，残りの全書籍を説明変数に選択し, 目的変数毎に回帰式を立てる．なお，説明変数の欠損値はスコアの中央値である 5.5 で補間する．

予測精度には平均絶対誤差（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 は真値を表す．一般に, モデルの汎化誤差の検証には交差検証法が用いられるが, 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ため，本研究ではブートストラップ法を適用する.

4 正則化と次元圧縮に関する実験の結果と考察

図 1 に，L1 正則化を適用した正則化法と次元圧縮法のスコアの予測精度を示す．図 1 より，正則化法と次元圧縮法のスコア予測精度の違いは, MAE の中央値で 0.19 ~ 0.24 であることがわかる．10 段階評価での差であるので, 2 つの手法のスコア予測精度に差はないと考える．L2 正

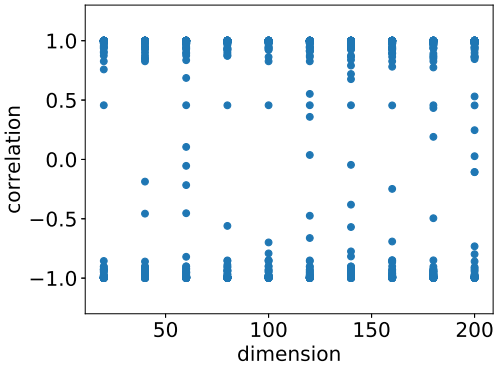


(a) 正則化法　(b) 次元圧縮法（ウィンドウサイズ 2）　(c) 次元圧縮法（ウィンドウサイズ 10）

図 1　L1 正則化を適用した正則化法と次元圧縮法のスコア予測精度

表 1　正則化法と次元圧縮法のモデルの学習時間 [s]

	正則化法	次元圧縮法
L1 正則化	34.2	23.6
L2 正則化	77.4	23.4



ウィンドウサイズ 2

図 2　正則化法と次元圧縮法の偏回帰係数の相関（L1 正則化）

則化を適用した場合も同様の結果であった．また, 圧縮次元数とウィンドウサイズによる MAE に違いはないため,Word2Vec のハイパーパラメータはスコア予測精度に影響を与えないと考える. しかし, 圧縮次元数とウィンドウサイズを増やすに連れ実行時間が長くなるため, 計算コストの観点では圧縮次元数とウィンドウサイズは小さいほうが良いといえる.

表 1 に, 正則化法と次元圧縮法について, 100 個の目的変数のモデル学習時間の合計を示す. Word2Vec のハイパーパラメータは, 圧縮次元数 20 次元, ウィンドウサイズ 2 に固定している. 表 1 より, 100 個の目的変数についてのモデル学習時間は次元圧縮法のほうが短いことがわかる. 目的変数 1 個あたりのモデル学習時間について考えると, 本実験の環境下では目的変数の数 n に対して実行時間 t は 正則化法 : $t = 0.342n$, 次元圧縮法 : $t = 20.3 + 0.0330n$ のような n の 1 次式で表すことができる. このことは, 予測したい回帰式が多数ある場合は次元圧縮法のほうが低計算コストで学習できる事を示唆している.

図 2 に, 正則化法と次元圧縮法の偏回帰係数の相関の散布図を示す. 図 2 より，圧縮次元数 20 次元, ウィンドウサイズ 2 の下では, 100 個の回帰式のうち約 50 個には高い正の相関が, 約 40 個には高い負の相関がみられる．正則化法の偏回帰係数が正しいと仮定すると，次元圧縮法の偏回帰係数は必ずしも類似していないといえる．現段階では, 次元圧縮法での推薦の透明性の実現は難しいと考える.

5　おわりに

本研究では，モデルベース協調フィルタリングで推薦の透明性の実現を目指し，重回帰分析の適用を試みた．過学習の解決法である正則化と次元圧縮について，スコアの予測精度などを検証する実験を行った．実験の結果より，重回帰分析を応用した協調フィルタリングで推薦の透明性を実現する場合, スコア予測能力やモデル学習時間の観点から L1 正則化が適していると考える. 今後の展望として，一般線形モデルから一般化線形モデルへの拡張を行う．目的変数の適切な確率分布や欠損値の適切な補間法などを明らかにしたいと考えている.

参考文献

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

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

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

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