行列因子分解による協調フィルタリング
—ゴルフ予約サイトデータ解析を事例にして—

鈴木　秀男

キーワード：推薦システム，スパース性，正則化

本稿は，五十嵐 丈浩さんによる 2015 年度慶應義塾大学大学院理工学研究科に提出した修士論文をもとに加筆修正したものです．

1. はじめに

情報技術の発達とともに，今日まで様々な推薦システムが提案され運用されてきました．推薦システムとは，例えば，EC サイトなどにおいて，顧客やユーザーにお勧めの商品を的確に提示して購買を促す仕組みのことです．よく知られた推薦システムの一つに協調フィルタリングがあります．協調フィルタリングは，あるユーザーが過去に行ったアイテム（商品や作品など）への評価，購買，閲覧履歴などのデータに基づき，そのユーザーに関する未知のアイテムの嗜好度を予測し推薦を行う手法です．例えば，あるユーザー A と嗜好が類似しているほかのユーザーが高く評価したり購買したりしたアイテム（ユーザー A にとっては未知のアイテム）を，ユーザー A への推薦の対象アイテムとします．一方，協調フィルタリングは，ユーザーとアイテムからなる評価行列を基礎に行われますが，その行の要素は，かなりの割合でゼロになります．すなわち，スパース性（ほんどのデータがゼロで，ごく一部が非ゼロであること）への対応の問題が指摘されています．この問題に対して，行列因子分解を行うことが有効であり，これまでいくつかの手法が提案されてきました（例えば [1, 2]）．図 1 において，行列因子分解の概念図を示します．

従来の推薦システムにおいて，ユーザーの熟達度とアイテムの難易度を考慮した推薦はあまり行われておりません．しかしながら，推薦の観点として，例えば，上手いゴルファーには難易度の高いゴルフ場を薦め，上手くないゴルファーには易しいゴルフ場を薦めることが望まれます．本研究では，ユーザースキルとアイテム難易度を考慮した行列因子分解を提案し，シミュレーションによる効果検証を行います．また実際のアンケート調査を通して提案手法の効果検証を行います．なお，本研究ではゴルフ場の予約データを用いて，ユーザーのゴルフスキルとゴルフ場の難易度を考慮する事例を扱っています．

すずき　ひでお
〒 223–8522 神奈川県横浜市港北区日吉 3–14–1
慶應義塾大学 理工学部管理工学科
hsuzuki@ae.keio.ac.jp

2. 提案モデル

行列因子分解を行うにあたり，以下の制約式を最小化するように特徴量ベクトル U_{*i}, V_{*j} を求めます．

$$
\begin{aligned}
f=&\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} S_{ij}^{user-item}(R_{ij}-{U_{*i}}^{\top}V_{*j})^2\\
&+\frac{\alpha}{2}\left[\sum_{i=1}^{N}\|U_{*i}\|^2+\sum_{j=1}^{M}\|V_{*j}\|^2\right]\\
&+\frac{\beta}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} S_{ij}^{user}\|U_{*i}-U_{*j}\|^2\\
&+\frac{\gamma}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} S_{ij}^{item}\|V_{*i}-V_{*j}\|^2
\end{aligned}
$$

ここで，R_{ij} はユーザー i のアイテム j に対する評価値，S_{ij}^{user} はユーザー i とユーザー j の類似度，S_{ij}^{item} はアイテム i とアイテム j の類似度を表します．第 1 項は実数値の評価行列と予測値の評価行列との最小二乗化項，第 2 項は過学習の防止やモデルの安定性を確保するための項，第 3 項は似ているユーザーほど彼らの特徴量を近づけるようにするための制約項，第 4 項は似ているアイテムほどそれらの特徴量を近づけるようにするための制約項です．上記の制約式を最小化することにより行列因子分解を行います．また，ユーザースキルとアイテム難易度を考慮したユーザー・アイテム間類似度行列 $S_{ij}^{user-item}$ は次のように与えられます．

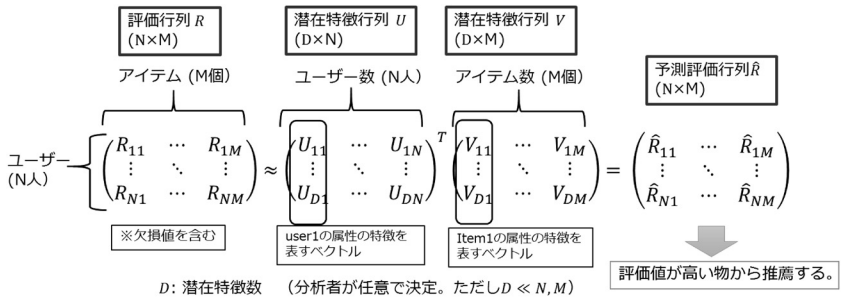

図 1　行列因子分解の概念図

$$S_{ij}^{user-item} = 1 - \left|U_i^{skill} - G_j^{level}\right|$$

ここで，U_i^{skill} と G_j^{level} は正規化したユーザー i のスキルとアイテム j の難易度を表します．

3. 実証実験

本研究では，ゴルフ予約サイトを運営する株式会社ゴルフダイジェスト・オンラインによって提供されたデータを用いて実験を行い，ゴルフ場推薦において，既存手法と精度比較を行いました．

3.1 実データによる検証：MAE の観点からの比較

ユーザー数 508 人，アイテム数 102 個（関東のゴルフ場）を対象として，ユーザーがアイテムに与えた予約回数を予測しました．評価指標には評価行列の実測値と予測値の差をとる平均絶対誤差 (MAE) を用いました．また，ユーザースキルとアイテム難易度の乖離を計る乖離度を用いました．これらの結果を表 1 に示します．

提案手法は既存手法の結果に比べ MAE においては 0.5%悪化しましたが，乖離度においては 18%改善しました．これは提案手法において $S_{ij}^{user-item}$ を既存手法の制約式の第 1 項に加えることで，ユーザースキルとアイテム難易度がマッチングしているものに重みを置いているためと考えられます．

3.2 アンケート調査に基づく検証

12 人のゴルファーに対して，シミュレーションと同じアイテムを対象にして，実際にアンケート調査を行いました．評価指標には，各ユーザーに対して五つのアイテムを提示して，何個のアイテムが選ばれたかの選択割合，そして選択されたゴルフ場難易度とユーザースキルとの乖離度を表す選択ゴルフ場乖離度を用いました．これらの結果を表 2 に示します．

提案手法は，既存手法に比べて，選択割合においては改善率 28%，選択ゴルフ場乖離度においても改善率 58%でよくなりました．また，両手法の評価値の差の検定を行ったところ 5%水準で有意であると確認されました．これは，提案手法はユーザーの嗜好を反映すると同時にユーザーに合った難易度のゴルフ場を推薦できているためと考えられます．

表 1　MAE と乖離度

	既存手法		提案手法	
	MAE	乖離度	MAE	乖離度
平均	0.035650	0.15977	0.035820	0.13114
標準偏差	0.000001	0.04401	0.000016	0.00593

表 2　選択割合と選択ゴルフ場乖離度

	既存手法		提案手法	
	選択割合	選択ゴルフ場乖離度	選択割合	選択ゴルフ場乖離度
平均	0.48333	0.19815	0.61667	0.11537
標準偏差	0.30101	0.13311	0.19924	0.09957

4. おわりに

提案手法によってユーザースキルとゴルフ場難易度の乖離は改善されました．また MAE では既存手法の方が提案手法よりも優れていましたが，アンケート調査では選択割合において提案手法の方が既存手法に比べ優れていることが示されました．推薦システムの手法を比較する場合，本来なら既存のデータに対して MAE で評価するのではなく，ユーザーに対して実際に推薦を行い手法の比較をするべきであるため，選択割合の指標で優れていた提案手法の有効性が示されました．

参考文献

[1] R. Salakhutdinov and A. Mnih, “Probabilistic matrix factorization,” *Advances in Neural Information Processing Systems*, **20**, pp. 1257–1264, 2008.

[2] Y. Zhen, W.-J. Li and D.-Y. Yeung, “TagiCoFi: Tag informed collaborative filtering,” In *Proceedings of the Third ACM Conference on Recommender Systems*, pp. 69–76, 2009.