社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (3)
2007－BIO－11 (3)
2007／12／20

# ディリクレ過程混合モデルに基づく離散データの共クラスタリング

桑田修平† 山田武士†† 上田修功††

ディリクレ過程混合モデルを用いた共クラスタリング手法を提案する．共クラスタリングとは，ユーザのアイテム購入履歴などのような行列形式で表現可能なデータに対して，行（ユーザ）と列（アイテム）を同時にクラスタリングする問題である．提案法は，ユーザ（もしくはアイテム）クラスごとにアイテム（もしくはユーザ）クラス数次元の多項分布を仮定し，互いに同じクラスを選択しあったときに購入行動が生じると仮定したモデルに基づいて共クラスタリングする．提案法は，ユーザ（アイテム）クラス数を事前に設定することなく共クラスタリングができ，特に，購買履歴のような欠損値を含むデータに対してより良いクラスタリング精度を示す．実データを用いた実験により，ディリクレ過程混合モデルに基づく従来手法（無限関係モデル）と比べて，より精度の高い共クラスタリング結果が得られることを示す．

# Co-clustering Discrete Data based on the Dirichlet Process Mixture Model

Shuhei Kuwata † ,Tekeshi Yamada †† and Naonori Ueda††

We propose a new co-clustering method based on the Dirichlet process mixture model (DPM). Co-clustering is the problem of simultaneously clustering rows and columns of a data matrix, such as purchase history data of users and catalog items. The proposed method assumes that each user (or item) class has a multinomial distribution over item (or user) classes to select, and a purchase occurs when both selections of user and item classes match. The proposed method can co-cluster users and items without knowing the true numbers of clusters. The experimental results show that the proposed method can provide better co-clustering results compared with IRM, another previously proposed co-clustering method based on the DPM, especially for the data matrix that contains missing data.

## 1. はじめに

企業にとって，ユーザの嗜好やニーズをより詳細に把握することがますます重要になってきている．ここで，ユーザのニーズを把握する 1 つのアプローチとして，ユーザだけでなく，アイテムも同時にクラスタリングする「共クラスタリング（Co-clustering)」を利用した手法が提案されている [1]．ユーザやアイテムの一方をクラスタリングする際に，他方のクラスタリング結果を利用しながらユーザとアイテムを相互にクラスタリングすることで，ユーザクラスとアイテムクラスから定まる“ユーザ・アイテムブロック”が求まる．ユーザのみをクラスタリングするアプローチと比べて，より粒度の細かい購買傾向に基づくクラスタリング結果が得られることになり，得られたユーザ・アイテムブロックごとに購買傾向の特徴を把握することによって，消費行動に関するより有用な情報を得ようとするアプローチである．

クラス数を予め設定する必要のある従来の共クラスタリング手法に対して，Kemp らにより提案された無限関係モデル（Infinite Relational Model）[3] は，確率モデルに基づいて，クラス数を事前に与えることなく複数の集合を同時にクラスタリングする手法であり，購買履歴を用いたユーザとアイテムの共クラスタリング問題にも適用可能である．このモデルでは，クラス数の決定基準がクラスタリングの際に用いられるモデル学習の基準に含まれており，クラス数は与えられたデータから学習される．しかし，無限関係モデルには欠損値の扱いが得意でないという問題点があり，一般に多くの欠損値が存在すると想定される購買履歴に適用するのは適切であるとは言えない．

本論文では，購買履歴のような欠損値を含むデータに適した確率モデルに基づく共クラスタリング手法を提案する．ここで，無限関係モデルと同様に，提案法もクラス数を事前に与える必要が無いという特長を有する．具体的には，ユーザのアイテム購入行動に対して，

† 株式会社 NTT データ 技術開発本部
R&D Headquarters, NTT data corporation
†† NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

無限関係モデルが「ある確率に従ってアイテムを購入する／購入しないを決定する」モデル化を行っているのに対して，提案法は「ユーザは複数候補の中からある確率に従ってアイテムを選択して購入する」モデル化を行う．つまり，アイテムを購入する事象のみに着目したモデル化を行うことで欠損値に対処する．提案法は，クラス数に対する事前分布としてディリクレ過程（Dirichlet Process: DP）を仮定した多項分布モデルに基づいており，DP 混合モデル（Dirichlet Process Mixture Model）[4] を応用したモデルと言える．

本稿の構成は以下の通りである．まず，**2** で，本研究の問題設定を示す．次に，**3** で従来法を説明し，**4** で無限関係モデルとその問題点について説明する．続いて，**5** で提案法の詳細を述べ，**6** で評価実験と考察を行う．最後に **7** でまとめと今後の課題を述べる．

## 2. 問題設定

$N$ 人のユーザと $M$ 個のアイテムからなる，$N \times M$ 行列の購買履歴データを $R$ とする．ここで，$R$ の第 $(i,j)$ 要素 $r_{i,j}$ は，ユーザ $i$ とアイテム $j$ の購買関係，

$$r_{i,j} = \begin{cases} 1, & \text{ユーザ } i \text{ はアイテム } j \text{ を購入済,} \\ 0, & \text{ユーザ } i \text{ はアイテム } j \text{ を未購入,} \end{cases}$$

を表す．本研究で扱う問題は，購買履歴データ $R$ を用いて，$N$ 人のユーザと $M$ 個のアイテムの両方をクラスタリング（共クラスタリング）することであり，ユーザのクラス数 $K$ とアイテムのクラス数 $S$ は未知とする．ここで，ユーザ $i$ $(=1,2,\ldots,N)$ の帰属クラスインデックスを $z_i \in \{1,\ldots,k,\ldots,K\}$，アイテム $j$ $(=1,2,\ldots,M)$ の帰属クラスインデックスを $w_j \in \{1,\ldots,s,\ldots,S\}$ とし，$Z=\{z_1,z_2,\ldots,z_N\}$，$W=\{w_1,w_2,\ldots,w_M\}$ とすると，具体的には，購買履歴データ $R$ が与えられた下で，ユーザのクラス割り当て $Z$ とアイテムのクラス割り当て $W$ を求めることが問題である．ただし，$r_{i,j}=1$ が欠損して $r_{i,j}=0$ となっている $r_{i,j}$ が $R$ に存在するものとする．

## 3. 従来研究

共クラスタリングの従来手法として，非負行列因子分解に基づく手法，グラフの分割に基づく手法，確率モデルに基づく手法などが挙げられる．これらの従来法に共通する点は，クラス数を予め設定しておく必要がある，という点である．しかし，クラス数は事前には未知である場合が多いため，クラス数を決定する際には，クラスタリングを行う基準とは別の基準を用意する必要がある．

## 4. 無限関係モデル

無限関係モデル [3] は，クラス数を予め設定することなく，ユーザとアイテムを共クラスタリングできる．クラス数に対しても確率分布を仮定することでクラス数の生成をモデルに組み込み，クラス数の決定をモデルパラメータの学習の枠組みで行う．無限関係モデルは，種類の異なる複数の集合間の関係を表現する確率モデルであり，購買履歴においては，履歴の類似性から，ユーザとアイテムをそれぞれ分割できる．

### 4.1 購入履歴の生成に対するモデル

ユーザの帰属クラスの割り当て $Z$ と，アイテムの帰属クラスの割り当て $W$ が与えられた下での，購買履歴データ $R$ の生成モデル $p(R|Z,W,\Theta)$ は，ベルヌーイ分布を用いて以下のように表現される．

$$p(R|Z,W,\Theta) = \prod_{i=1}^{N}\prod_{j=1}^{M} p(r_{i,j}|z_i,w_j,\Theta),$$
$$p(r_{i,j}|z_i,w_j,\Theta) = (\theta_{z_i,w_j})^{r_{i,j}}(1-\theta_{z_i,w_j})^{1-r_{i,j}}. \quad (1)$$

ここで，$\Theta$ はモデルパラメータを表す．このモデルでは，ユーザの帰属クラス $z_i$ とアイテムの帰属クラス $w_j$ により定まるユーザ・アイテムブロック $(z_i,w_j)$ ごとに購入確率 $\theta_{z_i,w_j}(\in\Theta, 0\le\theta_{z_i,w_j}\le 1)$ が割り当てられ，ユーザクラス $k$ のユーザは，アイテムクラス $s$ のアイテムを確率 $\theta_{k,s}$ で購入すると仮定している．

### 4.2 クラス数の生成に対するモデル

帰属クラスの割り当て $Z,W$ に対して，全ての可能な（一般には無限に存在する）分割に対する分布であるディリクレ過程（DP）を事前分布 $p(Z),p(W)$ として仮定する．文献 3) では，CRP を用いて DP を構成しており，例えば，$p(Z)$ は以下のように表される．

$$p(Z;\alpha) = \frac{\alpha^K \prod_{k=1}^{K}(n_k-1)!}{\alpha(\alpha+1)\cdots(\alpha+N-1)}.$$

$K$ は最終的に得られたユーザクラス数，$n_k$ はクラス $k$ の要素数，$\alpha(>0)$ は既知のハイパーパラメータをそれぞれ表す．$p(W)$ も同様の式で表される．

### 4.3 無限関係モデルの問題点

式 (1) が示すとおり，無限関係モデルは，購買履歴データ $R$ における $r_{i,j}=1$（購入済）と $r_{i,j}=0$（未購入）を対等に扱うモデルである．しかし，多くの購買履歴データの場合，$r_{i,j}=1$ に比べて $r_{i,j}=0$ となる要素数が圧倒的に多く（すなわちデータは疎で），また，ユーザは特定のアイテムを現在は未購入であっても存在を知らないだけで，将来的には購入する（$r_{i,j}=0$ の箇所が $r_{i,j}=1$ となる）可能性がある．すなわち，

これらを欠損値として扱う必要があり，購買履歴データ $R$ は数多くの欠損値を含むデータであると言える．つまり，$r_{i,j}=1$ と $r_{i,j}=0$ を対等に扱う無限関係モデルを，購買履歴データに適用する上では限界がある．

## 5. 提 案 法

無限関係モデルは**「ユーザは，アイテムごとに購入する／購入しないを確率的に決定する」**という仮定に基づくモデルであるが，提案法は**「ユーザは，複数のアイテムの中から購入するアイテムを確率的に選択し購入する」**という仮定に基づいてモデル化を行う．このような仮定の下では，購入する事象（$r_{i,j}=1$）のみが考慮される．つまり，$r_{i,j}=0$ に該当するデータを生成モデルに含めないことで，欠損値に対処する．ここで，ユーザの帰属クラスの割り当て $Z$ とアイテムの帰属クラスの割り当て $W$ の生成については，無限関係モデルと同様に DP を事前分布として仮定する．

### 5.1 購買履歴の生成に対するモデル

ユーザの帰属クラスの割り当て $Z$ と，アイテムの帰属クラスの割り当て $W$ が与えられた下での，購買履歴 $R$ の生成モデル $p(R|Z,W,\Theta,\Phi)$ を，多項分布を用いて以下のように表現する．

$$p(R|Z,W,\Theta,\Phi)=\prod_{i=1}^{N}\prod_{j=1}^{M}p(r_{i,j}|z_i,w_j,\Theta,\Phi),$$

$$p(r_{i,j}|z_i,w_j,\Theta,\Phi)=\left(\theta_{z_i,w_j}\phi_{w_j,z_i}\right)^{r_{i,j}}.$$

ここで，$\Theta,\Phi$ はモデルパラメータを表し，$\theta_k=\{\theta_{k,1},\theta_{k,2},\ldots\}(\theta_k\in\Theta,0\le\theta_{k,s}\le 1)$ は，ユーザクラス $k$ のアイテムクラス選択確率（アイテムクラス数次元の多項分布のパラメータ）を表し，$\phi_s=\{\phi_{s,1},\phi_{s,2},\ldots\}(\phi_s\in\Phi,0\le\phi_{s,k}\le 1)$ は，アイテムクラス $s$ のユーザクラス選択確率（ユーザクラス数次元の多項分布のパラメータ）を表す．モデルパラメータ $\Theta,\Phi$ は以下の式を満たす：$\sum_{s'=1}^{S}\sum_{k'=1}^{K}\theta_{k,s'}\phi_{s,k'}=1$.

このモデルは，以下の 3 つのステップを経て，履歴 $r_{i,j}=1$ が生成されると仮定するモデルである．

1. ユーザ $i$ はアイテムクラス選択確率 $\theta_{z_i}$ に従って，アイテムクラス $s'$ を選択，
2. アイテム $j$ はユーザクラス選択確率 $\phi_{w_j}$ に従って，ユーザクラス $r'$ を選択，
3. $s'=w_j$，かつ，$r'=z_i$ であるとき，ユーザ $i$ はアイテム $j$ を購入する（$r_{i,j}=1$ となる）．

前述のとおり，無限関係モデルでは，ユーザがアイテムを購入する事象と購入しない事象の両方を対等に考慮しているのに対して，提案モデルでは，ユーザがアイテムを購入する事象（ユーザとアイテムが互いに選択しあう事象）のみを考慮する．これにより，購買履歴データにおける欠損部分，つまり，今は購入していないが将来購入されるアイテムに対処する．

**表 1** ギブズサンプリングに基づく学習アルゴリズム

入力: $R,\alpha,\beta,\gamma,\eta$　　出力: $Z,W$

**初期化**：$Z,W$ を初期化する．

**ユーザクラス更新ステップ**：ランダムに選んだユーザ $i$ の帰属クラス $z_i$ の値を，以下の確率に従ってサンプリングした新たなクラス $k^*$ に更新する．

$$P(z_i=k^*|R,Z_{-i},W;\alpha,\beta,\gamma,\eta).$$

**アイテムクラス更新ステップ**：ランダムに選んだアイテム $j$ の帰属クラス $w_j$ の値を，以下の確率に従ってサンプリングした新たなクラス $s^*$ に更新する．

$$P(w_j=s^*|R,Z,W_{-j};\alpha,\beta,\gamma,\eta).$$

**クラスの分割・併合**：事後確率値に変化がなくなった場合，ユーザクラスもしくはアイテムクラスの 1 つを分割する，または，ユーザクラスもしくはアイテムクラスのうちの 2 つのクラスを併合する．

**事後確率収束判定**：$Z,W$ に関する事後確率が収束基準を満たせば終了．満たさない場合，ユーザクラス更新ステップに戻る．

### 5.2 提案法における同時分布

購買履歴データ $R$ の同時分布は，

$$p(R,Z,W,\Theta,\Phi;\alpha,\beta,\gamma,\eta)$$
$$=p(R|Z,W,\Theta,\Phi)p(Z;\alpha)p(W;\beta)p(\Theta;\gamma)p(\Phi;\eta),$$

となる．ここで，$\alpha,\beta$ は DP，$\gamma,\eta$ はディリクレ分布，の既知のハイパーパラメータをそれぞれ表す．また，$p(\Theta;\gamma),p(\Phi;\eta)$ に対して，多項分布の共役事前分布である（対称）ディリクレ分布を仮定する．

### 5.3 学習アルゴリズム

ギブズサンプリングアルゴリズムに基づき，事後確率 $p(Z,W|R;\alpha,\beta,\gamma,\eta)$ を最大化する $Z,W$ を求める．アルゴリズムを**表 1** に示す．ここで，$Z_{-i},W_{-j}$ は，$z_i,w_j$ をそれぞれ除いた $Z,W$ を現す．このアルゴリズムは，ユーザとアイテムを適当にクラスタリングをした初期状態から，事後確率の意味で良い共クラスタリング状態へと $Z,W$ を更新するアルゴリズムであり，最終的に局所解が求まる．

表 1 のクラスの分割・併合操作は，質の悪い局所解が求まるのを防ぐために導入したヒューリスティクスであり，本論文では，文献 2) で提案されている簡便な方法を利用する．本論文の実験では，事後確率の値が 1 万回連続で変わらない度に分割・併合操作を行い，100 万回連続で事後確率の値が変化しなくなったときに学習アルゴリズムを終了させた．

## 6. 実　　験

提案法の有効性，具体的には欠損値への適応度，を検証するため，映画の評価データを用いた評価実験を行った．比較対象として無限関係モデルを用いた．

### 6.1 実験設定

実データであるMovieLensデータ☆を用いて，提案法と無限関係モデルによる共クラスタリングの欠損値への適応度を定性的に評価した．MovieLensデータは，ユーザが5点満点で映画を評価した評点履歴データであり，本実験では，評価が与えられている（与えられていない）映画を，既購入（未購入）アイテムとみなし，購買履歴データを作成した．943ユーザ，1,682アイテム，スパース性約96%のデータである．

ここで，各手法におけるハイパーパラメータ値は，人工データでの実験結果を踏まえて，提案法：$\alpha=\beta=10, \gamma=\eta=0.01$，無限関係モデル：$\alpha=\beta=10, \gamma_0=\gamma_1=0.01$ とした．また，$Z, W$ の初期化の際のユーザクラス数／アイテムクラス数はそれぞれ100とした．

### 6.2 実験結果の定性的評価

購買履歴データ $R$ に対する提案法と無限関係モデルの欠損値への適応度を定性的に比較・評価する．ここで，**図1**，**図2**は，それぞれ，提案法と無限関係モデルを適用した結果である．図1，図2はそれぞれ行列 $R$ を表現しており，黒い点は $r_{i,j}=1$（既購入），白い点は $r_{i,j}=0$（未購入）をそれぞれ表す．ただし，これらの図は，共クラスタリングによって得られたユーザクラス，アイテムクラスごとにユーザとアイテムをソートしており，クラスの区切りを実線で表している．また，左上から要素数の大きい順に各クラスを並べている．各手法により得られたユーザクラス数／アイテムクラス数は，無限関係モデルにおいて36／36，提案法において49／33であった．図1，図2から，どちらの手法によっても，ユーザ・アイテムブロック単位で購買履歴（黒の部分）が密集している結果が得られることが分かる．ただし，無限関係モデルでは，購入履歴の少ないユーザ・アイテム群を一括りにして一番大きなユーザ・アイテムブロック（図2の左上部分）を構成してしまっており，購入履歴データの大部分を占める $r_{i,j}=0$ の影響を受けていることが分かる．これに対して，提案法は，$r_{i,j}=1$（黒）の部分が全体的に分布しており，欠損値を含んだデータに適応した結果が得られていることが分かる．

## 7. まとめ

本稿では，ディリクレ過程混合モデルに基づく離散データに対する共クラスタリング手法を提案した．提案法は，事前にクラス数を与えることなく，ユーザとアイテムを同時にクラスタリングする．映画の評価データを用いた実験により，無限関係モデルと比べて，より適切な共クラスタリング結果が得られることを確認した．今後は，頻度データに適したモデル化や，大規模データへ適用した際には計算時間が課題となるため，学習の効率化・高速化などについて検討したい．

☆ http://www.grouplens.org/taxonomy/term/14..

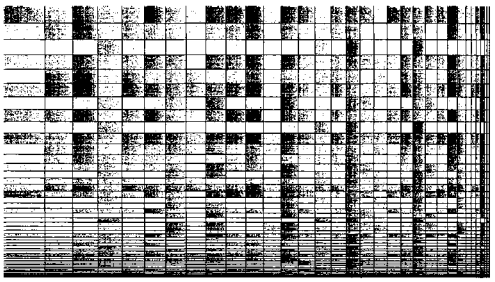

**図1** 提案法による共クラスタリング結果．

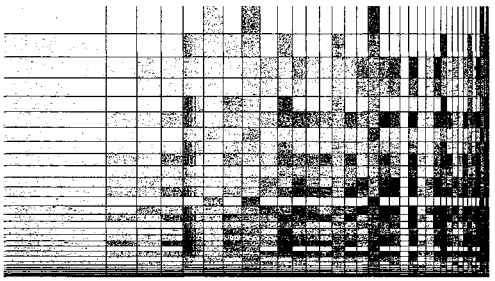

**図2** 無限関係モデルによる共クラスタリング結果．

## 参考文献


1) Deodhar, M. and Ghosh, J.: A Framework for Simultaneous Co-clustering and Learning from Complex Data, *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, San Jose, California, US, pp. 250–259 (2007).
2) Jain, S. and Neal, R. M.: A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model, *Journal of Computational and Graphical Statistics*, Vol. 13, pp. 158–182 (2004).
3) Kemp, C., Tenenbaum, J., Griffiths, T., Yamada, T. and Ueda, N.: Learning Systems of Concepts with an Infinite Relational Model, *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, Boston, Massachusetts, US, pp. 381–388 (2006).
4) McAuliffe, J. D., Blei, D. M. and Jordan, M. I.: Nonparametric empirical Bayes for the Dirichlet process mixture model, *Statistics and Computing*, Vol. 16, No. 1, pp. 5–14 (2006).