逐次学習更新則を導入したFCE法によるファジィ協調フィルタリング

1X08C068-3　柴谷雄大
指導教員　後藤正幸

1　研究背景

あるユーザに対して情報を推薦する情報推薦の代表的な手法として，嗜好の類似した他のユーザの情報を用いて自動的に推薦を行う協調フィルタリング (以下 CF) がある．CFには，様々な機械学習アルゴリズムが用いられているが，本研究では，クラスタリングを用いた手法に注目する．クラスリングには，ユーザが複数のクラスタに所属することで嗜好を複数加味できるファジィクラスタリングがあり，その代表的な手法として，クラスタ形状を自由に変化させ，様々なユーザに対応することを可能とした Fuzzy c-Elliptotypes（FCE）法 [1] が提案されている．

通常のクラスタリングでは，データベース上の全データを対象にクラスタを構築する一括学習の枠組みで議論されるのが一般的である．しかし，近年 EC サイトでは，膨大な量の新しいデータがデータベース上に蓄積され続け，ユーザの嗜好は常に変化している．このような現状に対応するための方法としては，新たに与えられたデータのみを学習する逐次学習が考えられる．この逐次学習の立場から FCE 法を CF に適応した研究として，新たなデータが与えられるごとにクラスタを更新する本多ら [2] の逐次学習アルゴリズムがある．しかし，本多らの手法では，ユーザの最適クラスタを 1 つに選択し推薦しているため，ユーザが類似したクラスタに所属した場合，最適クラスタ以外の影響を無視してしまい，推薦結果が悪くなってしまう．

そこで，本研究では逐次学習に基づく FCE 法の推薦において，ファジィクラスタリングの特性を利用し，ユーザが複数のクラスタに所属することを許容するモデルを提案し，より精度の高い推薦が可能であることを示す．

2　Fuzzy c-Elliptotypes（FCE）法

N 人のユーザが M 個のコンテンツを評価した（$N \times M$）データ行列 $X = (x_{ij})$ が与えられたときに，N 人のユーザを C 個のクラスタに分割することを考える．FCE 法は，c-means 法を複数クラスタに所属することを仮定した Fuzzy c-Means(FCM) 法 [1] と評価値と線形多様体との距離を用いた Fuzzy c-Varieties(FCV) 法 [1] を線形結合させることで，クラスタの形状を球状から線形多様体へと自由に変化させることができる手法である．また，FCV 法は局所的な主成分分析と等価であることが，Bezdak ら [1] により，証明されている．以上のことから，本多ら [3] はデータ行列 X に対して，主成分行列の要素である a_{cjk} と成分得点 f_{cik} を用いることで，FCE 法の目的関数を以下のように定式化している $(k = 1, 2, \ldots, p)$．$\boldsymbol{b}_c$ は第 c クラスタの中心を表す．

$$L_{fce} = \sum_{c=1}^{C}\sum_{i=1}^{N}\sum_{j=1}^{M} u_{ci}^{\theta}\left\{\alpha\left(x_{ij} - \sum_{k=1}^{p} f_{cik}a_{cjk} - b_{cj}\right)^2 + (1-\alpha)(x_{ij} - b_{cj})^2\right\} \quad (1)$$

u_{ci} は第 i ユーザの第 c クラスタへの所属度合いを表すメンバシップで，$\sum_{c=1}^{C} u_{ci} = 1$ を満たす．また，α が FCV 法と FCM 法の調整パラメータとなっており，θ はメンバシップのファジィ度を決める定数である．この θ が大きいほど各ユーザの所属クラスタへの所属が明確でなくなり，曖昧なユーザ分割が得られる．ただし，唯一の解を得るために，メンバシップの制約のほかに，以下の制約を付与する．F_c, U_c, A_c はそれぞれ f_{cjk}, u_{ci}, a_{cjk} を要素とする行列である．

$$F_c^T U_c^{\theta} F_c = I \quad ; \quad c = 1, ..., C \quad (2)$$

$$F_c^T U_c^{\theta} 1_{\mathrm{n}} = 0 \quad ; \quad c = 1, ..., C \quad (3)$$

$$A_c^T A_c = I \quad (4)$$

FCE 法では，(1) 式の目的関数を最小とするパラメータを求める事で，最適なクラスタリング結果が得られる．

3　従来手法

3.1　従来手法の逐次学習更新則

現在，EC サイトなどの普及により，時間経過とともに膨大な量のユーザやアイテムが，データベース上に蓄積されている．その結果，データベース上の情報は常に更新されているため，ユーザの購買性向や新たなアイテムに対するユーザの評価も時間経過と共に変化すると考える方が自然である．そこで本多ら [2] の手法では，ユーザ i がアイテム j を 1 件評価する度に，クラスタリング結果を更新する逐次学習アルゴリズムを構築している．逐次学習アルゴリズムは FCE 法を与えられた評価値の各パラメータの最小化を行い，以下の式を用いて最適なクラスタリング結果を得る．

$$L_{fcem}^{(i,j)} = \sum_{c=1}^{C} u_{ci}^{\theta}\left\{\alpha\left(x_{ij} - \sum_{k=1}^{P} f_{cik}a_{cjk} - b_{cj}\right)^2 + (1-\alpha)(x_{ij} - b_{cj})^2\right\} \quad (5)$$

これらのパラメータを以下の更新則を用いて更新する．

$$b_{cj} = b_{cj} + u_{ci}^{\theta}\gamma(x_{ij} - \alpha f_{cik}a_{cjk} - b_{cj}) \quad (6)$$

$$a_{cjk} = a_{cjk} + u_{ci}^{\theta}\gamma\left(\frac{1}{f_{cik}}(x_{ij} - b_{cj}) - a_{cjk}\right) \quad (7)$$

$$f_{cik} = f_{cik} + \gamma\left(\frac{1}{a_{cjk}}(x_{ij} - b_{cj}) - f_{cik}\right) \quad (8)$$

一方，メンバシップ u_{ci} については，最適値を容易に求める事が出来ない．そこで，u_{ci} を更新する際に用いる評価値とクラスタ中心との距離尺度 d_{ci}^2 を逐次更新することを考える．今，x_{ij} が与えられたときの d_{ci}^2 の更新則は以下のようになる．

$$d_{ci}^2 = d_{ci}^2 + \gamma\left\{\alpha\left(x_{ij} - \sum_{k=1}^{P} f_{cik}a_{cjk} - b_{cj}\right)^2 + (1-\alpha)(x_{ij} - b_{cj})^2 - \frac{1}{m}d_{ci}^2\right\} \quad (9)$$

(9) 式から,u_{ci} の更新則は以下になる．

$$u_{ci} = \left\{\sum_{l=1}^{C}\left(\frac{d_{ci}^2}{d_{li}^2}\right)^{\frac{1}{\theta-1}}\right\}^{-1} \quad (10)$$

3.2　協調フィルタリングへの適用

本研究では，FCE 法と逐次学習アルゴリズムを用いてクラスタリングを行う．このクラスタリング結果を用いて，未評価値を持つユーザに似たような評価値を持つユーザの嗜好情報をもとに，未評価値を推測する CF へ適用させる事ができる．推定対象となるユーザ i' の未評価値 $\hat{x}_{i'j'}$ について，

$$\hat{c}_{i'} = \arg\max_{c} u_{ci'} \quad (11)$$

となるクラスタを決定し，観測値のみを考慮して主成分得点 $f_{\hat{c}i'k}$ を求めることにより，

$$\hat{x}_{i'j'} = \sum_{k=1}^{p} f_{\hat{c}_{i'}i'k} a_{\hat{c}_{i'}j'k} + b_{\hat{c}_{i'}j'} + \bar{X} \qquad (12)$$

のように推定できる．$\bar{X}$ はデータ行列 X の平均値を表し，$\hat{x}_{i'j'}$ は未評価値に対する推測値である．

4 提案手法

4.1 背景

本多らの手法 [2] の問題点として，ユーザをメンバシップ値が最大となる唯一のクラスタに所属させているため，嗜好の類似したクラスタが複数存在した場合，最適クラスタ以外の嗜好を無視してしまうという点が挙げられる．この問題点を改善するため，本多ら [1] のパラメータ更新則を用いて FCE 法を適用し，メンバシップ値を考慮した推薦方法を提案する．メンバシップ値は各ユーザが各クラスタに対する所属度合いを表しているため，提案手法を用いることでユーザに複数の嗜好を持たせたまま，推薦を行うことができる．

4.2 メンバシップ値を用いた CF への適用

推定対象となるユーザ i' の未評価値 $\hat{x}_{i'j'}$ について，メンバシップの混合比を用いることにより複数クラスタに所属するものとする．$\hat{x}_{i'j'}$ は観測値のみを考慮して成分得点 $f_{ci'k}$ を求めることにより，

$$\hat{x}_{i'j'} = \sum_{c=1}^{C} u_{ci'}^{\theta} \left\{ \sum_{k=1}^{p} f_{ci'k} a_{cj'k} + b_{cj'} \right\} + \bar{X} \qquad (13)$$

のように推定できる．また，L から D までの $D-L+1$ 段階評価のデータ行列に対する推測の算出の場合, 以下の制約を付与し，$\tilde{x}_{i'j'}$ を予測値として用いる．

$$\tilde{x}_{i'j'} = \begin{cases} D & (\hat{x}_{i'j'} > D) \\ \hat{x}_{i'j'} & (L \leq \hat{x}_{i'j'} \leq D) \\ L & (\hat{x}_{i'j'} < L) \end{cases} \qquad (14)$$

以上から，メンバシップ値を用いて推薦する事で，ファジィクラスタリングの特性を生かした推薦結果を推定することができる．

5 実験方法

協調フィルタリングシステムとしての推薦能力の検証のために，MovieLens 映画お勧めシステムの web サイトで公開されている 5 段階評価のデータ [4] を用いて実験を行った．

5.1 実験条件

用いたデータセットは従来手法 [1] 同様，学習データ数を 80,000 件，テストデータを 20,000 件としたデータセットをランダムに 5 セット (u1〜u5) 用意した．評価指標は，実測値と推測値の差の絶対値の平均を取る MAE（平均予測誤差）を用いて，予測モデルとしての性能を比較する．MAE の値が小さいほど推薦精度が高い事を示す．

5.2 実験結果

従来手法・提案手法を用いて，80,000 件を一括学習した推薦精度の実験結果を表 1 に示す．図 1 は，学習データ数 80,000 件の中で，一括学習と逐次学習のデータ数の内訳けを 10,000 件ずつ変化させた時の結果である．図 2 は，40,000 件を一括学習した後，残りの 40,000 件を 1 件ずつ逐次学習した場合の MAE の推移である．表 1 より，従来と提案手法では，一括学習のみの推薦能力に変化がないことがわかる．しかし，図 1，2 から，提案手法は逐次学習数を増やすにつれ精度が良くなる事が分かる．以上のことから，提案手法が従来手法に比べ，精度の高い推薦を行うことができたと言える．

表 1．一括学習の推薦精度（MAE）

データ	u1	u2	u3	u4	u5	平均
従来手法	0.790	0.788	0.780	0.788	0.790	0.787
提案手法	0.790	0.788	0.780	0.788	0.790	0.787

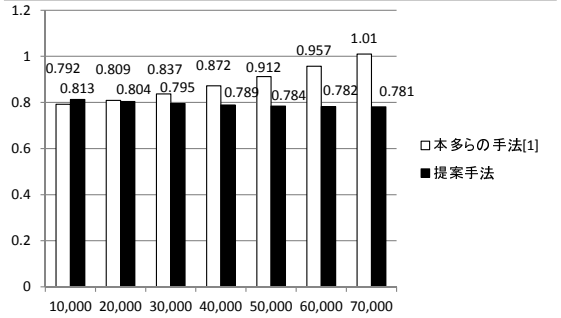

図 1．逐次学習数を変化させた結果（MAE）

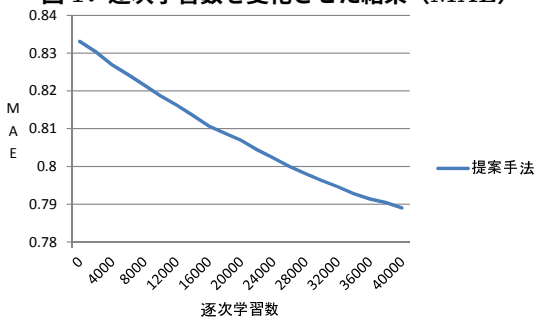

図 2．逐次学習数を 1 件ずつ増やした結果（MAE）

5.3 考察

本実験でクラスタリングを行なった結果，4 つのクラスタに対して，全ユーザのメンバシップ値がそれぞれ 0.25 程度となりほぼ均等に分割された．その結果，ユーザが単一クラスタに所属するという条件下で推薦する従来手法では推薦精度が低下したと考えられる．しかし，提案手法ではメンバーシップ値を用い，ユーザが複数クラスタに所属した状態で推薦するため，複数の嗜好を推薦結果に反映させる事が可能となった．また，表 1 と図 1，2 から，従来手法では，80,000 件を一括学習した推薦精度に比べ，40,000 件を一括学習し，40,000 件を逐次学習した推薦精度が悪い．しかし，提案手法では，80,000 件を一括学習した推薦精度と同程度の推薦結果を示していることから，逐次学習数を増やしても，推薦精度が低下しなかった事が示された．

6 まとめと今後の課題

本論文では，FCE 法を用いた逐次学習を拡張し混合比を用いた逐次学習による予測モデルを提案し，映画評価データに用いた評価実験により，その有効性を示した．今回の手法は，クラスタ数を実験的に決めたため，最適なクラスタ数を考慮した手法の構築が今後の課題である．

参考文献

[1] J.C.Bezdak,C.Coray,R.gunderson and J.Watson, “Detection and Characterization of Cluster Substructure ”, SIAM, Vol.40,No.2,pp.358-372

[2] 本多克宏, 市橋秀友, 野津亮, “線形クラスタリングのための逐次学習アルゴリズムと協調フィルタリングへの応用”, 22nd Fuzzy Symposium,pp.129-pp.132, 2006

[3] 上杉亮, 本多克宏, 市橋秀友, 野津亮, “線形ファジィクラスタリングに基づく混合データベースの局所的な主成分分析”, 日本知能情報ファジィ学会誌, vol.19, No.3, pp.287-298, 2007

[4] MovieLens Web Page; http://www.movielens.org/