機械学習モデルの解釈手法を用いたアイテム推薦理由の説明文の生成

森澤 竣† 山名 早人‡

†早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
‡早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1
E-mail: †‡{hiroshun, yamana}@yama.info.waseda.ac.jp

あらまし 推薦システムにおいて，アイテムの推薦理由をユーザに提示することは，推薦の効果や透明性，そしてユーザの満足度を向上させることが示されている．推薦理由を説明するためには，推薦モデルの解釈が必要となる．しかし，近年提案されている機械学習を用いた推薦モデルは解釈が難しいブラックボックスとなっているものが多く，学習済みモデルの状態から推薦理由を解釈することは困難である．筆者らはこれまでに，機械学習モデルの予測結果を解釈するアルゴリズムである LIME を推薦システムに適用し，解釈結果から推薦理由の説明をユーザに提示する手法の検討を行った．提案手法は大きく 2 つの手法から構成される．1 つ目は LIME を学習済み推薦モデルに対して適用し，ある予測における特徴量の重要度を算出する手法である．2 つ目は算出した特徴量の重要度を用いて，ユーザに提示するのに自然な推薦理由の説明文を自動生成する手法である．本稿では，2 つ目の推薦理由の説明文を自動生成する手法に関して，推薦システム上でユーザに提示する説明として妥当な説明文を生成するためにより具体的な検討を行った結果について述べる．評価実験では，LIME の推薦モデルへの適用による特徴量の重要度算出の精度を評価し，サンプリング数の値を上げると LIME による説明が推薦モデルに対して忠実になる結果が得られた．さらに，自動生成した推薦理由の説明文に対して，被験者によるアンケート評価を行い，提案手法はランダムで説明する特徴量を決定する手法に比べて，推薦の効果，透明性，信頼性，満足度の面で優位な結果が得られた．

キーワード 情報推薦，機械学習

1. はじめに

説明可能な推薦システムとは，アイテムの推薦に加えて，推薦理由をユーザに説明することができるシステムを意味する．推薦システムのアルゴリズムの一つとして有名な協調フィルタリングは解釈性が高く，説明を提供しているサービスが多く存在する．協調フィルタリングを使用した推薦システムでは，例えば，「あなたに似たユーザはこのような商品を購入しています」「あなたが購入した商品に関連したこのような商品を推薦します」といった説明をすることが可能である．こうした推薦理由の説明は，推薦の効果や透明性，そしてユーザの満足度を向上させることが示されている[1].

一方で，近年は行列分解や機械学習アルゴリズムを用いた新たな推薦モデルが多く提案されている．これらの最先端のアルゴリズムを使用した推薦システムは，より高い推薦精度を発揮するが，解釈性が低く，ユーザに対して説明を与えることが難しい．そこで，行列分解や機械学習アルゴリズムを用いながら解釈性の高い推薦モデルを実現するための研究が近年行われている．

機械学習の解釈性を高めるための研究は，情報推薦の分野のみならず，機械学習を活用するあらゆる分野において関心が高まっている．説明が可能な機械学習アルゴリズムを用いたシステムは，一般的に「説明可能な AI(XAI: Explainable AI)」と呼ばれ，DARPA[1]による研究を始めとして，様々な組織によって取り組まれている．また，Google Explainable AI[2]のように，実際にユーザに対して XAI の機能を提供するサービスも存在する．

機械学習を用いた説明可能な推薦モデルの研究としては，S. Seo ら[2]のアテンションニューラルネットワークを使って解釈を可能にする手法や，B. Abdollahi ら[3]の行列分解を使用した推薦モデルにおいて，説明可能でない学習結果にペナルティをかけることによって最終的な学習結果が説明可能となるように矯正する手法が存在する．しかし，[2][3]の手法は特定のデータやシステムに対する限定的な手法であり，応用が難しいといった問題点がある．

筆者ら[12]は，機械学習モデルの解釈を行うための手法として M. T. Ribeiro ら[4]によって提案された LIME(Local Interpretable Model-agnostic Explanations)を推薦システムに組み込むことによって説明を生成する手法を提案した．LIME は，任意の学習済み機械学習モデルに対して，特定の推論結果における各特徴量の重要度を算出するアルゴリズムである．

[1] DARPA, Explainable Artificial Intelligence, https://www.darpa.mil/program/explainable-artificial-intelligence
[2] Google, Google Cloud Explainable AI, https://cloud.google.com/explainable-ai/

本稿では，LIME によって算出された特徴量の重要度から推薦理由の説明文を自動生成する手法について，推薦システム上でユーザに提示する説明として妥当な説明文を生成するために，より具体的な検討を行った結果について，映画推薦モデルに対して適用した場合を例に述べる．さらに，本研究の目的である解釈性の低い推薦モデルに対して，提案手法を適用した際の説明の精度による評価と，被験者に対する説明の提示による評価を行った結果について述べる．また，[12]では，解釈が容易である協調フィルタリングモデルに対して提案手法を適用し説明の精度を測定したが，本稿では解釈性の低い推薦モデルに対して提案手法を適用し，説明の精度と被験者による説明の評価の両面における測定の結果について述べる．

本稿では，以下の構成をとる．2 節では説明可能な推薦システムの関連研究について述べる．3 節で提案手法について述べる．4 節では提案手法の推薦モデルへの適用時における，ユーザに対する説明の生成手法の具体例を述べる．5 節では評価実験の結果，および結果の考察について述べる．最後に，6 節でまとめを行う．

2. 関連研究

本節では，近年の行列分解や機械学習を使用した推薦モデルに対する，ユーザへの説明可能性を高めるための研究について，モデルのアルゴリズム別に分類して説明する．

2.1 行列分解を用いた説明可能な推薦モデル

本項では，行列分解(Matrix Factorization)を用いた説明可能な推薦モデルに関する研究について紹介する．行列分解モデルはユーザ・アイテム間の関係を表すモデルとして推薦システムに頻繁に用いられる．しかし，ユーザもしくはアイテムを表す潜在的な因子として仮定されている行列分解後の各要素は，それぞれの要素がどのような意味を持っているかを解釈することは難しい．

B. Abdollahi ら[3]は，行列分解を用いたレーティング予測モデルにおいて,「あなたに似たユーザはこの商品を高く評価しました.」といった説明の提示を前提とした，説明可能な行列分解モデルに関する研究を行った．ユーザ-アイテムの組み合わせに応じて，上記の説明を与えたときの説明の正確さを表す指標である Explainability(説明可能性)を，説明提示対象のユーザの過去のレーティングに対する予測レーティングの期待値を用いて定義した．さらに，行列分解の最適化における目的関数に Explainability を加えることによって，説明が正しくなるアイテムのみが推薦されるように改良を行った．また，[6]では，同様に目的関数に Explainability を組み込む手法を，制限付きボルツマンマシン(Restricted Boltzmann Machines)を用いた推薦モデルに適用する手法を提案した．

G. Peake ら[5]は，行列分解を用いたレーティング予測モデルにおいて，ユーザがアイテムに対して与えたフィードバックと，学習済み行列分解モデルの予測値の関係をアソシエーション分析を行うことによって，ユーザのどのフィードバックが出力値に対して大きく影響を与えたかを解釈する方法を提案した．また，各ユーザに対して，最も出力値に大きく貢献したフィードバックを用いて,「あなたは X を視聴したため，この映画を推薦します.」といった説明を行うことが可能である．

2.2 グラフ構造を用いた説明可能な推薦モデル

本項では，グラフ構造を用いた説明可能な推薦モデルに関する研究について紹介する．グラフベースの推薦モデルはソーシャルネットワーク関連の推薦モデルにおいてユーザ間あるいはユーザ・アイテム間の関係をグラフとして表すことによって活用されている．

X. He ら[7]は，ユーザとアイテム，そしてアイテムのアスペクト(特性)の関係を表す三部グラフを推薦モデルに導入した．アスペクトのノードは，ユーザレビューから抽出した単語レベルのアイテムの特性を表し，どのユーザがどのアイテムにどのようなレビューを記載したかによって，三部グラフを構成する．推薦モデルは，グラフ間の重みに応じてアイテムを表すノードをランク付けすることによって推薦を行う．また，グラフ構造を読み取ることによってユーザの好みのアスペクトを説明として用いることが可能である．

R. Heckel ら[8]は，購入履歴等のユーザとアイテムの関係を表す二部グラフに基づいた，重複ありのクラスタリングによる説明可能な推薦手法を提案した．各クラスタに含まれるユーザは同様の興味があると解釈することができるので,「アイテム A を購入したあなたと同様の興味を持つユーザ X はこれらの商品を購入しました」といった説明が可能である．

2.3 ニューラルネットワークを用いた説明可能な推薦モデル

本項では，ニューラルネットワークを用いた説明可能な推薦モデルに関する研究について紹介する．いずれの研究も，入力の特徴の重みを解釈しやすくするためにアテンションニューラルネットワークを使用している．アテンションニューラルネットワークとは，入力ベクトルの各要素に重みを与える層(Attention Layer)を追加したニューラルネットワークである．

S. Seo ら[2]は，入力にユーザが過去に書いたレビュー文章と，アイテムが過去に書かれたレビュー文章の単語を Bag of Words によって表現したベクトルを入力

として，ユーザがアイテムに与えたレーティングを予測するアテンションニューラルネットワークを提案した．Attention Layer における数値を，単語埋め込みベクトルに対応する各単語の重要度として解釈することができるので，数値の高い単語はユーザの興味のある特徴，もしくはアイテムに関係する特徴として，ユーザに説明することが可能である．

C. Chen ら[9]は，レビュー文章を一度畳み込みニューラルネットワークによってベクトル化し，これらのベクトルを入力として，ユーザがアイテムに与えたレーティングを予測するアテンションニューラルネットワークを提案した．Attention Layer における数値は，各レビュー文章の重要度として捉えることができるため，過去に書かれたレビュー文章をユーザに対する説明として提示することが可能である．

2.4 関連研究のまとめ

本項では，近年の機械学習モデルを使用した説明可能な推薦システムにおける既存研究を紹介した．既存研究の問題点の 1 つは，各手法がそれぞれの推薦アルゴリズムに対して解釈性を高める限定的な手法である点である．データやシステムに応じて適当な推薦アルゴリズムは変化するため，あらゆるモデルに対して適用可能な手法が求められる．また，もう 1 つの問題点として，既存モデルの解釈性を高めるために改良を加えることが，推薦モデルの精度を下げることにつながる点が挙げられる．例えば[3]では，目的変数に Explainability に基づいたペナルティを加えることによって，単なる行列分解よりも精度を下げるケースが存在する．

3. LIME を使用した推薦理由提示手法

本節では，筆者らが提案した，LIME を使用した推薦理由の提示手法[12]について述べる．3.1 項では，機械学習解釈モデルである LIME のアルゴリズムについての説明を行う．3.2 項では提案手法について説明を行う．

3.1 LIME のアルゴリズム

LIME(Local Interpretable Model-agnostic Explanations)[4]は，任意の学習済み機械学習モデルに対して，推論の結果を説明するためのアルゴリズムとして，M. T. Ribeiro らによって提案された．LIME は，推論結果の説明として，推論結果に影響を与えた任意の個数の特徴量のラベルと重要度を出力する．特徴量の選定と重要度の算出には，学習済み機械学習モデルにおける特徴空間内の特徴ベクトルと出力値を，解釈性の高い線形回帰モデルを用いて学習することによって行う．

[4]では，分類問題を推論する機械学習モデルに対して，LIME を適用して説明を生成するための手法が提案されている．さらに，[4]の筆者による実装[3]では，回帰モデルへ適用可能なように拡張されている．回帰モデルにも適用可能な LIME による説明生成のアルゴリズムを，アルゴリズム 1 に示す．

アルゴリズム 1　LIME を用いた説明の生成[4]

入力: 説明対象のモデルf, 入力ベクトルx
入力: サンプル数N, 類似度カーネル関数π_x
入力: 説明に用いる特徴量の数K
出力: 各特徴量の重要度$\boldsymbol{w}$

1. $\mathcal{Z} \leftarrow \{\}$
2. **for** $i \in \{1, 2, 3, \ldots, N\}$ **do**
3. $z_i \leftarrow \text{sample\_around}(\boldsymbol{x})$
4. $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle \boldsymbol{z_i}, f(\boldsymbol{z_i}), \pi_x(\boldsymbol{z_i}) \rangle$
5. **end for**
6. $w \leftarrow \text{K} - \text{Lasso}(\mathcal{Z}, K)$
7. **return** w

入力には，説明対象のモデル$f \in \mathbb{R}^d \rightarrow \mathbb{R}$と，入力ベクトル$\boldsymbol{x} \in \mathbb{R}^d$を与える．また，パラメータとして，サンプリングするベクトルの数Nと説明に用いる特徴量の数K，そして，$\boldsymbol{x}$との類似度を算出する類似度カーネル関数 π_xを与える．π_xは，局所性をコントロールする関数であり，式 3.1 で表される．σはカーネル幅，Dはコサイン距離やユークリッド距離といった距離関数を表す．

$$\pi_x(z) = \exp\left(-\frac{D(x,z)^2}{\sigma^2}\right) \qquad \text{(式 3.1)}$$

まず，推論結果の説明を行う入力ベクトル$\boldsymbol{x}$を中心として，ベクトルをサンプリングする(3 行目)．サンプリングした入力ベクトル$\boldsymbol{z_i}$と説明対象のモデルによる出力$f(\boldsymbol{z_i})$，そして入力ベクトル$\boldsymbol{x}$と$\boldsymbol{z_i}$の類似度$\pi_x(\boldsymbol{z_i})$を集合$\mathcal{Z}$に追加する(4 行目)．3，4 行目の操作をN回分行う．そして，z_iを説明変数，$f(\boldsymbol{z_i})$を目的変数として，K個の特徴量のみを用いた複数の線形回帰モデルの学習し，説明に最適なモデルの選定を行う(6 行目)．最適な線形回帰モデルを$\xi(\boldsymbol{x})$，候補の線形回帰モデル$g \in G$とすると，$\xi(\boldsymbol{x})$の選定は式 3.2 に基づいて行われる．$\xi(\boldsymbol{x})$には，$\boldsymbol{x}$の局所性π_xを考慮した，fのgによる近似の損失関数Lを用いる．Lの計算式を式 3.3 に示す．また，関数Ωは，モデルgの説明の複雑度を表し，gの重みのベクトルのうち非ゼロの要素数を返す関数が一般的に用いられる．

$$\xi(\boldsymbol{x}) = \boldsymbol{wx} = \underset{g \in G}{\text{argmin}}\ L(f, g, \pi_x) + \Omega(g) \qquad \text{(式 3.2)}$$

$$L(f, g, \pi_x) = \sum_{z \in \mathcal{Z}} \pi_x(\boldsymbol{z})\big(f(\boldsymbol{z}) - g(\boldsymbol{z})\big)^2 \qquad \text{(式 3.3)}$$

以上の方法によって選定された線形回帰モデルξにお

[3] GitHub - marcotcr/lime: https://github.com/marcotcr/lime

ける重み$\boldsymbol{w}$が出力となる．

アルゴリズム 1 による出力結果において，重みの絶対値が大きいほど，推論結果に大きく影響を与えた特徴量であると解釈できる．また，正の重みを持つ特徴は，出力値を正の値にするために寄与した特徴であり，負の重みを持つ特徴は，出力値を負の値にするために寄与した特徴であると解釈することができる．本手法は，特徴を表現した入力に対して出力を与えるあらゆるモデルに対して適用が可能である．

3.2 推薦理由提示手法

本項では，推薦モデルに対して LIME を適用し，算出された特徴量の重要度を用いてユーザに推薦理由の提示を行う提案手法について述べる．

図 3.1 に提案手法の模式図を示し，以下に各番号の手続きに対応した説明を記す．

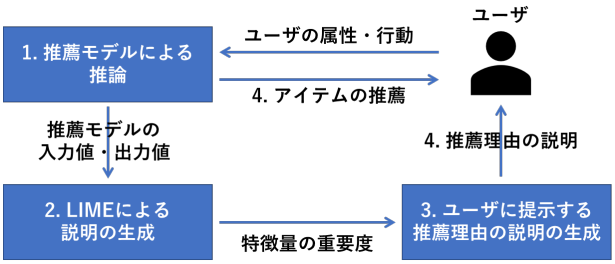

図 3.1 提案手法の模式図

1. ユーザの属性・行動データやアイテムのデータなどから，任意の学習済み推薦モデルを用いて推薦するアイテムを決定する．
2. LIME を用いて，1 で決定したアイテムの推薦に影響を与えた特徴量と，その重要度を算出する．
3. 各特徴量の重要度を用いて，ユーザに提供する推薦理由の説明文を生成する．具体的には，予め各特徴量に対応する定型文を用意しておき，重要度の高い特徴量のタイプに応じて説明に用いる定型文を出し分ける方法が考えられる．
4. 2 に基づいたアイテムの推薦と同時に，3 で生成した推薦理由の説明文をユーザへ提示する．

提案手法は，LIME を用いることによってあらゆる推薦モデルに対して適用が可能である．また，推薦モデル自体に変更を加えないため，推薦の精度に影響を与えない．以上の点は，2 節に記述した近年の機械学習を用いた説明可能な推薦モデルにおける問題点を解決する．

4. 推薦モデルに対する提案手法の適用の検討

本節では，実際の推薦モデルに対して，3 節に示した筆者ら[12]による提案手法を適用し，推薦システム上でユーザに提示する推薦理由の説明文を自動生成する手法について検討を行った結果について述べる．

4.1 使用した推薦アルゴリズム

提案手法の適用を行う推薦モデルとして，S. Rendle[10]によって提案された Factorization Machines (FM)を使用する．FM による回帰モデルを式 4.1, 式 4.2 に示す．入力ベクトル$\boldsymbol{x} = (x_1, \dots x_n) \in \mathbb{R}^n$に対する出力値 $\hat{y}$を，学習パラメータであるバイアス項$w_0 \in \mathbb{R}, \boldsymbol{w} = (w_1, \dots w_n) \in \mathbb{R}^n$と，変数間の相互作用項$\boldsymbol{v}_i = (v_{i,1}, \dots v_{i,k}) \in \mathbb{R}^k$によって表現したモデルである．

$$\hat{y}(\boldsymbol{x}) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle x_i x_j \qquad (式\ 4.1)$$

$$\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f} \qquad (式\ 4.2)$$

推薦モデルとして FM を用いる場合，$\boldsymbol{x}$をアイテムとユーザの one-hot ベクトルとして表現して学習させることによって，アイテムとユーザの学習を行う．さらに，入力変数にはアイテムとユーザの one-hot ベクトルに加えて，ユーザやアイテムの属性を数値化した特徴量も加えて学習することが可能である．

4.2 使用したデータセット・特徴量

データセットとして，Grouplens による MovieLens 1M Dataset[4]を用いた．MovieLens 1M Dataset は，ユーザが映画に与えた 5 段階のレーティングのデータセットである．ファイル ratings.dat には，6,040 人のユーザが 3,883 本の映画に対して与えたレーティングのログが 1,000,209 件含まれる．さらに，ファイル items.dat には各映画のジャンルの情報，users.dat には各ユーザの性別，年齢，職業の情報が含まれる．

データセットによって生成した FM の学習に用いる特徴量の一覧を表 4.1 に示す．本研究では，様々な特徴量による説明形式の評価を目的としているため，ユーザ・アイテムの補助情報を特徴量として使用した．さらに，各ユーザが過去にどの映画を過去に評価したかどうかを表す特徴量を「視聴履歴」として学習に使用した．

[4] MovieLens 1M Dataset | GroupLens https://grouplens.org/datasets/movielens/1m/

表 4.1 使用した特徴量

名称	表現方法	次元数
ユーザ ID	one-hot encoding を行う.	6,040
映画 ID	one-hot encoding を行う.	3,883
年齢	0~1 の間にスケーリングする.	1
性別	女性: 0, 男性: 1	1
ジャンル	one-hot encoding を行う.	18
職業	one-hot encoding を行う.	21
視聴履歴	視聴前: 0, 視聴済み: 1	3,883

4.3 FM の学習と評価

学習済み FM に対する提案手法を適用する前に，FM の学習に用いるパラメータチューニングと，予測の RMSE による評価を行った．データセットを timestamp の値の小さい順に 60%, 20%, 20%に分割し，それぞれを学習用データセット,検証用データセット,評価用データセットとした．パラメータは，検証用データに対するレーティング予測の RMSE の値が最も小さくなる値を使用した．その結果，式 4.1 に示した，変数間の相互作用項$\boldsymbol{v}_i$の次元数が8，学習における L2 正則化項の値が 5 となった．これらのパラメータを使用したときのテスト用データにおけるレーティング予測の RMSE の値は，0.930 となった．

4.4 提案手法の適用

本項では，学習済みモデルに対する提案手法の実際の適用例について，3.2 項に示した手続きの流れに準じて説明する．

(1) 推薦モデルによる推論

4.1~4.3 項に示したアルゴリズム，データセット，パラメータを用いて，FM の学習を行う．なお，パラメータチューニングではデータセットの分割を行ったが，ここでは全てのデータを学習に使用する．

また，学習済みモデルによって，各ユーザへ推薦する映画を決定する．推薦する映画は，各ユーザの未視聴の映画に対する予測レーティングの値が大きい上位 5 件とする．

(2) LIME による説明の生成

各ユーザに対して推薦する各映画における，予測時に使用した特徴量を使用して，LIME による重要な特徴量の選定と，その重要度の算出を行う．予測に使用した全ての特徴量は表 4.1 に示したとおりであるが，そのうち，ユーザに対する説明として適当でない特徴量は LIME によって選択する特徴量から除外する．本実験の場合，特徴量「ユーザ ID」は，全ユーザの one-hot 表現であり，いずれかのユーザを示す特徴量が重要であると判断されても，推薦理由の説明を行うことは困難である．同様に，特徴量「アイテム ID」も，全アイテムの one-hot 表現であり，推薦理由の説明を行うことは困難である．従って,「ユーザ ID」,「アイテム ID」を除外した,「年齢」「性別」「ジャンル」「職業」「視聴履歴」を LIME による説明の対象とする．

本実験における LIME によって生成した説明の例を表 4.2 に示す．ここで,「重み」は LIME によって近似された線形モデルの重みを表し，この重みの絶対値が大きいほど重要な特徴量であると解釈できる．また，「特徴量の値」は，推薦の予測に用いた各特徴量の値である．

表 4.2 LIME による説明の例

特徴量名	重み	特徴量の値
ジャンル_コメディ	-0.068	0
ジャンル_アクション	0.035	1
性別	0.032	1
視聴履歴_映画 A	0.021	0
視聴履歴_映画 B	0.013	1
職業_学生	0.009	1

(3) ユーザに提示する推薦理由の説明の生成

LIME によって抽出された重要度の高い特徴量とその重みから，ユーザに実際に提示する説明文の生成を行う. LIME によって抽出される特徴量には, 重みの正負と特徴量の値の組み合わせに応じて，説明に用いることが適当であるものと不適当であるものが存在する．

ここでは，表 4.2 に示した，LIME による説明の例に対して説明文を生成する場合の具体例について述べる．表 4.2 における特徴量「ジャンル_コメディ」では，重みが負の値で，特徴量の値は 0 である．つまり，対象ユーザはコメディ映画を好まないと予測しており，推薦対象の映画はコメディ映画ではないといった状況である．この場合,「この映画はコメディでないのでおすすめします.」といった説明をユーザに与えることが可能であるが，一般的に推薦の説明でネガティブな特徴量を説明することは望ましくない．従って，重みが負の値である特徴量は説明対象から除外する．また，表 4.2 における特徴量「視聴履歴_A」 では，重みが正の値で，特徴量の値は 0 である．つまり，映画 A を視聴したユーザは推薦対象の映画を好むと予測しているが，対象のユーザは映画 A をまだ見ていないといった状況である．この場合，ユーザに対する説明を与えることは困難である． 従って，値が 0 である特徴量は説明対象から除外する．以上の除外処理によって，残された「ジャンル_アクション」「性別」「視聴履歴_映画 B」「職業_学生」が説明として適当な特徴量であるといえる．

各特徴量に対する説明文は，特徴量のタイプに対応した定型文として提示することが妥当である．本実験で使用した各特徴量に対する説明文の例を表 4.3 に示す．

(4) アイテムの推薦・推薦理由の説明

推薦とともに，表 4.3 に示した説明文をユーザに対して提示する．なお，表 4.3 の説明におけるカッコ内の値は実際の特徴量の値に応じて変更する．

表 4.3 特徴量に対する説明
（カッコ内の値は実際の特徴量に応じた値が入る）

特徴量のタイプ	説明
年齢	あなたは(20)歳のため，この映画をおすすめします
性別	あなたは(男性)のため，この映画をおすすめします
ジャンル	この映画はジャンル(Drama)であるため，あなたにおすすめします
職業	あなたは(学生)のため，この映画をおすすめします
視聴履歴	あなたは(Titanic)を視聴したため，この映画をおすすめします

5. 評価実験

本節では，提案手法に対する評価実験の手法と実験結果について述べる．具体的には，4 節に示した映画推薦モデルに対する提案手法の適用時における，LIME による説明の精度の検証と，被験者に対して説明文を提示した際のアンケート評価による検証を行う．

5.1 説明の精度の評価

筆者らは[12]にて，解釈性の高い協調フィルタリングモデルに対する説明の精度の評価を行っている．その際には，モデル内部の数値を参照することによる実際の特徴量の重みを真値として評価を行うことが可能であった．一方，本稿で提案手法を適用する FM は各特徴量に対してベクトルのパラメータを持つため，特徴量の重要度を解釈することは難しい．

そこで，我々は推薦モデルに対する LIME による線形モデルの近さを評価するために，決定係数を評価指標として採用した．決定係数R^2は，標本値y_iに対する，回帰方程式による推定値f_iを用いて式 5.1 によって表すことができる．

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad (式\ 5.1)$$

ここで，標本値y_iはサンプリングされた特徴ベクトルを入力とした際の学習済み推薦モデルによる出力値，推定値f_iは LIME によって作成された線形モデルによる出力値である．なお，決定係数は 1 以下の実数を取り，1 に近いほど LIME による線形モデルが推薦モデルを説明できているといえる．

本実験では，各ユーザに対して top1 推薦を行う際の各予測を説明するモデルごとに決定係数を測定し，平均値を評価値とした．また，LIME におけるパラメータである，特徴空間内にサンプリングするベクトルの数(アルゴリズム 1 におけるN)を 1,000 から 10,000 の間の 1,000 刻みで変化させ，評価を行った．

決定係数による測定結果を図 5.1 に示す．サンプリング数の値を上げると LIME による説明が推薦モデルに対して忠実になることが確認できる．

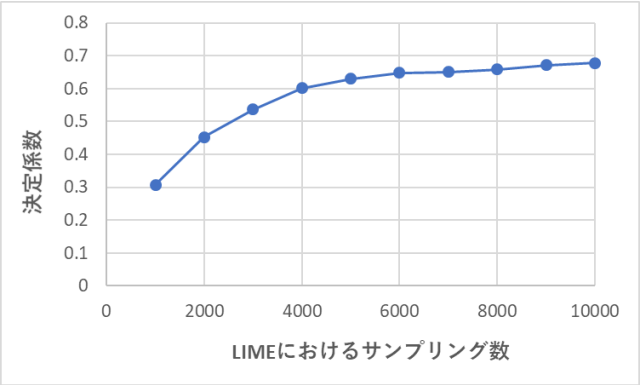

図 5.1 決定係数の測定結果

5.2 説明文の被験者による評価

被験者による説明の評価では，4 節で述べた学習済み FM モデルを用いて，被験者のコンテキストに応じた推薦映画を決定し，推薦の説明に用いる特徴量の選択方法として，LIME を活用した提案手法の他に，比較手法として，提案手法で提示されうる特徴量の中からランダムで説明する特徴量を決定する手法を使用した．被験者は，早稲田大学内から募集した 14 人の学生であり，実験は Web アプリケーションを被験者によって操作してもらう形で行った．

まず，被験者に対して，事前に年齢，性別，過去に視聴したことのある映画を問うアンケートに回答してもらい，この情報に応じて 4.1 項から 4.3 項で述べた学習済み FM モデルに対する入力ベクトルを作成し，推薦映画を決定した．各被験者の回答した，過去に視聴したことのある映画の件数は，最小 4 件，最大 86 件，平均 22.5 件であった．

提案手法による特徴量の選択では，まず推薦対象のユーザと映画を表すベクトルに対して，LIME におけるサンプリング数を 5,000 として特徴量の重要度を算出する．そして，4.4 項に示した，説明として適当であり，最も重要度の高い特徴量を選択する．ランダムによる特徴量の選択では，推薦映画と異なるジャンルや被験者が視聴していない映画といった，説明として適当でないものを除いた，提案手法で提示されうる特徴量の中からランダムに説明する特徴量を選択する．それぞれの手法によって選択した特徴量は，表 4.3 に示した説明形式によって，被験者に対して推薦映画とともに提示を行った．

被験者に対して表示した推薦映画と推薦理由の説明文の画面の例を図 5.2 に示す．被験者には，表示された説明に対して，以下に示す 6 つの質問を行い，それぞれの質問に対して，5 段階のリッカート尺度(全くそう思わない，あまりそう思わない，どちらとも言えない，ややそう思う，非常にそう思う)の選択肢を選んでもらった．質問項目は S. Chang ら[11]の被験者実験を参考とした，推薦システムにおける説明の満たすべき，推薦の効果(質問 1)，透明性(質問 2)，信頼性(質問 3,4)，満足度(質問 5,6)を検証するものである．

質問1. この説明によって,この映画に興味を持った
質問2. この説明は推薦システムの透明性を向上させる
質問3. この説明は信頼出来る
質問4. この説明は正しいと思う
質問5. この説明は分かりやすい
質問6. この説明は役に立つ

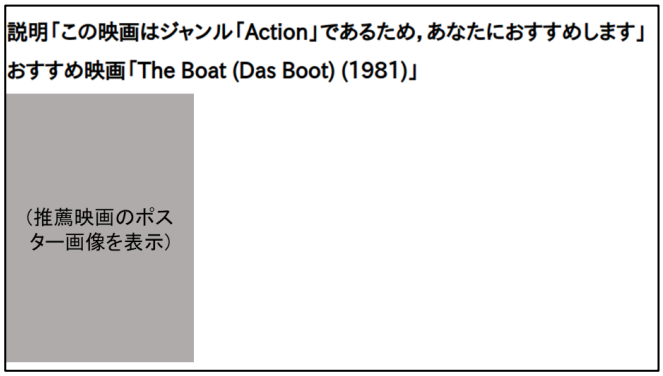

図 5.2 被験者に表示される推薦映画と説明文の例

被験者は Top 1 の推薦映画から順番に，Top N までの各映画に対する説明の評価を繰り返し，各被験者に最小 21，最大 50 の映画と説明のセットを評価してもらった．なお，提示される説明は，先述した LIME による提案手法と，ランダムで説明する特徴量を決定する比較手法が 1:1 の割合でいずれかの手法によって生成されるようにした．

提案手法と比較手法のそれぞれでアンケート結果を集計した割合を図 5.3 に示す．図 5.3 より，比較手法よりも提案手法のほうが各質問に対する評価が高いことがわかる．また，各質問に対する回答の頻度から，提案手法と比較手法の有意差を求めるためにマン・ホイットニーの U 検定を行い，p 値を算出した．その結果，質問 1 に p 値が 0.0313，質問 2 から質問 5 は p 値がすべて 0.0001 未満となり，有意水準 5%で有意であった．

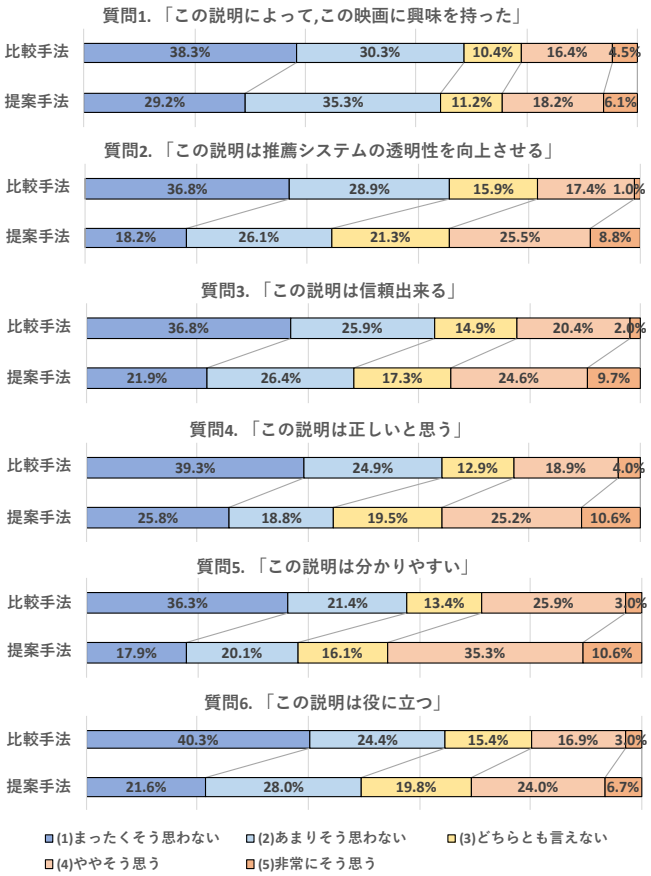

図 5.3 各質問に対する回答の割合

6. おわりに

本論文では，あらゆる推薦モデルに対して推薦理由の説明を可能とするために，機械学習の解釈手法である LIME を推薦モデルに対して適用する方法を提案した．また，解釈性の低い推薦モデルに対して，提案手法を適用し，推薦システム上でユーザに提示する説明として妥当な説明文を生成する手法について検討を行った．

評価実験では，LIME の推薦モデルへの適用による特徴量の重要度の算出の精度の評価を行い，サンプリング数の値を上げると LIME による説明が推薦モデルに対して忠実になることを検証した．

また，被験者による推薦理由の説明文の評価では，提案手法はランダムで説明する特徴量を決定する手法に比べて，推薦の効果，透明性，信頼性，満足度の面で有意な結果が得られた．

本手法を使用した研究の今後の予定としては，1)ニューラルネットワークなどの他のアルゴリズムを使用した推薦モデルに対する本手法の適用，2)他の特徴量選択手法を用いた説明生成との比較，3)実際の推薦システムへの本手法の適用によるオンライン評価の実施，などを検討している．

参 考 文 献

[1] J. L. Herlocker, J. A. Konstan, and J. Riedl, “Explaining Collaborative Filtering Recommendations,” in Proc. of ACM CSCW, pp. 241–250, 2000.

[2] S. Seo, J. Huang, H. Yang, and Y. Liu, “Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction,” in Proc. of ACM RecSys, pp. 297–305, 2017.

[3] B. Abdollahi and O. Nasraoui, “Using Explainability for Constrained Matrix Factorization,” in Proc. of ACM RecSys, pp. 79–83, 2017.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’ Explaining the Predictions of Any Classifier,” in Proc. of ACM KDD, pp. 1135–1144, 2016.

[5] G. Peake and J. Wang, “Explanation Mining: Post Hoc Interpretability of Latent Factor Models for Recommendation Systems,” in Proc. of ACM KDD, pp. 2060–2069, 2018.

[6] B. Abdollahi and O. Nasraoui, “Explainable Restricted Boltzmann Machines for Collaborative Filtering,” in Proc. of ICML Workshop WHI, 2016.

[7] X. He, T. Chen, M.-Y. Kan, and X. Chen, “TriRank: Review-aware Explainable Recommendation by Modeling Aspects,” in Proc. of ACM CIKM, pp. 1661–1670, 2015.

[8] R. Heckel, M. Vlachos, T. Parnell, and C. Duenner, “Scalable and Interpretable Product Recommendations via Overlapping Co-Clustering,” in Proc. of IEEE ICDE, pp. 1033–1044, 2017.

[9] C. Chen, M. Zhang, Y. Liu, and S. Ma, “Neural Attentional Rating Regression with Review-level Explanations,” in Proc. of WWW, pp. 1583–1592, 2018.

[10] S. Rendle, “Factorization Machines,” in Proc. of IEEE ICDM, pp. 995–1000, 2010.

[11] S. Chang, F. M. Harper, and L. G. Terveen, “Crowd-Based Personalized Natural Language Explanations for Recommendations,” in Proc. of RecSys, pp. 175–182, 2016.

[12] 森澤 竣, 真鍋 智紀, 座間味 卓臣, 山名 早人, “推薦システムにおける推薦理由提示手法の提案ー機械学習解釈モデルを用いてー”, 日本データベース学会和文論文誌 Vol.18-J, Article No.3, pp. 1-8, 2020.