隠れセミマルコフモデルに基づくユーザの嗜好持続性を考慮した商品分析手法に関する研究

1X18C057-5　土屋希琳
指導教員　後藤正幸

1. 研究背景・目的

近年，インターネット上の様々な動画配信サービス間では顧客獲得のための競争が日々激化しており，自社のサービスを長期的に利用してもらうための施策立案が非常に重要である．そのため，機械学習などを用いてユーザの視聴履歴データを分析し，各ユーザの視聴傾向に沿った施策を検討することは，重要なアプローチの 1 つである．しかし，動画配信サービスが取り扱う動画作品（アイテム）は日用品などとは異なり，視聴（消費）後のユーザのリアルタイムな嗜好に強く影響を与えると考えられる．すなわち，あるアイテムを消費した後のユーザの嗜好状態は，そのアイテムの影響により，「継続して同シリーズのアイテムや類似アイテムを好む」場合や「別系統のアイテムを好む」場合など，アイテムがユーザの嗜好持続時間に対して与える影響（アイテム下の嗜好持続時間確率）によって決定されると考えられる．したがって，長期的に自社のサービスを利用してもらうためには，アイテム下の嗜好持続時間確率に関する特徴を明らかにし，それに基づき新たに配信する作品の選択や現状の配信作品の嗜好持続時間の観点からの評価などが重要である．

時間経過によるユーザの嗜好の変化を捉えて，ユーザの購買行動をモデル化することが可能なトピックモデルの一種として，隠れセミマルコフモデル（HSMM）[1] が提案されている．HSMM は，潜在的な因子により，現在の嗜好状態から次の嗜好状態を予測するモデルであり，嗜好の持続を扱うことができる．従来 HSMM は，ユーザが次に消費するアイテムを予測するためのモデルとして活用されてきた．しかし，数ある時系列性を考慮したトピックモデルにおいて，HSMM の特徴はユーザの嗜好持続時間確率を推定可能な点であり，この値を含む HSMM によって推定されるパラメータを用いて，アイテム下の嗜好持続時間確率を算出・分析することができれば，従来の分析手法では得られなかった重要な知見を発見することが望める．

そこで本研究では，HSMM の特性を最大限活用し，アイテム下の嗜好持続時間確率の分布を用いたアイテムクラスタリングによる分析プロセスを提案する．これにより，配信作品の選択や評価が可能となり，自社のサービス向上に繋がることが期待できる．また，実際の動画配信サービス [2] における評価履歴データに提案手法を適用し，分析結果と考察を述べると共にその有用性を示す．

2. 隠れセミマルコフモデル

時系列性を考慮したトピックモデルの 1 つである隠れマルコフモデル（HMM）[3] は，ユーザの嗜好（状態）を全ての期間において常に一定の確率で遷移すると仮定している．故に，同一状態が持続する確率は時間の経過とともに指数関数的に減少する．これに対して HSMM は，ユーザの嗜好がある程度の期間持続するようなデータを適切にモデル化するために，持続時間の概念を導入している．なお，嗜好持続時間には，予め上限となる時間（最大持続時間 M）を設けた上でモデル化を行う．ここで，ユーザ u の時刻 t における嗜好の状態を $Z_u^t \in \{1, \cdots, k, \cdots, K\}$，その状態の持続時間を $D_u^t \in \{1, \cdots, d, \cdots, M\}$，また初期状態が k である確率（初期状態確率）を π_k としたとき，HSMM が推定するモデル構造のイメージを図 1 に示す．

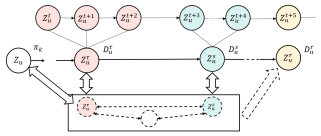

図 1: HSMM が推定するモデル構造のイメージ

3. 提案手法

3.1. 概要

これまでの HSMM の主たる応用では，ユーザの嗜好持続時間を考慮しつつ，ユーザが次に消費するアイテムを予測することを目的としていた．これに対し本研究では，HSMM から推定されるパラメータを用いて各アイテム下の嗜好持続時間確率を算出し，各アイテムの特徴を「そのアイテムを消費したユーザの嗜好持続時間に対して与える影響」の観点から分析する方法を提案する．アイテム下の嗜好持続時間確率が分析可能になると，動画配信サービスにおける配信作品の選択や評価などに活用可能である．そこで本研究では，HSMM によって得られるパラメータをもとにアイテム下の嗜好持続時間確率を算出する．さらに効率的に全体の傾向や特徴的なアイテムを確認するために，アイテム下の嗜好持続時間確率をもとにクラスタリング分析を行う．

3.2. 提案手法の詳細

任意の嗜好持続時間を D，アイテム集合を I，ユーザ u の時刻 t における消費したアイテム集合を I_u^t と定義する．このとき，アイテム i を消費したユーザの嗜好が d 期間持続する確率 $P(D = d | i \in I)$（アイテム下の嗜好持続時間確率）を式 (1) で定義する．

$$P(D = d | i \in I) = \frac{P(D = d, i \in I)}{P(i \in I)} = \frac{P(D = d, i \in I)}{\sum_d P(D = d, i \in I)} \quad (1)$$

$$\begin{aligned} P(D = d, i \in I) &= \sum_t \sum_k P(Z_u^t = k, D_u^t = d, i \in I_u^t) \\ &= \sum_t \sum_k P(i \in I_u^t, Z_u^{t-d+1:t} = k) \\ &= \sum_t \sum_k P(i \in I_u^t | Z_u^{t-d+1:t} = k) P(Z_u^{t-d+1:t} = k) \end{aligned} \quad (2)$$

さらに，式 (2) における $P(Z_u^{t-d+1:t} = k)$ はユーザ u が期間 $t - d + 1$〜t に状態 k である確率，$P(i \in I_u^t | Z_u^{t-d+1:t} = k)$ はその状態 k のユーザ u が時刻 t に

アイテム i を消費する確率を表す．これらの確率を，状態 j から状態 k への状態遷移確率 $A_{j,k}$，状態 k が d 期間持続する確率 $D_{k,d}$，持続時間 d の状態 k におけるアイテムを消費する確率 $p_{k,d}$，持続時間 d の状態 k におけるアイテムの消費回数 $r_{k,d}$，状態 k におけるアイテム i の消費確率 $\theta_{k,i}$ を用いて，それぞれ式 (3)，(4) のように表す．

$$\sum_t \sum_k P(i \in I_u^t | Z_u^{t-d+1:t} = k) = \sum_k \left(1 - \frac{(1-p_{k,d})^{r_{k,d}}}{(1-p_{k,d}(1-\theta_{k,i}))^{r_{k,d}}}\right) \quad (3)$$

$$\sum_t \sum_k P(Z_u^{t-d+1:t} = k) = \sum_t \sum_k \sum_{j \in S \setminus k} \sum_{d'=1}^{M} A_{j,k} D_{k,d} P(Z_u^{t-d} = k, D_u^{t-d} = d') = \sum_t \sum_k \sum_{j \in S \setminus k} \sum_{d'=1}^{M} A_{j,k} D_{k,d} P(Z_u^{t-d-d'+1:t-d} = k) \quad (4)$$

式 (2)〜(4) より，式 (1) は，HSMM における推定済みのパラメータや確率のみを含んだ計算に帰着する．ただし，初期状態における $P(Z_u^{t-d+1:t} = k)$ は，式 (5) で算出される．

$$P(Z_u^{1:d} = k) = \pi_k D_{k,d} \quad (5)$$

以上より，各アイテム下の嗜好持続時間確率を算出する．この嗜好持続時間確率を用いて，アイテムごとの特徴を詳細に把握することが可能となる．さらに，得られた全アイテム下の嗜好持続時間確率を用いてクラスタリング分析を行う．

4. 実データを用いた分析実験

4.1. 実験条件と分析方法

提案手法の有用性を確認するために，Netflix における評価履歴データ [2] に提案手法を適用し，得られた結果に対する考察を行う．対象期間は 2001/01〜2005/12，対象ユーザ数は 1,212，アイテム（作品）数は 5,264 である．また Zhang ら [1] を参考に，HSMM における時刻の単位は 1ヶ月，最大持続時間 M は 5ヶ月，状態数 K は 8 とする．さらに，本研究では得られたアイテム下の嗜好持続時間確率 $P(D=1|i \in I)$〜$P(D=5|i \in I)$ の高さの順位が一致したアイテムごとにクラスタリングし，各クラスのその特徴を観察する．

4.2. 結果と考察

提案手法により，4 つのクラスが得られた．各クラスに所属する作品例と作品数の全体に対する割合を表 1 に示す．また，各クラスに対する作品下の嗜好持続時間確率の分布の平均を図 2 に示す．

表 1: 各クラスに所属する主な作品と作品割合

クラス	作品例	作品割合
1	Back to the Future，Star Wars:Vol.1	7.05%
2	Die Hard，Harry Potter:Vol.1	31.74%
3	Titanic，Jurassic Park，Armageddon	61.19%
4	Day for Night	0.02%

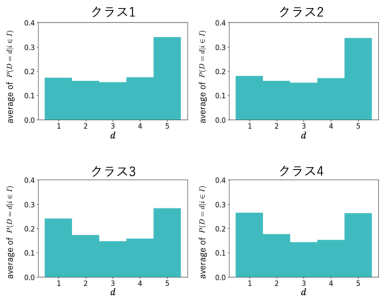

図 2: 嗜好持続時間確率分布に関するクラスごとの平均

これらより，クラス 1，2 は他のクラスと比較して，持続時間 $d=5$ の確率が非常に高い．また，長編シリーズの映画や同一ジャンルの作品が多く存在していたことから，一度このクラスに属する作品を視聴すると，同作品の次のシリーズや同一ジャンルの作品が長期的に視聴される傾向があると考えられる．一方，クラス 3，4 は他のクラスと比較して，持続時間 $d=1$ の確率が比較的高く，一話完結の有名な作品が多く所属している．よって，自身の嗜好による影響だけでなく，話題性や有名さの影響によって単発的に視聴される傾向がある作品が所属していると考えられる．

さらに表 1 より，他のクラスに比べ，持続時間 $d=1$ の確率が最も高いクラス 4 に所属している作品の数は非常に少ないとわかる．これらの結果から，Netflix では，一度作品を見ると同様な傾向の作品が継続的に視聴されるような作品を多く取り扱っていることが示唆される．Netflix は月額制のサービスであり，視聴後もユーザの嗜好が長く持続するような作品を一定数取り揃え，長期にわたってサービス利用されることが理想である．以上より，作品下の嗜好持続時間が長い可能性の高い理想的な作品を多く取り揃えていると考察できる．

5. まとめと今後の課題

本研究では，各アイテムの特徴を「そのアイテムを消費したユーザの嗜好持続時間に対して与える影響」の観点から分析可能な手法を提案し，実データに適用することでその有用性を示した．今後の課題として，他のデータセットへの適用や適切な時間単位の決定などが挙げられる．

参考文献

[1] Haidong Zhang, Wancheng Ni, Xin Li, and Yiping Yang. Modeling the heterogeneous duration of user interest in time-dependent recommendation: A hidden semi-markov approach. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 48, No. 2, pp. 177–194, 2016.

[2] Netflix Prize data. https://www.kaggle.com/netflix-inc/netflix-prize-data, 最終アクセス日 2021/8/24.

[3] Sahoo Nachiketa, Singh vir Param, and Mukhopadhyay Tridas. A hidden markov model for collaborative filtering. *Management Information Systems Research Center, University of Minnesota*, Vol. 36, No. 4, pp. 1329–1356, 2012.