ユーザの評価傾向を考慮した Sparse Factor Analysis による協調フィルタリング

1X09C019-1　小野　駿
指導教員　後藤 正幸

1　研究背景・目的

近年，情報技術の進展により，EC サイト等で扱う情報やアイテムの数は増加の一途をたどっている．このような現状から，ユーザの嗜好に合致したアイテムを自動で推薦するシステムの重要性が高まっている．推薦システムの代表的な手法として，類似ユーザの評価履歴データ等を用いて推薦を行う協調フィルタリング [1] があり，確率モデルや関数モデルを用いた手法など，様々な手法が既に提案されている．

関数モデルを用いた協調フィルタリングに関する研究として，Canny による Sparse Factor Analysis (以下 SFA)[2] がある．SFA は，予め与えられた数の潜在因子から評価データが生起するという仮定を置き，これらの潜在因子を説明変数とする回帰直線を推定して評価値予測を行う手法である．Canny の研究では，全データから 1 つの潜在回帰モデルを推定し，ユーザの評価傾向を同一として扱っている．しかし，ユーザには，全体的に高めの評価をし易いユーザや，低めの評価をする辛口ユーザなど，個々の評価傾向が存在すると考えられる．その場合，単一のモデルで全ユーザの評価傾向を表現することは困難であり，予測精度の低下が懸念される．

そこで，本研究では未観測であるユーザの評価傾向を，残差を用いて顕在化する．そして，予測精度の向上を目的とし，残差基準のユーザ分割を行うことで，ユーザの評価傾向を考慮してモデル化を行う SFA を提案する．提案手法を推薦システムのベンチマークデータへ適用し，その有効性を示す．

2　準備

2.1　推薦システム

推薦システムとは，ユーザの購買履歴や評価履歴からユーザの嗜好を推定し，嗜好に沿ったアイテムを推薦するシステムのことであり，本研究ではユーザのアイテムに対する評価データを対象としている．

いま，アイテム集合を $\mathcal{I} = \{I_i : 1 \leq i \leq n\}$，ユーザ集合を $\mathcal{U} = \{U_j : 1 \leq j \leq m\}$ と定義する．また，ユーザ U_j がアイテム I_i を C 段階評価で c 点の評価をした場合は c，未評価の場合は欠損値とし，これを要素として持つ評価データ行列を $Y \in \mathbb{R}^{n\times m}$ と定義する．ここで，Y_{ij} をこの行列の要素と定義する．

2.2　Sparse Factor Analysis (SFA)

SFA は線形回帰モデルに基づいているが，潜在変数である因子を仮定しているため，本研究では潜在回帰モデルと呼ぶこととする．因子集合を $\mathcal{K} = \{K_l : 1 \leq l \leq k\}$ と定義し，潜在回帰モデルを式 (1) のように表す．

$$Y = \Lambda X + N \tag{1}$$

$\Lambda \in \mathbb{R}^{n\times k}$ は各アイテムと因子間の相関の強さを表す行列であり，潜在回帰モデルに基づいて推定される回帰式の回帰係数を表す．$X \in \mathbb{R}^{k\times m}$ は各ユーザの因子に対する嗜好度合いを表す行列である．また，$N \in \mathbb{R}^{n\times m}$ は誤差項であり，この要素を v_{ij} とするとき，v_{ij} は $\mathcal{N}(0, \psi)$ の正規分布に従うものとする．SFA は式 (1) の潜在回帰モデルにおける各パラメータ Λ, ψ, X の推定値 $\hat{\Lambda}$, $\hat{\psi}$, $\hat{X}$ を EM アルゴリズムを用いて求め，$\hat{Y} = \hat{\Lambda}\hat{X}$ より予測評価値行列 $\hat{Y} \in \mathbb{R}^{n\times m}$ を算出し，評価値予測を行う手法である．なお，アイテムユーザ行列 Y は一般的に欠損が多いデータであるため，何らかの推定量を欠損値に代入して得られる行列 $\bar{Y}$ を用いて学習を行う．

2.3　従来手法学習・予測アルゴリズム

従来手法の学習・予測アルゴリズムを以下に示す．

Step1) 乱数により，Λ,ψ の初期値を与える．

Step2) 以下の式に従い X の更新を行う．

$$M = (\psi I + \Lambda^T \Lambda)^{-1} \tag{2}$$

$$X = M\Lambda^T \bar{Y} \tag{3}$$

Step3) 以下の式に従い Λ,ψ の更新を行う．

$$\Lambda = \bar{Y}X^T(XX^T + m\psi M)^{-1} \tag{4}$$

$$\psi = (1/nm)\mathrm{tr}(\bar{Y}\bar{Y}^T - \Lambda X\bar{Y}^T) \tag{5}$$

Step4) 各パラメータ Λ,ψ,X の値が収束するまで Step2, 3 を繰り返し，$\hat{\Lambda}$,$\hat{\psi}$,$\hat{X}$ を算出する．

Step5) $\hat{Y} = \hat{\Lambda}\hat{X}$ より予測評価値行列 $\hat{Y}$ を生成する．　□

3　提案手法

3.1　ユーザの評価傾向を考慮した SFA

SFA では，類似した潜在嗜好度を持つユーザは，類似した評価値を付与することを仮定し，全データから 1 つの潜在回帰モデルを推定している．しかし，実際には類似した嗜好を持ちながらも，他ユーザよりも高めに得点付けをするユーザと低めに得点付けをするユーザが混在していることが考えられる．このような各ユーザの評価傾向は，ユーザごとの平均回帰残差によって表現できる．もし、平均回帰残差の絶対値が大きいユーザが複数いる場合，ユーザの評価傾向によって層別しモデル化することによって，予測精度の向上が期待される．

そこで，本研究ではユーザの評価傾向を層別するために，平均回帰残差によってユーザのデータを複数クラスに分割し，ユーザクラスごとに SFA を行う手法を提案する．この概要を図 1 に示す．

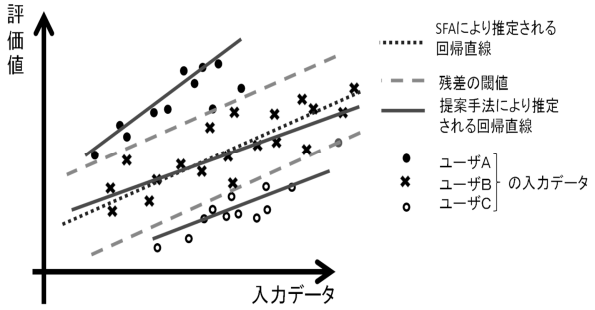

図 1. 提案手法の概要

3.2　ユーザ分割方法

本研究では前述の通り，ユーザの平均回帰残差を基準としてユーザを複数クラスに分割する．このとき，予測評価値よりも評価データの値が上回っているユーザは，全体的に高めの評価を行う傾向があると考えられる．一方，評価データの

値が下回っているユーザは，低めの評価を行う傾向があると考えられる．したがって，残差の正負を考慮し，正の残差の閾値，負の残差の閾値を設けることで，ユーザの評価傾向を3つに層別し，それに伴いユーザを3つのクラスに分割する．

ここで，ユーザ U_j の学習データ数を h_j と表記する．また，η_{ij} を，Y_{ij} が評価データである場合は1，未評価データ(欠損値) である場合は0を示すインジケータ関数として定義する．このときユーザの平均回帰残差は式 (6) で与えられるものとする．

$$S_j = \frac{1}{h_j}\sum_{i=1}^{n}(Y_{ij} - \hat{Y}_{ij})\eta_{ij} \tag{6}$$

3.3 提案手法学習・予測アルゴリズム

提案手法の予測アルゴリズムを以下に示す．

Step1) 従来の SFA を実行し，予測評価値行列 $\hat{Y}$ を生成する．

Step2) $\hat{Y}$ と学習データから，式 (6) を用いて各ユーザ U_j の平均回帰残差 S_j を算出する．

Step3) 設定した正の残差の閾値 $J_{pos} > 0$，負の残差の閾値 $J_{neg} < 0$ によりユーザを分割する．

Step4) 分割したユーザクラスで層別して SFA を行い，予測評価値を算出する．

□

ここで，Step3 における閾値は適当に設定する．設定が難しい場合には，各クラスのユーザ数が等しくなるような分割を採用することも可能である．

4 実験

提案手法の有効性を示すために，推薦システムのベンチマークデータでアイテム評価値の予測実験を行い，提案手法の予測精度の評価を行う．

4.1 実験条件

実験には，MovieLens の映画評価データ 10 万件を用いた．このデータセットはユーザ数 $m = 943$，アイテム数 $n = 1682$，$C = 5$ であり，実験に際してはランダムに学習データ 8 万件，テストデータ 2 万件に分割したものを 5 セット作成した．ユーザはすべての映画の中から，最低 20 件以上のアイテムを評価している．

5 つのデータセットに手法を適用することで未評価アイテムに対する予測評価値を算出し，MAE によって評価を行う．因子数 k は，各データセットに対して，従来手法を適用した際に最良の MAE を示したものを用いた．

ここで，提案手法における残差の閾値 J_{pos}, J_{neg} については，正の閾値 J_{pos} を 0.05 から 0.70 の範囲で，負の閾値 J_{neg} を -0.05 から -0.70 の範囲で 0.05 刻みに閾値を変化させて実験を行い，その中で最大の MAE と最小の MAE を比較に用いた．併せて，ユーザ分割を行う際に，各クラスの所属ユーザ数が同数となるようにユーザ分割を行う方法についても比較を行った．

4.2 評価方法

本研究では，推薦システムの評価指標として MAE(平均絶対誤差) を用いる．MAE は次の式 (7) で表される．

$$\mathrm{MAE} = \frac{1}{D}\sum_{j=1}^{m}\sum_{i=1}^{n}|t_{ij} - \hat{Y}_{ij}|\delta_{ij} \tag{7}$$

ここで，t_{ij} はテストデータの評価値を表し，テストデータの個数を $D = 20000$ とする．また，δ_{ij} はテストデータが存在する要素である場合は 1，それ以外の要素は 0 の値を示すインジケータ関数である．

4.3 実験結果と考察

従来手法，提案手法における各々の MAE を図 2 に示す．

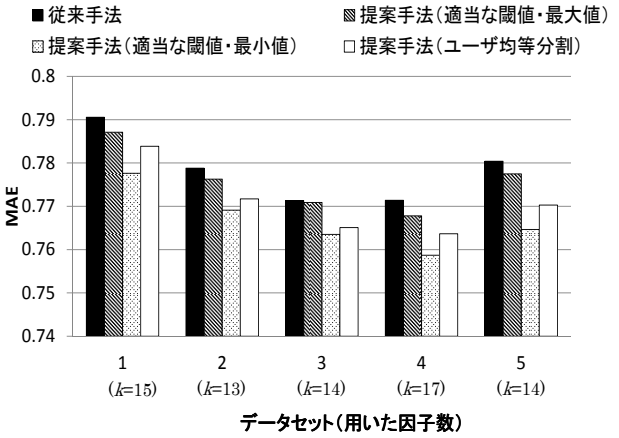

図 2. 実験結果

図 2 より全てのデータセットにおいて従来手法よりも提案手法の MAE が下回っていることから，提案手法の有効性を確認することができた．この理由として，本研究では全ユーザの評価傾向が複数あることを仮定し，複数の回帰直線を推定したが，この仮定が有効であったためと考えられる．

また，各クラスのユーザ数が同じになるようにユーザ分割を行う方法の MAE は，任意の範囲で閾値を設定した際の MAE の最大値よりも小さい値を示した．これは，ユーザの分割を行う際に，極端にユーザ数が少ないユーザクラスが生成されると，そのユーザクラスで推定される潜在回帰モデルが過学習を起こし，MAE を低下させる要因の 1 つとなるが，ユーザを均等に分割することでこの点を避けることができたためと考えられる．したがって，有効な閾値の決定方法がない場合は，ユーザ数が同じになるようにクラス分割する方法も実用的であると考えられる．

5 まとめと今後の課題

本研究では，ユーザの評価傾向を考慮した Sparse Factor Analysis に基づく協調フィルタリングの手法を提案し，実験によりその有効性を示した．

今後の課題として，残差の閾値の自動決定アルゴリズムの検討，因子数 k を決定するアルゴリズムの検討が挙げられる．

参考文献

[1] P. Esnick, N. Iacavou, M. Suchak, P. Bergstorm and J. Riedl, “An Open Architecture for Collaborative Filtering of Netnews,” *Proc. ACM Conf. on Comp. Supported Cooperative Work*, pp.175–186, 1994.

[2] J. Canny, “Collaborative Filtering with Privacy via Factor Analysis,” *Proc. 25th Annual ACM SIGIR Conf. on research and Development in Information Retrival*, pp.238–245, 2002.