混合ベータ分布を導入した協調フィルタリング手法に関する研究

1X11C010-1　板垣直矢
指導教員　後藤正幸

1　研究背景と目的

Web サービスの進歩に伴い，多くの EC サイト等では，ユーザーに対しての利用・販売促進を目的として，膨大な情報の中からユーザーの嗜好に合致するアイテム (商品) を自動的に推薦する推薦システムが導入されるようになっている．推薦システムの手法の 1 つに，ユーザー間の評価履歴の類似性を利用して未評価アイテムの評価値を予測する協調フィルタリング (以下 CF) がある．一般に，CF を用いた推薦システムにおいて，アイテム推薦は予測評価値の高い順に行われるため，高精度な評価値予測は大変重要な課題である．

一方，評価値予測のための CF の手法として，Hofmann により Gaussian Probabilistic Latent Semantic Analysis(以下 Gaussian pLSA)[1] が提案されている．Gaussian pLSA は，嗜好の類似したユーザーの集合である潜在クラスを仮定し，これを用いて未評価アイテムの評価値を予測する pLSA[2] の拡張モデルである．ユーザーがあるアイテムに対して評価値を与えることを確率事象として考え，潜在クラスごとに評価値が正規分布に従うことを仮定している．しかし，実際に観測データから評価履歴の度数分布を調べると，その形状は最頻値が分布の中心とは限らず多様である．加えて，評価値の定義域は $(-\infty, \infty)$ でなく 1 から 5 までの 5 段階評価といった閉区間で与えられるため，区間外の評価値の出現確率を考慮してしまう正規分布は評価値の出現分布には適さないと考えられる．そこで，閉区間上の多様な分布形を表現可能であり，パラメータ数も正規分布と同数であるベータ分布に注目する．本研究では，ベータ分布を pLSA に導入したモデルを提案し，予測評価値の精度向上を目指す．また，ベンチマークデータを用いた評価実験によってこの手法の有用性を示す．

2　準備

2.1　変数の定義

推薦システムのユーザーを $u_i(i \in \{1, ..., m\})$，被評価対象であるアイテムを $a_j(j \in \{1, ..., n\})$，嗜好の類似したユーザーの集合である潜在クラスを $z_k(k \in \{1, ..., K\})$，評価値を表す変数を r とそれぞれ定義する．ここで，m，n，K はそれぞれ全ユーザー数，全アイテム数，潜在クラス数である．また，学習データ数を N^L とし，x_l, y_l, v_l をそれぞれ l 番目の学習データの組 $(l \in \{1, ..., N^L\})$ におけるユーザー，アイテム，評価値とする．すなわち，$x_l \in \{u_1, ..., u_m\}$，$y_l \in \{a_1, ..., a_n\}$ であり，v_l は離散値である．確率を $P(\cdot)$，確率密度関数を $p(\cdot)$ で表す．

2.2　混合分布を導入した pLSA

Probabilistic Latent Semantic Analysis(pLSA)[2] は，嗜好の類似したユーザーの集合である潜在クラスを仮定し，未評価アイテムの評価値を潜在クラスによる条件付確率分布で表したモデルである．ユーザーの潜在クラスに対する所属確率 $P(z_k|u_i)$ と，潜在クラスにおける評価値の出現傾向を表す確率密度関数 $p(r|a_j, z_k)$ を用いてユーザー u_i のアイテム a_j に対する評価値 r の確率密度関数 $p(r|u_i, a_j)$ を式 (1) で定義する．

$$p(r|u_i, a_j) = \sum_{k=1}^{K} P(z_k|u_i) p(r|a_j, z_k) \tag{1}$$

ただし，$P(z_k|u_i)$ は $\sum_{k=1}^{K} P(z_k|u_i) = 1$ を満たすものとする．学習データから式 (1) のパラメータを推定し，推定された密度関数 $\hat{p}(r|u_i, a_j)$ の r に関する期待値を用いて，ユーザー u_i のアイテム a_j に対する予測評価値 $\hat{r}_{i,j}$ を式 (2) で求める．

$$\hat{r}_{i,j} = \int_{-\infty}^{\infty} r\hat{p}(r|u_i, a_j) dr \tag{2}$$

3　従来手法 : Gaussian pLSA

3.1　定式化

Gaussian pLSA[1] は，ユーザーの評価傾向が正規分布で表現可能であるとし，pLSA の潜在クラスにおける評価値の確率分布に正規分布を仮定したモデルである．ユーザー u_i のアイテム a_j に対する評価値の出現に関する確率分布 $p(r|u_i, a_j)$ は，潜在クラスに依存する評価値の平均 $\mu_{j,k}$ と分散 $\sigma_{j,k}^2$ を用いて式 (3) で定式化される．

$$p(r|u_i, a_j) = \sum_{k=1}^{K} P(z_k|u_i) \frac{1}{\sqrt{2\pi}\sigma_{j,k}} \exp\left[-\frac{(r-\mu_{j,k})^2}{2\sigma_{j,k}^2}\right] \tag{3}$$

予測評価値 $\hat{r}_{i,j}$ は，学習データから推定したパラメータ $\hat{P}(z_k|u_i), \hat{\mu}_{j,k}$ を用いて式 (4) で求められる．

$$\hat{r}_{i,j} = \sum_{k=1}^{K} \hat{P}(z_k|u_i)\hat{\mu}_{j,k} \tag{4}$$

3.2　パラメータの推定

潜在クラスは，観測されない仮想的なユーザークラスタであるため，潜在クラス z_k に依存するパラメータの推定量を学習データから陽に求めることができない．そこで，EM アルゴリズムを用いてパラメータを推定する．

EM アルゴリズムは E-Step，M-Step から成り，この 2 ステップの繰り返し学習により尤度を最大化するパラメータを推定する方法である．E-Step では潜在クラスに関する尤度の期待値を計算し，M-Step ではこの尤度を最大化するパラメータを求める．以下に Gaussian pLSA による EM アルゴリズムのパラメータ推定式を示す．

E-Step)

$$P^{(t)}(z_k|x_l, y_l, v_l) = \frac{p^{(t)}(v_l|y_l, z_k)P^{(t)}(z_k|x_l)}{\sum_{k'=1}^{K} p^{(t)}(v_l|y_l, z_{k'})P^{(t)}(z_{k'}|x_l)} \tag{5}$$

M-Step)

$$P^{(t+1)}(z_k|u_i) = \frac{\sum_{l:x_l=u_i} P^{(t)}(z_k|x_l, y_l, v_l)}{\sum_{k'=1}^{K}\sum_{l:x_l=u_i} P^{(t)}(z_{k'}|x_l, y_l, v_l)} \tag{6}$$

$$\mu_{j,k}^{(t+1)} = \frac{\sum_{l:y_l=a_j} v_l P^{(t)}(z_k|x_l, y_l, v_l)}{\sum_{l:y_l=a_j} P^{(t)}(z_k|x_l, y_l, v_l)} \tag{7}$$

$$\sigma_{j,k}^{(t+1)^2} = \frac{\sum_{l:y_l=a_j} (v_l - \mu_{j,k}^{(t+1)})^2 P^{(t)}(z_k|x_l, y_l, v_l)}{\sum_{l:y_l=a_j} P^{(t)}(z_k|x_l, y_l, v_l)} \tag{8}$$

ただし，添え字 (t) は t 回目の繰り返し時点のパラメータであることを表す．繰り返し学習の終了条件は，対数尤度の変化が一定値以下になった時点とし，その時のパラメータを評価値予測に用いる．

4 提案手法

4.1 着眼点

Gaussian pLSA では，ユーザーの評価値の出現傾向に正規分布が仮定されていた．しかし，実際の評価値は 1 から 5 までの 5 段階評価といった形で与えられるため，その形状は最頻値が分布の中心付近とは限らず多様である．加えて，評価値の定義域が $(-\infty, \infty)$ でなく閉区間であり，区間外の評価値の出現確率を考慮してしまう正規分布は評価値の出現分布には適さないと考えられる．そこで，本研究ではベータ分布に注目する．ベータ分布は閉区間上の多様な分布形を表現可能である．また，推定すべきパラメータが正規分布と同数であるため，モデルの複雑さは正規分布と同程度である．本研究では，ベータ分布を pLSA に導入したモデルを提案し，予測評価値の精度向上を目指す．

4.2 定式化

提案モデルでは，各潜在クラスにおけるベータ分布のパラメータを $\alpha_{j,k}$，$\beta_{j,k}$ として，ユーザー u_i のアイテム a_j に対する評価値の出現に関する確率分布 $p(r|u_i, a_j)$ を式 (9) で定式化する．

$$p(r|u_i, a_j) = \sum_{k=1}^{K} P(z_k|u_i) \frac{\Gamma(\alpha_{j,k} + \beta_{j,k})}{\Gamma(\alpha_{j,k})\Gamma(\beta_{j,k})} r^{\alpha_{j,k}-1}(1-r)^{\beta_{j,k}-1} \tag{9}$$

ここで，$\Gamma(\cdot)$ はガンマ関数を表す．予測評価値 $\hat{r}_{i,j}$ は推定された各ベータ分布の平均 $\hat{\alpha}_{j,k}/(\hat{\alpha}_{j,k} + \hat{\beta}_{j,k})$ と $\hat{P}(z_k|u_i)$ を用いて，式 (10) で求める．

$$\hat{r}_{i,j} = \sum_{k=1}^{K} \hat{P}(z_k|u_i) \frac{\hat{\alpha}_{j,k}}{\hat{\alpha}_{j,k} + \hat{\beta}_{j,k}} \tag{10}$$

4.3 パラメータの推定

従来手法と同様に，パラメータの推定には EM アルゴリズムを用いる．

E-Step)

$$P^{(t)}(z_k|x_l, y_l, v_l) = \frac{p^{(t)}(v_l|y_l, z_k)P^{(t)}(z_k|x_l)}{\sum_{k'=1}^{K} p^{(t)}(v_l|y_l, z_{k'})P^{(t)}(z_{k'}|x_l)} \tag{11}$$

M-Step)

$$P^{(t+1)}(z_k|u_i) = \frac{\sum_{l:x_l=u_i} P^{(t)}(z_k|x_l, y_l, v_l)}{\sum_{k'=1}^{K}\sum_{l:x_l=u_i} P^{(t)}(z_{k'}|x_l, y_l, v_l)} \tag{12}$$

$$\alpha_{j,k}^{(t+1)} = \psi^{-1}\left(\psi(\alpha_{j,k}^{(t)} + \beta_{j,k}^{(t)}) + \gamma_{j,k}^{(t)}\right) \tag{13}$$

$$\beta_{j,k}^{(t+1)} = \psi^{-1}\left(\psi(\alpha_{j,k}^{(t)} + \beta_{j,k}^{(t)}) + \delta_{j,k}^{(t)}\right) \tag{14}$$

ただし，

$$\gamma_{j,k}^{(t)} = \frac{\sum_{l:y_l=a_j} \log v_l P^{(t)}(z_k|x_l, y_l, v_l)}{\sum_{l:y_l=a_j} P^{(t)}(z_k|x_l, y_l, v_l)} \tag{15}$$

$$\delta_{j,k}^{(t)} = \frac{\sum_{l:y_l=a_j} \log(1-v_l) P^{(t)}(z_k|x_l, y_l, v_l)}{\sum_{l:y_l=a_j} P^{(t)}(z_k|x_l, y_l, v_l)} \tag{16}$$

ここで，$\psi(\cdot)$ はディガンマ関数，$\psi^{-1}(\cdot)$ は，ディガンマ関数の逆関数である．$\psi^{-1}(\cdot)$ は解析的に求めることができないため，ニュートン法により探索的に求める．また，従来手法と同様，繰り返し学習の終了条件は対数尤度の変化が一定値以下になった時点とし，その時点のパラメータを評価値予測に用いる．

5 実験

提案手法の有効性を検証するため，従来手法との評価値の予測精度を比較する実験を行った．

5.1 実験条件

データセットには，GroupLens[3] による MovieLens-100k の映画評価データを用いた．MovieLens-100k はユーザー数 943，アイテム数 1,682，評価履歴 100,000 件の評価値データであり，評価値は 1 から 5 までの 5 段階である．このデータを，各ユーザーが各々1 件の評価履歴を持つようなユーザー数に当たる 943 件をテストデータ，残りを学習データとして 99,053 件に分割した．この学習データを用いてモデルのパラメータを推定し，予測により得られた評価値をテストデータによって評価した．また，従来手法，提案手法ともに潜在クラスの数を 2〜20，30，40，50 と変化させ，それぞれの場合について評価を行った．

5.2 評価指標

評価指標には，テストデータと予測評価値の平均絶対誤差 (MAE) を用いた．MAE はテストデータにおける評価値 $r_q^{Test}(q \in \{1, ..., N^{Test}\})$ と，これに対応する予測値 $\hat{r}_q$，テストデータ数 N^{Test} を用いて以下の式で表される．

$$\mathrm{MAE} = \frac{1}{N^{Test}} \sum_{q} |\hat{r}_q - r_q^{Test}| \tag{17}$$

5.3 結果と考察

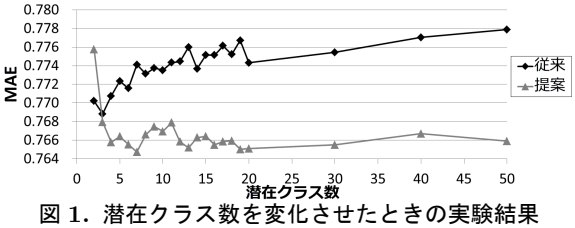

図 1．潜在クラス数を変化させたときの実験結果

図 1 に潜在クラス数を変化させたときの予測精度の挙動を示す．提案手法は潜在クラスが一定数以上ある際には，従来手法と比較して精度が向上していることがわかる．潜在クラス数が少ない時は，すべての評価値の出現頻度が一様に近くなるため，尖度の小さな分布形の方が当てはまりが良くなる．そのため，潜在クラス数が少ない場合には，閉区間で定義され，尖度の大きいベータ分布よりも，開区間で定義され尖度の小さい分布形を考慮できる正規分布の方が精度がよくなったと考えられる．また，潜在クラス数 7 の場合の提案手法が最小の MAE を示している．これは，潜在クラス数が適当な数の場合は，各潜在クラスの評価値の分布間で多様な分布形を考慮でき，さらに十分な数の分布を混合することができたことで，ユーザーの多様な評価傾向をより良く表現できた為と考えられる．

6 まとめと今後の課題

本研究では，推薦システムの CF 手法において，混合正規分布を用いた pLSA である Gaussian pLSA を，混合ベータ分布に拡張したモデルを提案した．さらに評価値予測で提案手法が従来手法と比べて高性能な予測ができることを示した．

今後の課題として，ユーザーごとの評価傾向の違いに着目しモデルの改善を行い，このモデルにおいてさらに高性能な評価値の予測を実現するということがあげられる．

参考文献

[1] Thomas Hofmann.“Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis.”Proc. 26th Ann. International ACM SIGIR Conf. 2003.

[2] Thomas Hofmann. “Probabilistic Latent Semantic Indexing.”Proc. 22nd Ann. International ACM SIGIR Conf. 1999.

[3] Grouplens Research, MovieLens DataSets. www.grouplens.org