A-008

欠損値推定による協調フィルタリング手法
An Improved Method of Collaborative Filtering with Predicting Unobserved values

高島 秀佳† 山岸 英貴† 平澤 茂一†
Hideyoshi Takashima Hidetaka Yamagishi Shigeichi Hirasawa

1. はじめに

近年，膨大なデータから利用者や消費者等のユーザの要求を自動的に推定し，その要求を満たす情報を積極的に推薦するシステムが研究されている．例えば，流通業では書籍や CD について，過去のユーザの評価データからユーザの要求を満たす商品を推薦している．本論文ではこのような推薦システムの基本技術である協調フィルタリング [1] について考察する．

協調フィルタリングの代表的アルゴリズムに相関係数法 [2] がある．この手法はあるアイテムが好きか嫌いかという多数のユーザの多段階評価値について，評価値の相関係数を用いて特定のユーザの評価値を予測する．しかし，一般にある時点までに得られている評価値は少なく，欠損値が多い．そのため，予測を行う際，ユーザ間の相関係数が大きくても予測結果に反映されない場合がある．

そこで，本研究では相関係数の大きいユーザの欠損値を推定し，評価情報を補足する手法を提案する．また，シミュレーションにより提案手法の有効性を示す．

2. 協調フィルタリング

協調フィルタリングとはあるアイテムに対するユーザの評価値を，そのユーザの別のアイテムに対する評価値と，他ユーザの評価値に基づいて予測する手法である．図 1 に協調フィルタリングの概念図を示す．図 1 では，評価ユーザ u_i のアイテム v_x に対する 5 段階評価値 M_{ix} を予測する様子が示されている．ここで空白の要素は欠損値であることを示している．図 1 より，u_i と評価傾向が最も類似しているユーザは u_2 であり，u_2 の v_x に対する評価が高いことから，u_i の v_x に対する評価も高いと予測される．また，u_1,u_3 はいずれも u_i とは評価傾向が異なっており，u_i との相関係数が低いことから，予測の際にはそれほど考慮されない．

アイテム

ユーザ	v_1	v_2	v_3	v_4	$\cdots$	v_x	$\cdots$	v_z
u_1	4	2	5		$\cdots$	4	$\cdots$	1
u_2	1		2	4	$\cdots$	5	$\cdots$	4
u_3	4			2	$\cdots$	3	$\cdots$	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_i	1		2	5	$\cdots$	M_{ix}	$\cdots$	5
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_n	2		2	5	$\cdots$	4	$\cdots$	4

図 1: 協調フィルタリングの概念図

3. 従来手法

3.1 相関係数法

行がユーザを表し，列がアイテムを表す行列を行列 $M = (M_{ix})$ で示す．その (i, x) 要素 M_{ix} を i 番目のユーザ u_i の x 番目のアイテム v_x への評価値とする．ここで，ユーザが未評価の要素（空白）については欠損値として扱う．通常，行列 M は欠損値を多く含み，ほとんどの要素には値が与えられない場合が多い．相関係数法では評価対象ユーザ u_i の評価対象アイテム v_x 以外のアイテムに対する評価値と他ユーザの評価値に基づいて，欠損値 M_{ix} に対して予測値 $\hat{M}_{ix}$ を算出する．

ここで評価値の予測に用いる以下の値を定義する．M_i はユーザ u_i の評価値の平均値で，以下の式で定義する．

$$M_i = \frac{\sum_k M_{ik}}{\sum_k 1}. \tag{1}$$

ただし，ここでの $\sum_k$ は評価値が既知のアイテムのみについて取る．また，評価対象ユーザ i とデータベース中のユーザ j の相関係数 C_{ij} を以下の式で定義する．

$$C_{ij} = \frac{\sum_k (M_{ik} - M_i)(M_{jk} - M_j)}{\sqrt{\sum_k (M_{ik} - M_i)^2 \sum_k (M_{jk} - M_j)^2}}. \quad \in [-1, 1] \tag{2}$$

ただし，ここでの $\sum_x$ は評価対象ユーザ i とデータベース中のユーザ j の両方で評価値が既知のアイテムについてのみとる．以上の値を用いて未知の評価値 M_{ix} を以下のように算出する．

$$\hat{M}_{ix} = M_i + \sum_j \frac{C_{ij}(M_{jx} - M_j)}{|C_{ij}|}. \tag{3}$$

ただし，ここでの $\sum_j$ はアイテム x を評価しているユーザのみについてとる．

3.2 エントロピー

相関係数法はすべてのアイテムを同等に扱っているが，実際にはターゲットアイテム（推測されるアイテム）に大きく影響を与えるアイテムとそうでないアイテムがあると考えられる．そこで，評価値のばらつきが大きいアイテムほど予測にとって有益なアイテムであると仮定し，ばらつきを測る尺度としてエントロピー [1] を用いる．まず，アイテム x に対するエントロピーを以下の式のように定義する．

$$H_x = -\sum_v \frac{N_v(x)}{N(x)} \log \frac{N_v(x)}{N(x)} \tag{4}$$

ただし，$N_v(x)$ はアイテム x を評価値 v と評価したユーザ数，$N(x)$ はアイテム x を評価したユーザの総数．

さらに，エントロピーを用いて，アイテム x に対する重みを以下のように定義する．

$$w_x = \frac{H_x}{H_{max}} \tag{5}$$

ただし，H_{max} は全アイテムのエントロピーの最大値．

この重みを用いることにより，相関係数の算出式である式 (2) を以下のように再定義する．

$$C_{ij} = \frac{\sum_k w_k (M_{ik} - M_i)(M_{jk} - M_j)}{\sqrt{\sum_k w_k (M_{ik} - M_i)^2 \sum_k w_k (M_{jk} - M_j)^2}}. \quad \in [-1, 1] \tag{6}$$

†早稲田大学大学院理工学研究科経営システム工学専攻

3.3 従来手法の問題点

式 (3) において，アイテム v_x（評価対象アイテム）を評価していないユーザは計算に含まれない．一方，ある時点までに確定している評価情報は少なく，通常，行列 M の要素は欠損値を大変多く含む．その結果，評価対象ユーザとの相関係数は高いが予測値の推定に含まれないユーザが多数出てくることになる．そこで，提案手法ではそのようなユーザに対して評価対象アイテムの欠損値を補うことによって予測精度の向上を目指す．

4. 提案手法

本節ではユーザ u_i のアイテム v_x に対する評価値 M_{ix} を予測する問題を考える．予測精度向上のために相関係数が高いが評価対象アイテム v_x への評価値が欠損しているユーザ u_j に対し，その欠損値の補足手法を提案する．以下にそのアルゴリズムを示す．

(Step1)$n = 0$ とし，相関係数法により，評価対象 M_{ix} に対する予測を行い予測値 $\hat{M}_{ix}(n)$ を得る．

(Step2) データベース中のユーザ $u_j(\neq u_i)$ が「評価対象アイテムを予測していない」かつ「評価対象ユーザとの相関係数 C_{ij} の絶対値が閾値 s より大きい」ならば，欠損しているユーザ u_j の評価対象アイテム v_x に対する予測値 $\hat{M}_{jx}$ を，ユーザ u_j を評価対象ユーザとして相関係数法を用いて推定する．

(Step3) 得られた補足値も用いて評価対象 M_{ix} に対する予測を行い，$\hat{M}_{ix}(n+1)$ を得る．

(Step4) もし $|\hat{M}_{ix}(n+1) - \hat{M}_{ix}(n)| > \epsilon$ ならば Step5 へ進み，$|\hat{M}_{ix}(n+1) - \hat{M}_{ix}(n)| \leq \epsilon$ ならば $\hat{M}_{ix}(n+1)$ を予測値として採択する．

(Step5)$n = n+1$ とし，評価対象アイテムも相関係数計算に含み相関係数を再計算する．相関係数を更新した後 Step2 に戻る．

5. シミュレーションによる評価

5.1 利用データ

評価データには協調フィルタリングの検証用データとしてよく用いられる Movie Lens データ [3] を用いた．Movie Lens データは 943 人の 1682 本の映画に対する 5 段階評価データである．評価されている項目は全部で 100,000 個あり欠損率は 93.7 %である．

5.2 実験方法

本実験では「All but 1」方式で実験を行う [2]．

「All but 1」方式は全評価データのうち 1 つをランダムに選びその箇所をマスクし，残りの評価データからその箇所を予測し評価するという方式で，既知のユーザの評価情報が十分にある状態での推薦システムの振る舞いを見る．

また閾値は $s = 0.8$(予備実験より求めた値)，$\epsilon = 0.0001$ を用い，実験回数 N は $N = 10,000$ 回とした．

5.3 評価方法

提案手法を平均絶対誤差 MAE と F 値で評価した．

(基準 1) MAE は以下の式で表される．

$$MAE = \frac{\sum_{l=1}^{N} |\hat{M}_{ix}(l) - M_{ix}(l)|}{N}. \quad (7)$$

ここで N は実験回数，$\hat{M}_{ix}(l)$ は l 回目の実験の予測値，$M_{ix}(l)$ は l 回目の実験の正解値である．

(基準 2) 推薦されるべきアイテムが正しく推定される割合を表す評価基準として R（再現率）と P（正解率）があり，再現率は適合アイテムを漏れなく推薦できる度合い，正解率は適合アイテムだけを推薦できる度合いを表している．ここでは評価値が 4 以上のアイテムを推薦されるべき適合アイテムとし，以下の式で計算される．

$$R = \frac{\text{推薦された適合アイテム数}}{\text{データ中の全適合アイテム数}}, \quad (8)$$

$$P = \frac{\text{推薦された適合アイテム数}}{\text{推薦されたアイテム数}}. \quad (9)$$

ここで再現率と正解率はトレードオフの関係になっており，今回は両方が向上していることを評価する評価基準として F 値を用いる．F 値は以下の式で計算される．

$$F = \frac{2PR}{P+R}. \quad (10)$$

5.4 実験結果と考察

表 1 に実験結果を示す．

表 1: All but 1 方式の実験結果

	MAE	再現率	適合率	F 値
従来手法	0.775	0.779	0.750	0.764
提案手法	0.733	0.798	0.758	0.777

(結果) 表 1 より All but 1 方式では MAE が 0.042，F 値が 0.013 向上した．

(考察) エントロピーを重みとして用い，アイテムに対する評価値のばらつきを考慮した相関係数を用いることにより，予測にとって有益なユーザを選出することができ，また，そのユーザに対して欠損している評価値を推定によって補うことで，従来手法に比べて予測精度が向上したと考えられる．

また，予備実験より「補足値の更新を行わない」方法や「すべてのユーザの評価対象アイテムに対する評価値を補足する」方法よりも予測精度が向上するという結果が得られた．

6. おわりに

本研究では，評価対象ユーザとの相関係数は高いが評価対象アイテムへの評価値が欠損しているユーザに対して，その欠損値を推定して補うことで予測精度を向上させる協調フィルタリング手法を提案した．またシミュレーションを行い，従来手法に比べ予測精度が向上することを示した．

今後は更に効果的なユーザ選択の方法を考えるとともに，他のデータへの適用性も検討していきたい．

参考文献

[1] Kai Yu, xhong Wen, Xiaowei Xu, Martin Ester, " Feture Weighting and Instance Selection for Collaborative Filtering ", *Proc. of the 2nd International Workshop on Management of Information on the Web*, 2001.

[2] John S. Breese, David Heckerman, Carl Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering ", 1998 ,*in Proc. of the 14^{t}h Conference on Uncertainty in Artificial Intelligence*, pp.43-52, 1998.

[3] http://www.cs.umn.edu/Research/GroupLens