混合制約付き潜在ディリクレ配分法に基づく協調フィルタリングに関する研究

1X08C060-4　坂本 俊輔
指導教員　後藤 正幸

1　研究背景・目的

近年，情報技術の進展により，EC サイト等の Web サービスで扱う情報やアイテムの数が膨大になっている．ユーザの嗜好の多様化も伴い，ユーザの嗜好を満たした情報やアイテムを自動で推薦するシステムの重要性が高まっている．このような推薦システムの代表的な手法として，ユーザ間の過去の購買履歴情報を用いて推薦を行う協調フィルタリングがあり，確率モデルやベクトル空間を用いた手法など，様々な手法が既に提案されている．

確率モデルを用いた協調フィルタリングに関する研究として，潜在ディリクレ配分法 (以下 LDA) を協調フィルタリングに適用した岩田らの研究が挙げられる [1]. 岩田らの研究ではユーザ，アイテム間に潜在クラスを導入し，アイテムの生起確率を潜在クラスの条件付き確率によって表現している．しかし，LDA の確率モデルは，各潜在クラスに対し，全ユーザと全アイテムの所属確率が割り当てられているため，潜在クラスの数を変化させると，そのパラメータ数は，ユーザ数とアイテム数の合計に比例して大幅に増減してしまう．そのため，適切な潜在クラス数を選択しても，複雑すぎるモデルであったり，逆にシンプルなモデルである可能性がある．

一方，ベイズ統計の分野では，考え得る全てのモデルを混合することでベイズ最適な予測が与えられ，予測精度が向上することが知られている [2]．しかし，先に述べたように，LDA ではモデルの複雑さが大幅に変化してしまうため，ベイズ最適な予測が有効となるモデルクラスを構成できない．

以上の議論から，本研究ではまず，協調フィルタリングにおいてより当てはまりの良い統計モデルを探索し，推薦精度を向上させるため，特定の潜在クラスへの所属確率パラメータの値を 0 と制約した，制約付き LDA を提案する．さらにそれらのモデルを混合した混合制約付き LDA を提案する．この方法は，あるモデルクラスの下でモデルを混合するので，ベイズ最適な予測を与える．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，提案手法の有効性を示す．

2　準備

2.1　推薦システム

推薦システムとは，購買履歴からユーザの嗜好を推定し，アイテムを推薦するシステムのことである．いま，ユーザを $u \in \{1, \cdots, U\}$, アイテムを $i \in \{1, \cdots, I\}$ とする．このとき，ユーザがアイテムを購買した場合は 1, 未購買の場合は 0 をとる購買履歴データの行列を $\boldsymbol{R} = (R_{u,i})$, $1 \leq u \leq U$, $1 \leq i \leq I$ と定義し，未購買アイテムの中からユーザが好むと予測されるアイテムを推薦する．

2.2　潜在ディリクレ配分法 (LDA)

LDA は，アイテムが潜在クラスに基づいて生成される過程を確率的に表現したモデルであり [1]，潜在クラスによってユーザの嗜好の多様性を表現することができる．いま，潜在クラスを $k \in \{1, \cdots, K\}$, K を潜在クラス数とする．LDA では，ユーザ u がアイテム i を購入する確率を，ユーザ u がある潜在クラス k に所属する確率と，その潜在クラスでアイテム i が生起する確率の 2 つの要素に行列分解することで算出する．ユーザ u がアイテム i を購買する確率 $P(i|u)$ を以下の式で表す．

$$P(i|u) = \sum_{k=1}^{K} \theta_{u,k} \phi_{k,i}. \tag{1}$$

ここで，$\theta_{u,k}$ は，ユーザ u が潜在クラス k に所属する確率を表し，これをまとめて $\boldsymbol{\theta} = (\theta_{u,k})$ と表す．また $\phi_{k,i}$ は，潜在クラス k の下でアイテム i が生起する確率を表し，$\boldsymbol{\phi} = (\phi_{k,i})$ とする．(1) 式における $\theta_{u,k}$ と $\phi_{k,i}$ はそれぞれ任意の正の値をとるパラメータ α, β であるディリクレ分布から生成されると仮定する．

以上のモデルの構造から，LDA では各ユーザ，アイテムは全ての潜在クラスへの所属確率を持つという特徴を持つ．$U = 3, I = 3, K = 3$ のときの LDA のモデルの構造の例を図 1 に示す．四角はそれぞれユーザ集合，潜在クラス集合，アイテム集合を表し，ノードは各ユーザ，潜在クラス，アイテムを表す．ユーザ—潜在クラス間の全リンクに $\theta_{u,k}$ の値が，潜在クラス—アイテム間の全リンクに $\phi_{k,i}$ の値が割り当てられる．

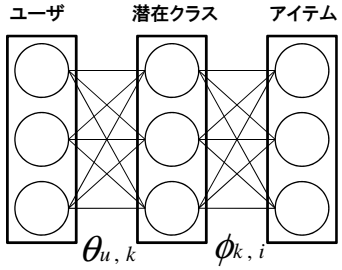

図 1．モデルの構造

3　提案手法

本研究では，推薦精度を向上させるため，特定の潜在クラスへの所属確率に制約を加えた制約付き LDA を提案すると共に，それらを混合した混合制約付き LDA を提案する．本提案では，混合の重みづけの際に対数事後確率の漸近式である BIC を用いた [2]．BIC は (2) 式で算出する．n はデータ数を表す．

$$\mathrm{BIC}_{(K)} = -\log\left(\sum_{u=1}^{U}\sum_{k=1}^{K}\sum_{i=1}^{I}\hat{\theta}_{u,k}\hat{\phi}_{k,i}\right) + \frac{K(U+I)}{2}\log n. \tag{2}$$

3.1　制約付き潜在ディリクレ配分法

従来の LDA ではユーザ，アイテムは，全ての潜在クラスへ所属する可能性を残してモデル化されている．だが一般的にユーザ，アイテムが全ての潜在クラスに所属することは考えにくく，潜在クラスへの所属確率がほぼ 0 のリンクも存在するはずである．また，(2) 式の第 2 項で用いるパラメータ数は潜在クラス数 K を 1 増加させただけで，ユーザ数とアイテム数の合計分増加し，モデルが一気に複雑になってしまう．そのため，BIC 基準の下でパラメータ数が $K(U+I)$ の近傍に，より当てはまりの良い統計モデルが存在する可能

性がある．これは，学習後 $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ の値の小さい要素を 0 として制約することで得られる．

具体的には，従来の LDA に対して，(2) 式により BIC を算出し，BIC 基準の下で最適な潜在クラス数 K を定める．次に，潜在クラス数 K+1 の LDA に対して，$\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ の値が小さい任意の $C_{\boldsymbol{\theta}}$ 個，$C_{\boldsymbol{\phi}}$ 個の要素を 0 とする．このとき，各潜在クラスにおいてパラメータの和が 1 となるように基準化する．$C_{\boldsymbol{\theta}}$, $C_{\boldsymbol{\phi}}$ の個数については，まず，パラメータ数を $(K+1)(U+I)$ から $(U+I)$ 個減らす方法，すなわち，$C_{\boldsymbol{\theta}}+C_{\boldsymbol{\phi}}=U+I$ を満たす $C_{\boldsymbol{\theta}}$, $C_{\boldsymbol{\phi}}$ の組み合わせを探索し，最も BIC の値の小さいモデルを選択する．

3.2 混合制約付き潜在ディリクレ配分法

提案された制約付き LDA のモデルの中で，BIC の値が類似しているモデルを複数選択して混合を行う．この方法は，あるモデルクラスの下で制約付き LDA を事後確率により混合するので，ベイズ最適な予測を与える．モデル m の重みを ω_m とすると，重み ω_m は，(3) 式のように与えられる．

$$\omega_m = \frac{\exp(-\mathrm{BIC}_{(m)})}{\sum_{m=1}^{M}\exp(-\mathrm{BIC}_{(m)})}. \tag{3}$$

ただし M は混合数を表す．このとき混合制約付き LDA モデルにおけるアイテム i の購買確率は，

$$P(i|u) = \sum_{m=1}^{M}\omega_m \sum_{k=1}^{K_m}\theta_{u_m,k_m}\phi_{k_m,i_m}, \tag{4}$$

で与えられる．K_m, θ_{u_m,k_m}, ϕ_{k_m,i_m} は，それぞれモデル m における潜在クラス数，潜在クラス所属確率，アイテム生起確率を表す．

3.3 学習・予測アルゴリズム

混合制約付き LDA の学習・予測アルゴリズムを以下に示す．

Step1) 各ハイパーパラメータ α, β を用いて，$\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ の初期値を生成する．

Step2) $\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ の現在値からギブスサンプリングを行って，潜在クラスへの所属確率の事後分布を近似する．

Step3) Step2 の結果から各ディリクレ分布のハイパーパラメータを更新し，再び $\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ を生成する．値が収束するまで Step2, 3 を繰り返す．

Step4) $C_{\boldsymbol{\theta}}+C_{\boldsymbol{\phi}}=U+I$ を満たす全ての $C_{\boldsymbol{\theta}}$, $C_{\boldsymbol{\phi}}$ の組み合わせに対し，制約付き LDA を作る．

Step5) 制約付き LDA の中で BIC の値が類似する M 個のモデルを選定し，(3) 式による重みづけを行い，モデルを混合する．

Step6) (4) 式の予測値が大きい順にアイテムを推薦する．

□

4 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推薦アイテムの予測実験を行い，提案手法の推薦精度の評価を行う．

4.1 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 943, アイテム数 1682 であり，学習データを 8 万件，テストデータを 2 万件とした．各ユーザに購買確率が大きいアイテムを上位 N 件推薦し，推薦したアイテムがテストデータに含まれる割合を表す Top-N 精度で評価する．潜在クラス数 K が 2〜10 までの BIC を算出したところ，BIC が最小となる潜在クラス数 K は 4 となった．そのため，制約付き LDA は，従来の $K=5$ の LDA のモデルから $C_{\boldsymbol{\theta}}+C_{\boldsymbol{\phi}}=2625$ となるようパラメータ数に制約を加えた．$C_{\boldsymbol{\theta}}$ の値は 1〜200 とし，それらの中から BIC が最小となったモデルを選択した (提案手法 1)．さらに得られた制約付き LDA の中で BIC の値が類似したモデルを 3 つ抽出し (M=3)，混合モデルを構成した (提案手法 2)．

4.2 実験結果と考察

従来手法，提案手法 1, 提案手法 2 の N=1, 2, 3, 5, 10 における Top-N 精度を以下の図 2 に示す．図 2 より全ての N に対し，提案手法の精度が勝っていることから，その有効性を示すことができた．

提案手法 1, 2 が従来手法よりも良い結果を示したのは，従来の LDA は全潜在クラスにユーザ，アイテムが所属するという特性により，$\boldsymbol{\theta}$, $\boldsymbol{\phi}$ の要素の小さい部分がノイズとなってしまったためと考えられる．また，提案手法 1, 2 に関して，N=1, 2, 3 で提案手法 1 が最も精度が良かったのは，各ユーザにとってランキング上位 3 件のアイテムは各モデルで変化が少なく，1 つのモデルを選択するだけでユーザの嗜好を反映できたためと考えられる．一方，N=5, 10 のときは提案手法 2 が最も精度が良かった．これは，上位 5 件以下のアイテムは各モデルで異なり，1 つのモデルだけでユーザの嗜好を表現することが不十分であったためだと考えられる．モデルを混合した結果，このようなユーザの嗜好の多様性を表現することができ，N=5, 10 における Top-N 精度での推薦精度が向上したと考えられる．

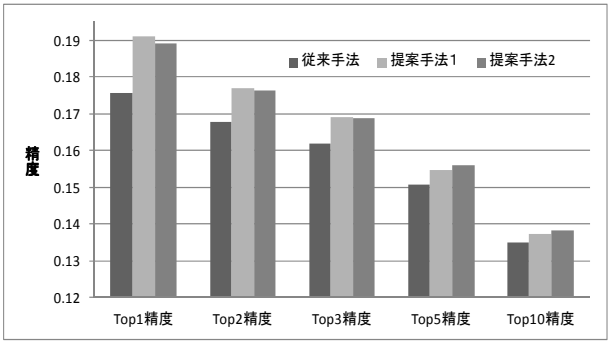

図 2. 実験結果

5 まとめと今後の課題

本研究では，LDA を用いた協調フィルタリングにおいて，潜在クラスへの所属確率に制約を加えた，制約付き LDA を提案すると共に，制約付き LDA を混合した混合制約付き LDA モデルを提案し，実験によりその有効性を示した．

今後の課題として，制約付き LDA における適切な $C_{\boldsymbol{\theta}}$, $C_{\boldsymbol{\phi}}$ の個数を決定するアルゴリズム，BIC の値の類似したモデルを探索するアルゴリズムの検討と最適な混合数の決定問題が挙げられる．

参考文献

[1] 岩田具治, 渡辺晋司, 山田武士, 上田修功,“ 購買行動解析のためのトピック追跡モデル,” 電子情報通信学会 *D*, vol. J. 93-D, no. 6, pp. 978-987, 2010.

[2] 松嶋敏泰,“ 統計モデル選択の概要,” オペレーションズ・リサーチ学会誌，vol. 41, no. 7, pp. 369-374, 1996.