アンサンブルを導入したEMNMFによる協調フィルタリング

1X09C087-6　西川貴将
指導教員　後藤正幸

1　研究背景・目的

近年，ECサイト上には大量の商品が存在し，ユーザ嗜好も多様化しているため，各ユーザの嗜好に合致した商品を自動で推薦するシステムの重要性が高まっている．これらの手法の一つとして，類似ユーザの評価履歴や購買履歴からユーザの未購買アイテムに対する予測評価値を算出し，そのランキング順に推薦する協調フィルタリング (以下，CF) がある．

その手法の一つに Zhang ら [1] による次元縮約を用いたEM Non-negative Matrix Factorization (以下，EMNMF) がある．これはアイテムとユーザがそれぞれ共通のクラスタに属すると考え，それらへの所属度合いを示す二つの低次元かつ全要素が非負の行列の積に分解し，分解後の行列の積により評価値を近似し予測を行う方法である．EMNMF の学習では，アイテムの未評価要素に対して各アイテムの平均評価値を初期値として代入し，繰り返し更新を行うアルゴリズムによって収束させる．しかし，未評価要素に与える初期値によって，推定アルゴリズムによる収束先が異なるため，その設定法については検討の余地がある．また，クラスタ数についても事前に決定する必要があるが，最適なクラスタ数の決定のための明確な基準も示されていない．

一方，学習理論の分野では，アンサンブル学習 [2] と呼ばれる方法が脚光を浴びている．これは，学習に用いるモデルを 1 つに選択するのではなく，複数のモデルを混合することで予測精度を向上させる手法である．アンサンブル学習を用いることで，本研究で対象とする EMNMF に対しても，上記の問題を解決できる可能性がある．そこで本研究では，予測精度の向上を目的とし，異なる初期値やクラスタ数による複数の予測結果をアンサンブルする評価値予測法を提案する．提案手法を推薦システムのベンチマークデータに適用し，その有効性を示す．

2　推薦システム

推薦システムとは，ユーザの評価履歴や購買履歴からユーザの嗜好を特定し，アイテムの推薦を行う手法である．本研究では評価履歴に基づく CF に着目する．いま，アイテム集合を $\mathcal{I} = \{I_i : 1 \leq i \leq N\}$，ユーザ集合を $\mathcal{J} = \{J_j : 1 \leq j \leq M\}$ と定義する．アイテム I_i に対し，ユーザ J_j が Y 段階評価で y 点の評価をした場合は y を成分とし，未評価の場合は欠損とするアイテム・ユーザ行列を $A = [a_{ij}] \in \mathcal{R}^{N \times M}$ と定義する．行列 A に対し各推薦手法を用いることで，未評価アイテムの中からユーザが好むと予測されるものの推薦を行う．

3　EMNMF

EMNMF は，スパースであるアイテム・ユーザ行列を二つの低次元かつ全要素が非負の行列の積で近似し，未評価アイテムの予測評価値を得る手法である．いま，この二つの行列をそれぞれアイテム I_i がクラスタ $k(1 \leq k \leq K)$ に所属する度合いを表した行列 $U = [u_{ik}] \in \mathcal{R}^{N \times K}$，ユーザ J_j がクラスタ k に所属する度合いを表した行列 $V = [v_{kj}] \in \mathcal{R}^{K \times M}$ と定義する．ただし，u_{ik} は $[0, 1)$ の範囲で，v_{kj} は任意の実数値をとるものとし，U，V の積によって算出されるアイテム・ユーザ行列の近似行列を $X = [x_{ij}] \in R^{N \times M}$ で表す．すなわち，$X = UV$ である．また，アイテム・ユーザ行列 A の評価値が存在する要素の集合を $\mathcal{A}^o$，未評価要素の集合を $\mathcal{A}^u$ と定める．このうち，$a_{ij} \in \mathcal{A}^o$ に対応する成分は定数とし，$a_{ij} \in \mathcal{A}^u$ に対応する成分を変数とした行列を $A' = [a'_{ij}] \in R^{N \times M}$ とする．EMNMF は，A' が更新される度に，式 (1) を最小化するように行列 U, V を学習する．なお，$||\cdot||_F^2$ は行列のフロベニウスノルムを表す．

$$\min_{U,V} ||A' - UV||_F^2 \tag{1}$$

$$s.t. \quad \forall u_{ik} \geq 0, \ \forall v_{kj} \geq 0 \tag{2}$$

上記の最適化問題は，次の式 (3) から式 (6) を用いて A' と U, V を繰り返し更新することで解の導出が可能となる．行列 A', U, V, X について，t 回目の更新後の各行列を $A'^{(t)}, U^{(t)}, V^{(t)}, X^{(t)}$，および各行列の (i, j) 成分を $a'^{(t)}_{ij}, u^{(t)}_{ij}, v^{(t)}_{ij}, x^{(t)}_{ij}$ とする．また，$t = 0$ は各行列および各成分の初期値を表すものとする．式 (5) では式 (3)，(4) によって更新した行列 $U^{(t)}$ を正規化している．

$$u^{(t)}_{ik} = u^{(t-1)}_{ik} \frac{(A'^{(t-1)} V^{(t-1)T})_{ik}}{(U^{(t-1)} V^{(t-1)T} V^{(t-1)T})_{ik}} \tag{3}$$

$$v^{(t)}_{kj} = v^{(t-1)}_{kj} \frac{(U^{(t-1)T} A'^{(t-1)T})_{kj}}{(U^{(t-1)T} U^{(t-1)T} V^{(t-1)})_{kj}} \tag{4}$$

$$u^{(t)}_{ik} \leftarrow \frac{u^{(t)}_{ik}}{\sqrt{\sum_{i=1}^{N} (u^{(t)}_{ik})^2}} \tag{5}$$

$$a'^{(t)}{}_{ij} = \begin{cases} a_{ij} & (a_{ij} \in \mathcal{A}^o) \\ x^{(t-1)}_{ij} & (a_{ij} \in \mathcal{A}^u) \end{cases} \tag{6}$$

式 (6) より，更新後の行列 A' の未評価部分には $t-1$ 回目の処理で算出された予測評価値が代入される．最終的に上記の最適化問題の解として得られた行列 A' の値を用いて推薦を行う．

4　提案手法

4.1　着眼点

従来の EMNMF では，行列 A' の未評価成分に対する初期値として各アイテムの平均評価値を用いている．しかし，マーケティング分野ではアイテムだけでなくユーザの異質性をモデル化する方法の有効性が示されており，各ユーザの平均評価値も考慮に加えることで，より良い予測評価値を算出できる可能性がある．また，事前に決定するクラスタ数 K によっても予測精度が異なるため，適切な値を設定する必要があるが，最適なクラスタ数を一意に決定することは難しい．

そこで，EMNMF に対し前述のアンサンブル学習 [2] を導入することで，アイテム，ユーザ双方の初期値，またクラスタ数の差異によって得られる多様なモデルを混合する方法を考える．具体的には，アイテム，ユーザ双方の平均評価値を初期値として用い，それぞれに対し，複数のクラスタ数を設定して EMNMF を行い，得られる異なった予測結果をアンサンブルする．それらに加え，「アイテム，ユーザのどちらの平均評価値によって得られる予測結果をより考慮するか」を表す重みを導入する．これにより与えられたアイテムおよびユーザの評価傾向から，最もアンサンブルの効果が高まる両者の比率を決定できるようにする．

4.2　アンサンブルを用いた EMNMF

本研究では，複数のクラスタ数 K と行列 A' の初期値に加え，アイテム，ユーザの各平均評価値を用いた予測結果を

重みづけした混合を行う方法を提案する．いま，提案手法で使用するクラスタ数の集合を $\mathcal{K} = \{K_c : 1 \leq c \leq C\}$ とする．また，A' の未評価要素に対する初期値を，b_{ij}^L と定義する．提案手法ではアイテム，ユーザの平均評価値をそれぞれ用いるため，$L = 0$ のとき，b_{ij}^0 はアイテム I_i の平均評価値とし，$L = 1$ のとき，b_{ij}^1 はユーザ J_j の平均評価値をとるものとする．ここで，クラスタ数 K_c，パラメータ L を用いて作成された予測評価行列を $A'^{(K_c,L)} = [a'^{(K_c,L)}_{ij}]$ と定義する．提案手法で算出される予測評価行列 A'^* の各成分 $a'^*{}_{ij}$ は以下の式 (7) で算出される．

$$a'^*{}_{ij} = \frac{1}{C}\sum_{c=1}^{C}\left\{\alpha\left(a'^{(K_c,0)}_{ij}\right) + (1-\alpha)\left(a'^{(K_c,1)}_{ij}\right)\right\} \quad (7)$$

この時，α を $a'^{(K_c,0)}_{ij}$ と $a'^{(K_c,1)}_{ij}$ 間の重みパラメータ $(0 \leq \alpha \leq 1)$ とする．

4.3 アルゴリズム

提案手法のアルゴリズムを以下に示す．

Step1) 使用するクラスタ数 $K_1, ..., K_C$ を設定し，$c = 1$, $L = 0$ とする

Step2) アイテム・ユーザ行列 A から初期行列 $A'^{(0)}$ を作成する．$a'^{(0)}_{ij}$ を式 (8) で定義する．

$$a'^{(0)}{}_{ij} = \begin{cases} a_{ij} & (a_{ij} \in \mathcal{A}^o) \\ b_{ij}^L & (a_{ij} \in \mathcal{A}^u) \end{cases} \quad (8)$$

Step3) 行列 U，V の初期行列 $U^{(0)}$，$V^{(0)}$ の各成分を $[0, 1)$ の一様乱数に従い生成し，$t = 1$ とする．

Step4) 行列 U，V，A' の各成分を更新する．このとき，$A'^{(t)}$ の (i, j) 成分 $a'^{(t)}_{ij}$ は式 (9) で導出される．

$$a'^{(t)}{}_{ij} = \begin{cases} a_{ij} & (a_{ij} \in \mathcal{A}^o) \\ x_{ij}^{(t-1)} & (a_{ij} \in \mathcal{A}^u) \end{cases} \quad (9)$$

Step5) 各行列の更新の際に

$$||A'^{(t)} - U^{(t)}V^{(t)}||_F^2 \quad (10)$$

を計算する．この時，式 (10) が収束していなければ $t = t + 1$ として Step4 へ，収束した場合は Step6 へ移行する．

Step6) $A'^{(t)}$ を $A'^{(K_c,L)}$ として保存する．この時，$L = 0$ の場合は $L = 1$ として Step2 へ戻る．$L = 1$ の場合，$c = C$ ならば Step7 へ，そうでなければ $c = c + 1$, $L = 0$ とし，Step2 へ戻る．

Step7) Step6 で保存した全ての行列 $A'^{(K_c,L)}$ の値を用いて，アイテム・ユーザ行列の予測評価値行列 A'^* を作成する．A'^* の各成分 $a'^*{}_{ij}$ の値は式 (7) で計算される．

Step8) 行列 A'^* の各成分を予測評価値とし，予測評価値の高いアイテムをユーザに推薦する．

□

5 実験

提案手法の有効性を示すため，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，提案手法と比較手法の予測精度の評価を行う．

5.1 実験条件

実験では，公開データセット MovieLens の映画評価データ 10 万件を用いた．ユーザ数 $M = 943$, アイテム数 $N = 1682$, 評価値は 5 段階評価であり，学習データを 8 万件，テストデータを 2 万件としてランダムに分けた実験を 10 回行なった．

評価指標はテストデータと予測評価値の MAE (平均絶対誤差) を用いた．

$$\text{MAE} = \frac{1}{D}\sum_{j=1}^{M}\sum_{i=1}^{N}|T_{ij} - x^*{}_{ij}|\delta_{T_{ij}} \quad (11)$$

ここで T_{ij} はテストデータ中のユーザ J_j がアイテム I_i に対してつけた評価値を表し，D はテストデータ数である．また，$\delta_{T_{ij}}$ は T_{ij} がテストデータの場合は 1 ，それ以外の場合は 0 を示すインジケータ関数とする．MAE は誤差の指標であり，低いほど精度が良いことを示している．予備実験により，従来手法において最良の MAE を示したのは $K = 30$ の場合であったため，提案手法における K は，30 とその前後 2 ($\mathcal{K} = \{28, ..., 32\}$) とし ($C = 5$)，比較手法としてそれぞれの K における従来手法を用いた．また α による性能の変化を確認するため，α は 0 から 1.0 まで変化させて実験を行った．

5.2 実験結果と考察

提案手法の α を 0 から 1.0 まで変化させたときの結果，および従来手法で K=28 から 32 まで変化させたときの各結果を図 1 に示す．図 1 から全てのケースで提案手法が従来手法を上回っていることがわかる．この結果は，異なるクラスタ数による EMNMF の結果をアンサンブルすることの有効性を示している．一方，異なる初期値をアンサンブルすることの効果については，$\alpha = 0.5$ 付近において精度が高いことから示されている．これにより，アイテム・ユーザ双方の特徴を考慮した A' の初期値を用いて，より多くの要素に対し優れた初期値を与えることができたと考えられる．ただし，$\alpha = 0.4$ の精度が一番良くなったことから，式 (7) において初期値の違いに重みづけを行うためのパラメータ α を導入したことに意義があったといえる．

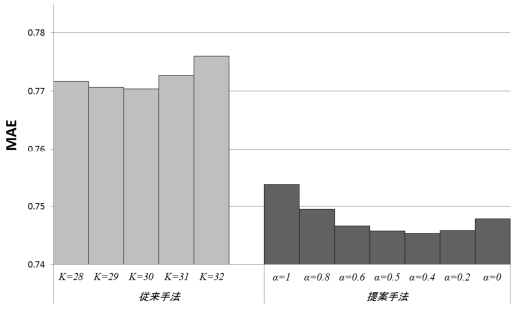

図 1. 実験結果

6 まとめと今後の課題

本研究では，EMNMF を用いた CF において，複数のクラスタ数および初期値をアンサンブルする手法を提案し，その有効性をベンチマークデータを用いた実験により示すことができた．また，混合の際にアイテム，ユーザ間の重み α を導入し，与えられたデータ毎にアンサンブルの効果をより高められる混合比率を決定できるようにした．今後の課題として，アンサンブルに用いる K および C の選択方法の検討が挙げられる．

参考文献

[1] Zhang S., Wang W., Ford J., Makedon F., “Learning from Incomplete Ratings Using Non-negative Matrix Factorization,” *6th SIAM Conference on Data Mining(SDM)*, pp.549–553, 2006.

[2] Robert S., “The Strength of Weak Learnability,” *Machine Learning*, Vol.5(2), pp.197–227, 1990.