2E1-R-1-3

ユーザクラスタリングとアイテムクラスタリングを用いた協調フィルタリングの提案

A Collaborative Filtering Based on User-Clustering and Item-Clustering

間瀬　英之　　大和田　勇人
Mase Hideyuki　　Hayato Ohwada

東京理科大学　理工学研究科　経営工学専攻
Department of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science

*2東京理科大学　理工学部　経営工学科
Department of Industrial Administration, Faculty Of Science and Technology, Tokyo University of Science

協調フィルタリングにはクラスタリングを使用する手法が存在する．従来の手法では欠損値を含んだユーザ×アイテム行列にクラスタリングを行っており，また，一つのクラスタリングにより予測する手法が多い．そこで，本研究ではすべての欠損値を補間した後にハイブリッド・クラスタリングを用いた新しい手法の提案を行った．評価実験より，本提案手法は有効性を示すことができた．

1. はじめに

インターネットの発達によって膨大な量の情報が存在するようになったが，ユーザが自分の嗜好に合った情報を獲得することは困難になっている．そのような状況を解決するための一つのアプローチとして，多くのユーザ嗜好情報からユーザの嗜好に合った情報を予測する協調フィルタリングがある [1].

協調フィルタリングにはクラスタリングを使用する手法が存在する．従来の手法では欠損値を含んだユーザ×アイテム行列にクラスタリングを行っており，また，一つのクラスタリングにより予測する手法が多い．そこで，本研究ではすべての欠損値を補間した後にハイブリッド・クラスタリングを用いた新しい手法の提案を行う．

2. 問題意識とアイデア

従来のクラスタリングを使用した協調フィルタリングでは，ユーザがアイテムを評価していない欠損値を含んだままクラスタリングを行っていた [2]．しかし，このままでは実は類似しているユーザやアイテムをクラスタできていない可能性があり，結果的に予測の精度が低下してしまうおそれがある．そこで，本研究ではユーザ×アイテム行列全体に対してデータ補間による欠損値の補間し，クラスタリングを行っていく．例えば，表（1）のユーザ×アイテム行列をユーザクラスタリングすると，表（2）のように $user_1$ と $user_3$ は評価の傾向が似ていることから，同じクラスタに分類される．しかし，データ補間により表 3 のような行列ができた後にクラスタリングを行った場合は $user_1$ と $user_3$ だけでなく，$user_4$ も評価の傾向が類似している可能性が高いと考えられる．これにより $user_1$ や $user_3$ のアイテムに対する予測において，実は $user_4$ も考慮した方が嗜好が類似した人の情報が増え，より精度が高くなるという場合が考えられる．また，これにより欠損値が増えると予測の精度が低下するというスパーシティの問題も解決される [**?**].

次に，従来のクラスタリングを使用した協調フィルタリングでは一つのクラスタリングにより予測する手法が多い．そこで，本研究ではユーザクラスタリングとアイテムクラスタリングを同時に使用したハイブリッド・クラスタリングによる新しい手法により予測を行っていく．これにより，従来の一つのクラスタリングを使った手法 [3] に比べ，より類似したものの中で予測を行うことができ，予測精度を向上させることができる可能性が高いと考えられる．例えば，$user_3$ の $item_3$ の評価を予測を考える．表（4）のようにユーザクラスタリングにより $user_1$，$user_3$，$user_4$ が類似していることから予測値

表 1: ユーザ×アイテム行列

	$item_1$	$item_2$	$item_3$	$item_4$	$item_5$	$item_6$
$user_1$	◎	—	—	○	×	—
$user_2$	×	×	×	◎	○	—
$user_3$	◎	—	◎	○	×	—
$user_4$	—	×	—	○	×	◎
…	…	…	…	…	…	…

表 2: 従来手法

	$item_1$	$item_2$	$item_3$	$item_4$	$item_5$	$item_6$
$user_1$	◎	—	—	○	×	—
$user_2$	×	×	×	◎	○	—
$user_3$	◎	—	◎	○	×	—
$user_4$	—	×	—	○	×	◎
…	…	…	…	…	…	…

表 3: 提案手法

	$item_1$	$item_2$	$item_3$	$item_4$	$item_5$	$item_6$
$user_1$	◎	×	○	○	×	○
$user_2$	×	×	×	◎	○	×
$user_3$	◎	○	◎	○	×	○
$user_4$	○	×	○	○	×	◎
…	…	…	…	…	…	…

連絡先: 間瀬　英之，東京理科大学大学院 理工学研究科 経営工学専攻，千葉県野田市山崎 2641，04(7124)1501，h-mase@ohwada-lab.net

表 4: 従来手法

	$item_1$	$item_2$	$item_3$	$item_4$	$item_5$	$item_6$
$user_1$	◎	×	○	○	×	○
$user_2$	×	×	×	◎	○	×
$user_3$	◎	○	—	◎	×	○
$user_4$	○	×	○	◎	×	◎
…	…	…	…	…	…	…

表 5: 提案手法

	$item_1$	$item_2$	$item_3$	$item_4$	$item_5$	$item_6$
$user_1$	◎	×	○	○	×	○
$user_2$	×	×	×	◎	○	×
$user_3$	◎	○	—	◎	×	○
$user_4$	○	×	○	◎	×	◎
…	…	…	…	…	…	…

は○と予測されるだろう．しかし，表（5）の提案手法のようにユーザクラスタリングとアイテムクラスタリングによって，$item_1$，$item_3$，$item_4$ も類似している情報も加えることで $item_3$ は $item_1$ や $item_4$ と同じようにユーザから高い評価を受けているので，予測値は○ではなく◎の評価を受ける可能性がある．このように，ユーザだけでなくアイテムにおけるクラスタリングの結果を加えることで，より精度の高い予測を行える可能性があると考えられる．

以上をまとめると，本提案手法には以下のような 2 つの特徴がある．

- ユーザ×アイテム行列全体のデータ補間
- ユーザクラスタリングとアイテムクラスタリングを同時に使用したハイブリッド・クラスタベースによる予測

3. 提案手法

本章では，本研究の提案手法について説明する．図 1 に処理全体の流れを示す．提案手法では，まず，ユーザ×アイテム行列全体にアイテムベースの協調フィルタリングを用いて，データ補間を行う．その後，ユーザクラスタリングとアイテムクラスタリングを組み合わせた新しい手法を用いて，予測を行う．

3.1 アイテムベースの協調フィルタリングによるデータ補間

本研究では，データ補間に関してデータセットの全体にアイテムベースの協調フィルタリングを 2 回行う．1 回目のアイテムベースの協調フィルタリングで欠損値を補間したユーザ×アイテム行列を作成し，さらにそのユーザ×アイテム行列から，2 回目のアイテムベースの協調フィルタリングを行う．これにより，さらに高い精度の予測が行えると考えられる．アイテムベースの協調フィルタリングによるデータ補間の流れを図 2 に示す．

次に，アイテムベースの協調フィルタリングアルゴリズムについて説明をする．アイテムベースの協調フィルタリングは予測するユーザが評価したアイテムセットを見て，それらが予測するアイテムにどれだけ類似しているか計算することである．すなわち，アイテムベースの協調フィルタリングは (1) アイテム間の類似度の計算，(2) 類似度に基づく予測値の計算，という 2 つの手順で予測を行っていく．

3.1.1 アイテム間の類似度計算

協調フィルタリングアルゴリズムで使われる類似度の計算にはいくつかの方法があるが，本論文ではピアソン相関係数を使用する．ピアソン相関係数は式 (1) となる．

$$Sim(i,j) = \frac{\sum_{u \in U_r}(r_{ui} - \bar{r_i})(r_{uj} - \bar{r_j})}{\sqrt{\sum_{u \in U_r}(r_{ui} - \bar{r_i})^2 \sum_{u \in U_r}(r_{uj} - \bar{r_j})^2}} \quad (1)$$

r_{ui} はユーザ u によるアイテム i の評価値，$\bar{r_i}$ はアイテム i の評価値の平均，U_r はそれぞれのアイテム i と j の評価を行ったユーザの集合を表す．

3.1.2 予測値の生成

予測するアイテム t に対する予測するユーザ a の予測評価値は式 (2) のように，予測するユーザ a が評価したアイテムの評

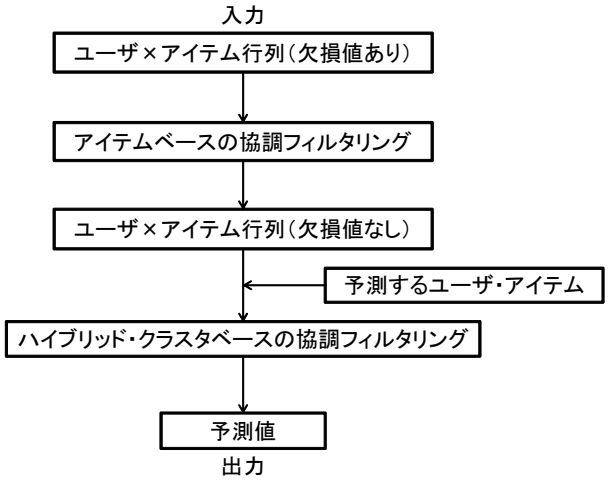

図 1: 提案手法の流れ

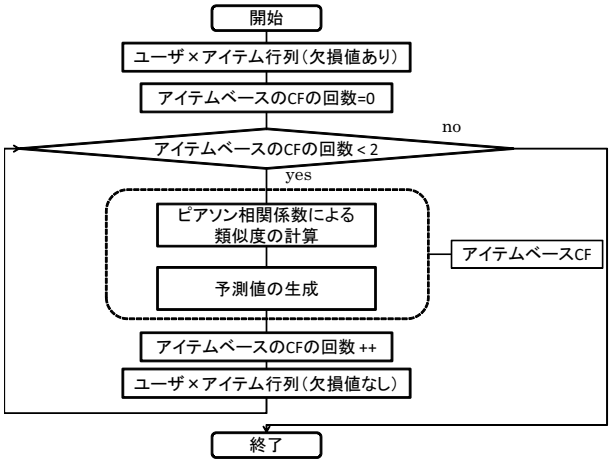

図 2: アイテムベース協調フィルタリングによるデータ補間の流れ

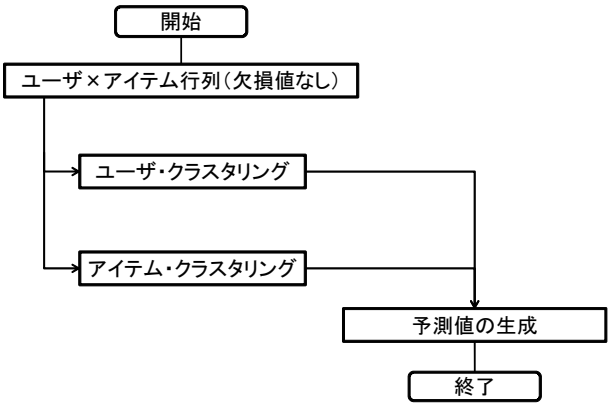

図 3: ハイブリッド・クラスタベースの協調フィルタリングの流れ

価値に類似度の重みを足し合わせて計算を行う．

$$P_{at} = \bar{r_t} + \frac{\sum_{i \in I_r} Sim(i,t)(r_{ai} - \bar{r_i})}{\sum_{i \in I_r} Sim(i,t)} \quad (2)$$

I_r は予測するユーザ a が評価したアイテムの集合である．

3.2 ハイブリッド・クラスタベースの協調フィルタリング

アイテムベース協調フィルタリングによるデータ補間を行い，欠損値のないユーザアイテムユーザ×アイテム行列が出来た．このユーザ×アイテム行列から 2 つのクラスタリングを用いて，予測を行っていく．以下に，その新しい手法について説明をしていく．

まず，ハイブリッド・クラスタベースの協調フィルタリングの流れを図 3 に示す．

3.2.1 クラスタリング手法

予測対象のユーザに類似したユーザとアイテムに類似したアイテムを選択し，クラスタを生成する．本論文では，クラスタリングアルゴリズムで最も基本的な K 平均法を用いてクラスタを生成する．ここで，K とはクラスタ数である．初めに，各データに対してランダムにクラスタを割り振る．次に，割り振ったデータをもとに各クラスタの中心を計算する．データの 1 つとクラスタの中心との距離を求め，そのデータを最も近い中心のクラスタに割り当て直す．先程の処理ですべてのデータのクラスタの割り当てが変化しなかった場合は処理を終了する．それ以外の場合は新しく割り振られたクラスタからクラスタの中心を再計算して先程の処理を繰り返す．もし，ユーザが K グループに分かれるとしたら，ユーザ $U = u_1, u_2, ..., u_n$ のクラスタリング結果は，$C_u^1, C_u^2, ..., C_u^k$ と表され，アイテム $I = i_1, i_2, ..., i_n$ の場合は $C_i^1, C_i^2, ..., C_i^k$ と表される．

3.2.2 予測値の生成

ユーザクラスタリングとアイテムクラスタリングの後に，予測するユーザ a のアイテム t の予測評価値は式 (3)，式 (4) の手順で計算する．

$$avg_u = \sum_{u \in C_a} \frac{\sum_{i \in C_t} r_{ui}}{|C_t|} \quad (3)$$

C_a は予測するユーザ a と同じクラスタに所属するユーザの集合，C_t は予測するアイテム t と同じクラスタに所属するアイテムの集合であり，r_{ui} はそれらのユーザとアイテムによる評価値である．また，$|C_t|$ は予測するアイテム t と同じクラスタに所属するアイテムの総数である．

式 (3) では，r_{ui} による評価値から C_a に属する各ユーザの平均値を求める．

$$P_{at} = r_{at} + \frac{\sum_{u \in C_a} (r_{ut} - avg_u)}{|C_a|} \quad (4)$$

$|C_a|$ は予測するユーザ a と同じクラスタに所属するユーザの総数である．式 (4) では，式 (3) で求めた各ユーザの平均値と各ユーザの予測するアイテムに対する評価値との差を取り，r_{at} に加える．

以上をアルゴリズムにして表すと図 4 のようになる．

Algorithm 1 Hybrid Cluster-Based CF Algorithm

Input: *target user*, *target item*

Output: *the prediction*

1: C_a from $C_u^1, C_u^2, ..., C_u^k$
2: C_t from $C_i^1, C_i^2, ..., C_i^k$
3: $\sum_{i \in C_t} r_{ui}, \sum_{u \in C_a} r_{at}$ from *smoothing rating matrix*
4: calc $avg_u \Leftarrow C_a, C_t, r_{ui}$
5: calc $P_{at} \Leftarrow C_a, r_{ut}, r_{at}, avg_u$

図 4: ハイブリッド・クラスタベースの協調フィルタリングアルゴリズム

4. 実験

本章では，本研究の提案手法の有効性を示すための評価実験について述べる．4.1 で実験手順を，4.2 で実験結果及び考察を述べる．

4.1 実験手順

まず，本実験では映画の評価をしたデータセット *MovieLens*[*1] に対して，500 のユーザと 1000 のアイテムが含まれる 50000 件のユーザ×アイテムの評価値を抽出し，データセットとした．

次にデータセットに対して，5-分割交差検定を行った．5-分割交差検定では標本群を 5 個に分割し，そのうちの 1 つをテストデータ，残る 4 つを訓練データとする．交差検定は，5 個に分割された標本群それぞれをテストデータとして 5 回検定を行う．そのようにして得られた 5 回の結果の平均より 1 つの推定を得る．

本実験では図 1 における入力で訓練データを入力し，テストデータを用いて予測値を出力した．また，クラスタリングにおいてはクラスタ数 K を 5，10，20，40，60，80，100 と変化させ，実験を行った．

最後に出力した予測結果に，もっとも広く使われている統計的評価基準として MAE(Mean Absolute Error)[4] を用いて評価を行った．MAE の計算式は以下の式（5）ようになる．

∗1 http://www.cs.umn.edu/Research/GroupLens/

$$MAE = \frac{\sum_{u \in T} |r_u(t_j) - \bar{r_u}(t_j)|}{|T|} \quad (5)$$

$r_u(t_j)$ は実際の評価値，$\bar{r_u}(t_j)$ は予測値で，$|T|$ は評価されたアイテムの数である．

MAE は低ければ低いほど，予測した評価値が実際の評価値に近いので予測精度が高く，提案手法の有効性を示すことができる．

4.2 実験結果及び考察

クラスタ数と訓練データごとの MAE は表 6 となった．

表 6: クラスタ数と訓練データごとの MAE

	k=5	k=10	k=20	k=40	k=60	k=80	k=100
train1	0.798	0.791	0.790	0.789	0.789	0.789	0.789
train2	0.817	0.811	0.810	0.809	0.808	0.808	0.808
train3	0.804	0.799	0.795	0.795	0.794	0.795	0.794
train4	0.806	0.799	0.796	0.795	0.795	0.795	0.794
train5	0.821	0.813	0.810	0.809	0.808	0.808	0.809

次に表 6 より，交差検定による各クラスタ数ごとの実験結果は表 7 となり，MAE は 0.79-0.81 の範囲となった．

表 7: クラスタ数ごとの MAE

	k=5	k=10	k=20	k=40	k=60	k=80	k=100
MAE	0.809	0.803	0.800	0.799	0.799	0.799	0.799

Peng ら [5] によると,「MAE が 1 以下であることは，予測値がユーザの実際の評価値と一致しているか，1 ランクずれているだけである．よって，MAE が 1 以下であることは，高い精度の予測を行えたと言える.」としている．また，*Gui－RongXue* らや *SongJieGong* の *MovieLens* を使った評価実験でも MAE は 0.79-0.86 の範囲であった．よって，本提案手法は高い精度の予測が行うことができ，提案手法の有効性を示すことができたと言えるだろう．

また，図 5 からクラスタ数が大きくなるにつれ，MAE が下がる傾向にあることが分かる．これは，クラスタ数が多い方が予測するユーザ・アイテムにより嗜好の近いユーザとアイテムの集合ができたからと考えられる．

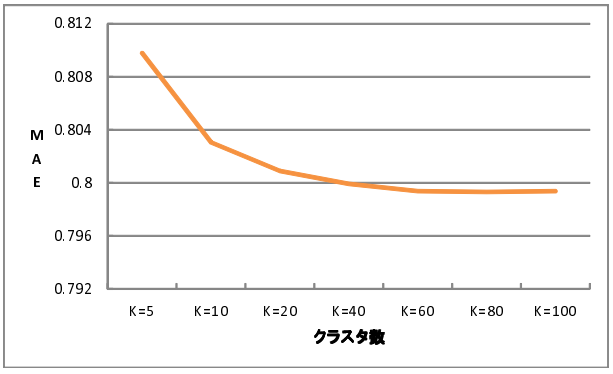

図 5: クラスタ数と訓練データごとの MAE

5. 関連研究

Gui－RongXue ら [2] はスパーシティの問題を解決するためにデータ補間の手法を導入した，ユーザクラスタリングとユーザベースの協調フィルタリングを組み合わせた新しい手法を提案した．しかし，彼らの手法はデータ補間をユーザ×アイテム行列全体に行っていなく，クラスタリングも欠損値を含んだまま行っている．そこで，本研究ではユーザ×アイテム行列全体にデータ補間を行い，行列全体を考慮したクラスタリングを行うことで，より高い精度の予測ができるのではないかと考えた．

また，*SongJieGong*[3] はユーザクラスタリングとアイテムクラスタリングを使った新しい手法を提案した．しかし，彼はユーザクラスタリングとアイテムクラスタリングを別々に使用している．そこで，本研究ではユーザクラスタリングとアイテムクラスタリングを同時に使い，予測するユーザとアイテムに対してより類似したもの同士で予測値の計算を行うことで，さらに高い精度の予測ができるのではないかと考えた．

6. まとめ

本論文では欠損値を補間したユーザ×アイテム行列にハイブリッド・クラスタリングを用いた手法の提案を行った．結果として，提案手法は高い精度の予測を行うことができ，提案手法の有効性を示すことができた．

今後の展望としては計算時間の問題がある．これは本研究においてユーザ×アイテム行列全体にデータ補間を行ったため，ユーザ数とアイテム数が膨大な量になるとデータ補間の際に膨大な時間が掛かってしまうと考えられるからである．それらを解決できれば，より高い性能の協調フィルタリングが期待できるだろう．

参考文献

[1] Sarwar B, Karypis G, Konstan J, Riedl J. Item-Based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International World Wide Web Conference. 2001. 285-295.

[2] Xue, G., Lin, C., & Yang, Q., et al. Scalable collaborative filtering using cluster-based smoothing. In Proceedings of the ACM SIGIR Conference 2005 pp.114-121.

[3] SongJie Gong. A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering. JOURNAL OF SOFTWARE, VOL.5, NO.7, JULY 2010

[4] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl. Evaluating Collaborative Filtering Recommender Systems. ACM Transactions on Information Systems 22 (2004), ACM Press, 5-53

[5] Peng Li, Seiji Yamada. A Movie Recommender System Based on Inductive Learning Algorithms. The 18th Annual Conference of the Japanese Society for Artificial Intelligence, 2004, 3H2-03