アイテム推薦理由の説明のための特徴量選択手法の検証

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 近年，推薦システムは複雑なアルゴリズムを用いることによって精度を向上させた一方で，ユーザはアイテムの推薦理由を理解するのが困難となった．このため，推薦システムに対する透明性やユーザの満足度を向上させるために，推薦理由をユーザに対して提示することが必要となっている．近年の研究では，ブラックボックスな機械学習モデルを解釈性の高いモデルで近似する手法である LIME を推薦システムに適用することによって，推薦モデルを解釈し説明する手法が提案された．筆者らはこれまでに，LIME を推薦モデルに適用する際に，解釈モデルに用いる特徴量の数を最適な値にチューニングすることによって，解釈モデルの推薦モデルに対する忠実度(model fidelity)を向上させる手法を提案した. 本論文では, 提案手法における特徴量の選択手法の違いが model fidelity に与える影響を検証した．また，生成された推薦理由の説明文を被験者に提示し，アンケート評価を行うことにより，提案手法が推薦システムの透明性，信頼性，満足度の面で関連研究に対して優れていることを示した．最後に，従来用いられている評価指標である model fidelity と被験者の推薦システムの透明性，信頼性，満足度の評価値との間に，弱い正の相関があることを明らかにした．一方で，負の相関を持つ被験者も存在し，model fidelity の評価指標としての改善の余地があることを明らかにした．

キーワード 情報推薦，機械学習，解釈性，説明可能性

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可能なシステムである. 推薦理由の説明は, 推薦の効果や透明性, およびユーザの満足度を向上させることが過去の研究によって示されている[1].

近年，多種多様なアイテムの中からユーザが好みのアイテムを見つけ出すための支援として，推薦システムが不可欠となり，機械学習を中心とした高精度な推薦アルゴリズムが提案されている．これらの最新のアルゴリズムは推薦精度を向上させる一方で，推薦理由の解釈性が低いといった問題点が存在する. そのため, 高い推薦精度を維持しつつ，解釈性の高い推薦を行うことは重要な課題となっている.

説明可能な推薦システムに関する近年の研究は, model intrinsic approach(以下、モデル内在型と呼ぶ)と model agnostic approach(以下、モデル独立型と呼ぶ)に分類することができる[2]. モデル内在型は，推薦モデル内部の学習パラメータが解釈可能なアルゴリズムを採用することによって，説明を可能とするアプローチである. 例として, 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や，Chen ら[4]のアテンションニューラルネットワークを使用したレビューテキストの単語レベルの重要度を解釈する手法が存在する．モデル内在型は，各手法によって生成される説明が推薦モデルに対して忠実であることである一方, 特定のデータやモデル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.

モデル独立型は，推薦モデルをブラックボックスとして扱い，解釈可能なモデルを用いて，推薦モデルの入力と出力のセットを学習することで推薦理由を説明するアプローチである．モデル独立型は，推薦モデルと解釈を行うモデルが同一ではないため，完全に正確な説明を行うことができない．一方で，既存の推薦システムに手を加えることなく汎用的に説明を可能とすることができるため、近年着目を集めているアプローチである．Nóbrega ら[5]は，任意の機械学習モデルの推論に対して特徴量の重要度を計算することが可能な LIME(Local Interpretable Model-agnostic Explanations)[6]を推薦モデルに適用することによって推薦理由をユーザに説明する LIME for Recommender Systems(LIME-RS)を提案した. LIME は, 特徴空間内の局所的な部分に対してシンプルな回帰モデルによる学習を行うことによって，回帰係数を特徴量の重要度として解釈する手法である.

筆者ら[15]は，LIME-RS を拡張し，すべての特徴量を用いて解釈のための回帰モデルを学習するのではなく，最適な数の特徴量を選択し，回帰モデルに用いることによって，より解釈モデルの推薦モデルに対する忠実度を高める手法を提案した．また，筆者らが提案した手法は LIME-RS と比較して, 推薦モデルに対する解釈モデルの忠実度を測定するオフライン評価指標で

ある model fidelity@50 において，2.5-2.7%の改善を示した．さらに，LIME によって算出された特徴量の重要度を元に，ユーザに対して推薦理由を提示するための説明文の生成方法を提案した．

本稿では，[15]において提案した，最適な数の特徴量の選択において，異なる選択手法を用いたときの model fidelity の差異を検証する．また，提案手法と LIME-RS のそれぞれの解釈モデルを用いて生成した推薦理由の説明文を，被験者に評価してもらうことによって，説明文ベースでの評価の差異を測定する．

本稿では，以下の構成を取る．2 節では関連研究について述べる．3 節で提案手法に用いられる LIME のアルゴリズムについて述べ，4 節で提案手法について述べる．5 節では評価実験の結果について述べる．最後に，6 節でまとめを行う．

2. 関連研究

本節では，モデル独立型を使用した説明可能な推薦システムにおける研究を紹介する．モデル独立型は，推薦モデルをブラックボックスとして扱い，解釈可能なモデルを用いて，推薦モデルの入力と出力のセットを学習することで推薦理由を説明するアプローチである．モデル独立型は，推薦モデルと解釈を行うモデルが同一ではないため，完全に正確な説明を行うことができない．一方で，既存の推薦システムに手を加えることなく説明を可能とすることができるといった利点がある．

Peake ら[7]は，潜在因子推薦モデルにアソシエーションルールを用いることによって解釈を行う手法を提案した．学習済み潜在因子モデルにおける，入力特徴量の値と推薦アイテムとの関係にアソシエーション分析における指標である，支持度，確信度，リフト値を算出することによって，推薦アイテムに対する特徴量の寄与度を解釈した．そして，寄与度の高い特徴量を推薦の説明に使用することを提案した．

任意の機械学習モデルの推論結果を解釈する手法として，Ribeiro ら[6]によって提案された LIME は，あらゆる機械学習モデルに対しても適用可能であるといった汎用性の高さから，情報推薦の分野においても適用が行われている．Zhu ら[8]は，推薦モデルの開発者が特徴量を分析することを目的として，推薦システムにおけるクリック予測モデルに LIME を適用した．この手法はユーザにとっては理解が困難な特徴量やアイテムに対してネガティブに作用する特徴量も説明に含まれるため，ユーザへの説明を提供するためには活用が難しい．一方，Nóbrega ら[5]は，ユーザへの推薦理由の説明を目的として，説明可能な特徴量によって解釈モデルの学習を行う，推薦モデルに LIME を適用する手法である LIME-RS を提案した．LIME-RS は Peake ら[7]の手法と比較して，推薦モデルに対する解釈モデルの忠実度を測定するオフライン評価指標である model fidelity@10 において，0.086 の改善を示した．しかし，LIME-RS では，特徴量の数が増えることによって，解釈モデルの学習が困難となるといった問題点が存在する．また，[7]では，推薦モデルの推論における特徴量の重要度の算出方法のみを提案し，対象となるユーザに提示する説明を生成する手法は具体的に検討されていない．

3. LIME のアルゴリズム

本節では，提案手法に用いられる，LIME(Local Interpretable Model-agnostic Explanations)[6]のアルゴリズムについて述べる．LIME は特徴量空間内の局所空間に対して線形回帰を行うことによって，説明対象の機械学習モデルの入出力のペアを学習し，各特徴量の重要度を算出する．アルゴリズム 1 に，回帰モデルの推論に対して LIME によって説明を行う方法を示す．

アルゴリズム 1　LIME を用いた説明の生成 ([6]をもとに作成)

入力: 説明対象のモデルf, 入力ベクトル$\boldsymbol{x}$
入力: サンプル数S, カーネル幅σ, 特徴量の数K
出力: 各特徴量の重要度$\boldsymbol{w}$

$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, \ldots, S\}$ **do**
　$\boldsymbol{z_i} \leftarrow$サンプリングされた$\boldsymbol{x}$の周辺のベクトル
　$\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle \boldsymbol{z_i}, f(\boldsymbol{z_i}), \pi(\boldsymbol{x}, \boldsymbol{z_i}) \rangle$
end for
$g \leftarrow \mathcal{Z}$に対して$L(f, g)$を最小化するK個の特徴量を使用した回帰モデル
$w \leftarrow g$の回帰係数
return w

LIME は，説明対象のモデル$f \in \mathbb{R}^d \rightarrow \mathbb{R}$と入力ベクトル$\boldsymbol{x} \in \mathbb{R}^d$，そして$f$の解釈モデルの学習のためにサンプリングされる入力ベクトルの数S, 局所性を調節するためのカーネル幅σ，特徴量の数Kを設定する．まず，推論結果を説明する対象となるベクトル$\boldsymbol{x}$を中心としてベクトルをサンプリングする．次にサンプリングした入力ベクトル$\boldsymbol{z_i}$と得られた出力$f(\boldsymbol{z_i})$，$\boldsymbol{x}$と$\boldsymbol{z_i}$の局所性$\pi(\boldsymbol{x}, \boldsymbol{z_i})$のセットを$\mathcal{Z}$に追加し，これらの処理を$S$回繰り返す．

続いて，$\boldsymbol{z_i}$を入力，$f(\boldsymbol{z_i})$を正解ラベルとして，K個の特徴量を使用して線形回帰モデル$g \in G$を学習する．ここで，fを最も説明することができる最適な解釈モデルgは，式 3.1 に示す損失関数Lを最小化する．なお，サンプルはカーネル幅σを使用したカーネル関数$\pi(\boldsymbol{x}, \boldsymbol{z})$で重み付けし，局所性をもたせることによって，複雑なモデルに対する線形回帰モデルの近似を可能としている．

$$L(f, g) = \sum_{\boldsymbol{z} \in \mathcal{Z}} \pi(\boldsymbol{x}, \boldsymbol{z}) \big(f(\boldsymbol{z}) - g(\boldsymbol{z})\big)^2 \qquad (式\ 3.1)$$

最後に，gの回帰係数$\boldsymbol{w}$を出力とする．ここで，wの

絶対値が大きい要素ほど，対応する特徴量が予測結果に与える影響が大きいと解釈することが可能である．また，正の要素に対応する特徴量は，出力値を正の値にするために寄与した特徴であり，負の要素に対応する特徴量は，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．

4. 提案手法

本節では，筆者ら[15]が提案した推薦の解釈モデルに用いる特徴量の数を最適な値にチューニングすることによって，推薦モデルに忠実な解釈モデルを学習し説明を生成する提案手法について説明する．本手法はユーザに対して推薦理由の説明に LIME を用いた最先端の手法である LIME-RS の 2 つの問題点を以下のアプローチによって解決した．(1)特徴量が増えると解釈モデルの学習が困難となる問題点については，解釈モデルに使用する特徴量の数をチューニングすることによって解決した．(2)ユーザに提供する説明の生成方法が考慮されていない点については，[15]で初めて検討し，LIME によって得られる重要な特徴量から推薦理由の提示として適さない特徴量を除外して，テンプレートを用いて説明を生成する方法を提案した．図 4.1 に本手法の模式図を示し，以下に各ステップの詳細を説明する．

1) ユーザへのアイテム推薦

任意の学習済み推薦モデルによって，ユーザの属性や行動データ等に基づき，ユーザに推薦するアイテムを決定する．

2) LIME による推薦モデルの解釈

このステップでは，推薦モデルがターゲットユーザに対してアイテムを推薦する理由を，LIME を拡張した手法によって解釈する．

a) **入力ベクトルのサンプリングと推薦モデルによる出力値の取得**: 3 節で示したとおり，LIME における入力ベクトルのサンプリングでは，単純に特徴空間内からランダムにベクトルを得た．一方，本手法では実在するユーザとアイテムを表現した特徴ベクトルをサンプリングする．入力ベクトルのサンプリング手法をアルゴリズム 2 に示す．これにより，推薦モデルの実在しないユーザ・アイテムベクトルの入力による予期しない出力値を学習することを防ぐことができる．そして，サンプリングしたベクトルを推薦モデルに入力し，出力値を得る．

アルゴリズム 2　入力ベクトルのサンプリング

入力: ユーザ集合U，アイテム集合I
出力: サンプリングしたベクトル$\boldsymbol{z_i}$
1. $j \leftarrow I$からランダムにアイテムを抽出
2. $u \leftarrow U$からランダムにユーザを抽出
3. $\boldsymbol{z_i} \leftarrow j, u$を特徴ベクトルに変換
4. **return** $\boldsymbol{z_i}$

b) **最適な数の説明可能な特徴量を用いた局所的な線形回帰モデルの学習**: ステップ 2.a で得られた入力ベクトルと出力値を用いて，局所回帰モデルを訓練し，推薦の解釈モデルを作成する．なお，以下に示す推薦の説明として利用することができない特徴量は訓練に使用しない．

- ユーザ ID を one-hot 表現した特徴量
- アイテム ID を one-hot 表現した特徴量
- ユーザが理解することが困難な特徴量

また，LIME-RS ではすべての説明可能な特徴量を使用して解釈モデルを訓練する一方で，提案手法では全体の説明可能な特徴量からK個の特徴量のみを選択し，訓練に使用する．特徴量の数Kは，チューニングが必要なハイパーパラメータである．特徴量の数をチューニングすることによって，説明がシンプルになるだけでなく，冗長な特徴量の学習を防ぐことによって，元の推薦モデルに対する，解釈モデルの忠実度を高めることができる．K個の特徴量の選択手法は，回帰モデルの損失に基づいて特徴量を 1 つずつ追加していく forward selection や，特徴量を 1 つずつ減らしていく backward elimination など複数の戦略が存在する．

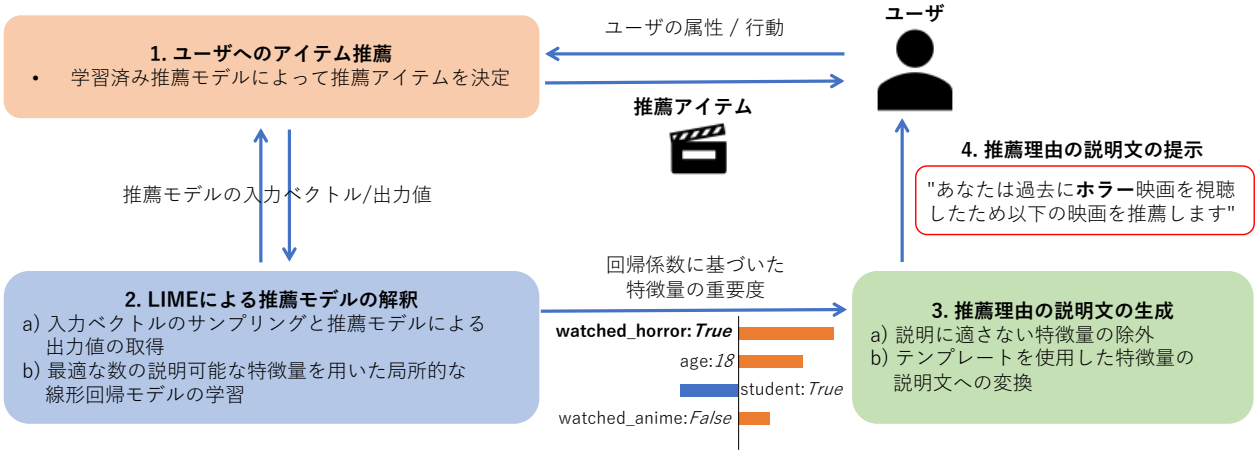

図 4.1 提案手法の模式図

3) 推薦理由の説明文の生成

このステップでは，解釈モデルによって得られた特徴量の重要度を用いて，ユーザに提示するための説明文を生成する．

a) **説明に適さない特徴量の除外**: ステップ 2.b で選択されたK個の説明可能な特徴量には，対象のユーザや推薦アイテムに対する説明として不適当なものが含まれている可能性が存在する．具体的には以下の 2 パターンのいずれかに合致する特徴量は説明に不適当であるため，説明の対象から取り除く．

- ネガティブな特徴量: 例えば，図 4.1 における特徴量「student」重要度は負であり，「あなたは学生のため，この商品を好まないと予測されました.」と説明することができる．しかし，このようなネガティブな説明は，アイテム推薦の説明としては不適当である．
- ユーザ･アイテムに適合しない特徴量: 例えば，図 4.1 における特徴量「watched_anime」の値は「False」となっており，アニメを見ていないユーザに対して「あなたはアニメを見たため，この商品を好むと予測していました.」といった説明がされてしまう．解釈モデルはシンプルな線形回帰モデルを使用しているため，このような明らかに間違った特徴量の重要度が高くなってしまうことが起こり得る．

b) テンプレートを使用した特徴量の説明文への変換: ステップ 3.a によってフィルタリングされた後の残りの特徴量を使用して，各特徴量の種類に対応するテンプレートに基づいて説明文を生成する．テンプレートは各特徴量が推薦に寄与したことを説明するための文章をあらかじめ人手で作成する必要がある．以下にテンプレートのサンプルを示す．各テンプレートにおけるカッコの中は，対応する特徴量に応じた値で埋められる．

- ユーザの属性: 「あなたの年齢( )に基づいて以下の商品をおすすめします.」
- ユーザの行動: 「あなたが購入したアイテム( )に基づいて以下の商品をおすすめします.」
- アイテムの属性: 「カテゴリ( )に基づいて以下の商品をおすすめします.」

4) 推薦理由の説明文の提示

生成された説明文を推薦アイテムとともに対象ユーザに提示する．ステップ 3.b において，説明文はフィルタリング後の特徴量の数だけ作成されるが，実際にユーザに提示する特徴量の数は特徴量の重みに応じて上位N個といったように任意の値を設定することが可能である．

提案手法は LIME をベースとしているため，あらゆる機械学習モデルを採用した推薦アルゴリズムに対しても適用可能であり，推薦アルゴリズムに変更を加えることなく説明を生成することが可能である．

5. 評価実験

評価実験では，提案手法における特徴量選択手法の違いが与える影響を検証するために，解釈モデルの推薦モデルに対する忠実度である model fidelity を測定する．また、提案手法による説明の提示がユーザの推薦システムに対する透明性や満足度に与える影響を検証するため，被験者による説明文の評価実験を行う．さらに，上記の実験結果をもとに model fidelity とユーザの説明文の評価の相関関係を明らかにする．

5.1. データセット

データセットとして，MovieLens 1M Dataset[1]を使用した．データセットには，6,040 人のユーザが 3,883 本の映画に対して，5 段階のレーティングで評価した 1,000,209 件のログが含まれる[9]．さらに，データセットには各映画のジャンル，各ユーザの性別，年齢，職業といった補助情報も含まれる．これらのデータセットを元に，推薦モデルに使用するため，表 5.1 に示す特徴量の作成を行った．なお，データセットはレーティングのタイムスタンプの順番に応じて，学習データ 60%，検証データ 20%，テストデータ 20%の割合で分割した．検証データは，推薦モデルおよび提案手法における解釈モデルのチューニングに使用した．

表 5.1 使用した特徴量

名称	表現方法	次元数
ユーザ ID	推薦対象のユーザに one-hot encoding を行う．	6,040
映画 ID	推薦対象のアイテムが 1，その他のアイテムが 0 となるよう，one-hot encoding を行う．	3,883
年齢	最小値(18)が 0，最大値(56)が 1 となるよう，min-max normalization を行う．	1
性別	女性が 0，男性が 1 を表現する．	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	過去のタイムスタンプのデータにおいて，ユーザがレーティング済みの映画を 1(視聴済み)，そうでない映画を 0(未視聴)として one-hot encoding を行う．	3,883

[1] MovieLens 1M Dataset, GroupLens, https://grouplens.org/datasets/movielens/1m/
Nóbrega ら[5]の研究で使用された MovieLens 10M Dataset は，ユーザの補助情報(性別，年齢，職業)を含まないため，MovieLens 1M Dataset を使用した．

5.2. 推薦アルゴリズム

提案手法があらゆる推薦アルゴリズムに対して適用可能なことを示すために，2 種類のアルゴリズムをベースとした推薦モデルを構築し，評価実験を行う．1 つ目の推薦アルゴリズムは，Rendle[10]によって提案された factorization machine(FM)である．FM の実装には，Python ライブラリの fastFM[11]を使用し，alternating least squares による学習を行う．なお，検証データによるチューニングの結果，相互作用項のランクを 8，L2 正則化項の値を 100 となるようにハイパーパラメータを設定した．

2 つ目の推薦アルゴリズムは neural network(NN)である．本実験で使用するモデルは，それぞれ 1024 ノードと 64 ノードを持つ dropout 機構を含む全結合 NN を使用した．NN の実装には，Python ライブラリの PyTorch[12]を使用した．検証データによるチューニングの結果，バッチサイズを 32，学習率を 0.001，dropout ratio を 0.2 となるようにハイパーパラメータを設定した．

5.3. ベースライン手法

提案手法と比較するために，ベースライン手法として，以下に示す 2 種類のモデル独立型の手法を選択した．

- LIME-RS[5]: LIME を推薦システムに対して適用した最先端の手法．解釈モデルのハイパーパラメータは検証データにおける model fidelity@50(算出方法については 5.5.1 項参照)によってチューニングを行った結果，FM では$S = 10{,}000$，$\sigma = 0.01$，NN では$S = 10{,}000$，$\sigma = 0.03$を採用した(各記号の意味はアルゴリズム 1 を参照)．
- Local AR[7]: 特徴量と推薦アイテムの関係を association rules(AR)によって解釈する手法．各パラメータは[7]に準じて，AR を算出するための近傍ユーザの数を 10, criteria は，min_supp = 0.1, min_conf = 0.1, and min_lift = 0.1 と設定した．

5.4. 評価に使用する特徴量選択手法

解釈モデルの忠実度の評価では，K個の特徴量の選択を以下に示す 3 種類の手法によって行う．

- forward selection: 解釈に使用する回帰モデルの損失を最も小さくする特徴量をKに達するまで 0 から 1 つずつ追加する方法．
- backward elimination: 解釈に使用する回帰モデルの損失を最も小さくする特徴量をKに達するまですべての説明可能な特徴量の数(3,924)から 1 つずつ削減する方法．
- lasso regression: 解釈に使用する回帰モデルに L1 正則化項αを追加することによって，冗長な特徴量の重みを 0 にする方法．この手法では特徴量の数Kの代わりにαを指定する．

なお，ハイパーパラメータは検証データにおける model fidelity@50 によってチューニングを行った結果，FM では$S = 10{,}000$，$\sigma = 0.01$，NN では$S = 10{,}000$，$\sigma = 0.03$を採用した．さらに，特徴量の数Kは forward selection と backward elimination で共通して 20 を，lasso regression におけるαは 0.3 を採用した．

5.5. 解釈モデルの忠実度の評価

5.5.1. 評価指標

解釈モデルの評価では，Peake ら[7]によって提案された推薦モデルに対する忠実度を計測する指標である model fidelity を計測する．model fidelity@N の計算手法を式 5.1 に示す．

$$\text{model fidelity@}N = \frac{|\text{recommended\_items} \cap \text{explainable\_items}|}{|\text{recommended\_items}|} \quad \text{(式 5.1)}$$

ここで，recommended_itemsはオリジナルの推薦モデルによって算出される各ユーザの top-N推薦アイテムの集合である．一方，explainable_itemsは解釈モデルによって算出される各ユーザの top-N推薦アイテムの集合である．すなわち，model fidelity は元の推薦リストに対する解釈モデルの再現率を算出することによって，解釈モデルの推薦モデルに対する忠実度を計測する．model fidelity はデータセット内の各ユーザに対して計測し，最終的な評価の値は全ユーザの平均値とする．

5.5.2. 評価結果

3 つの異なる特徴量の選択手法を採用した提案手法とベースライン手法における model fidelity@10 と model fidelity@50 の評価結果を表 5.2，表 5.3 にそれぞれ示す．提案手法はいずれの特徴量の選択手法を用いた場合も，FM，NN のどちらにおいてもベースライン手法を上回る結果となった．提案手法で特徴量の各選択手法の結果を比較すると，forward selection が最も model fidelity が高い結果となった．

5.6. 被験者による説明文の評価

5.6.1. 被験者実験の概要

本実験では，提案手法によって生成された推薦理由の説明文を，被験者に対して提示し，異なる尺度による説明の質を問うアンケートに回答してもらうことによって説明文ベースの評価を行う．なお，推薦モデルの解釈手法は，ベースラインの中で model fidelity が最も高かった LIME-RS と，提案手法の中で model fidelity が最も高かった forward selection による特徴量選択を使用した手法の 2 つのみを使用し，提案手法の LIME-RS に対する優位性を検証する．

被験者は大学内で募集した 21 名の学生であり，web アプリケーションによって評価の収集を行う．被験者側の実験の操作とシステム側の内部の動作の流れを以

下に示す.

1. (被験者側) 被験者は年齢, 性別を入力する. また, MovieLens データセットに存在する映画のリストから, 過去に視聴したことのある映画を選択する.
2. (システム側) 被験者が入力した年齢, 性別, 過去に視聴したことのある映画をもとに, 表 5.1 に示した特徴量に変換を行う. 次に, 特徴量を元に, 5.5 項で使用したものと同様の, MovieLens データセットによる学習済み推薦モデルを用いて推薦映画を決定する. なお, 推薦アルゴリズム FM と NN のそれぞれで MovieLens データセットに存在するすべての映画に対する予測レーティングを算出し, top-50 推薦映画を決定する.
3. (システム側) 画面に表示する推薦映画と説明文のセットを作成する. 推薦アルゴリズムと説明文の生成手法は, (FM, 提案手法), (NN, 提案手法), (FM, LIME-RS), (FM, LIME-RS)の 4 つの組み合わせのうち, ランダムにいずれかが採用される. 推薦映画の説明文の生成は, 4 節に示した手法と同様に行われ, 提案手法または LIME-RS において最も重要度が高い 1 つの特徴量に対応する説明文を生成する. なお, 各特徴量に対応する説明文のテンプレートは表 5.4 に示すとおりである.
4. (被験者側) ステップ 4 によって得られた推薦映画と説明文を見て, 説明文の評価を行う. 各説明文の評価は, 表 5.5 に示す 6 つの質問によって行われ, 被験者は各質問に対して 5 段階のリッカート尺度(1.全くそう思わない, 2.あまりそう思わない, 3.どちらとも言えない, 4.ややそう思う, 5.非常にそう思う)によって回答を行う. 質問文は Chang ら[13]の研究を参考としており, 各質問は Tintarev ら[14]によって定義された推薦における説明の目的の 7 分類のうち, 推薦の効果, 透明性, 信頼性, 満足度を問うものである. なお, ステップ 4 における推薦アルゴリズムと説明文の生成手法の組み合わせにおいて, どのパターンが適用されたかは被験者に対して開示されない.
5. ステップ 4, 5 を実験開始から 1 時間が経過, もしくは 50 個の推薦映画と説明文のセットの評価が完了するまで繰り返す.

表 5.2 model fidelity@10 の評価結果

	推薦アルゴリズム	
	FM	**NN**
提案手法 (Forward Selection)	**0.8096**	**0.7966**
提案手法 (Backward Elimination)	0.8092	0.7948
提案手法 (Lasso Regression)	0.8072	0.7908
LIME RS	0.7790	0.7682
Local AR	0.7768	0.7560

表 5.3 model fidelity@50 の評価結果

	推薦アルゴリズム	
	FM	**NN**
提案手法 (Forward Selection)	**0.8020**	**0.7702**
提案手法 (Backward Elimination)	**0.8020**	0.7698
提案手法 (Lasso Regression)	0.8008	0.7680
LIME RS	0.7808	0.7516
Local AR	0.7564	0.7200

表 5.4 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

表 5.5 説明文を評価するための質問

質問文	説明の目的の分類[14]
1.この説明によって,この映画に興味を持った	推薦の効果
2.この説明は推薦システムの透明性を向上させる	透明性
3.この説明は信頼できる	信頼性
4.この説明は正しいと思う	信頼性
5.この説明は分かりやすい	満足度
6.この説明は役に立つ	満足度

5.6.2. 評価結果

被験者の各説明文に対する評価の回答結果をもとに, 提案手法とベースライン手法における回答の違いを検証した. 図 5.1 に推薦モデルに FM を使用した場合における手法・質問別の 5 段階評価の割合を示す. また, 図 5.2 に推薦モデルに NN を使用した場合における手法・質問別の 5 段階評価の割合を示す. さらに, 帰無仮説「提案手法に対する評価が LIME-RS に対する評価よりも優位とは言えない」としたときのマン・ホイットニーの U 検定の結果を表 5.6 に示す. 表 5.6 から, p 値を 5%としたとき, 推薦モデルに FM を使用した場合は, 質問 2 から質問 6 に対しては帰無仮説が棄却され, 提案手法は LIME-RS よりも優位な評価が得られたといえる. 同様に, 推薦モデルに NN を使用した場合は, 質問 2 から質問 5 に対しては帰無仮説が棄却され, 提案手法は LIME-RS よりも優位な評価が得られたといえる. すなわち, 提案手法によって生成された説明文は, LIME-RS によって生成された説明文と比較

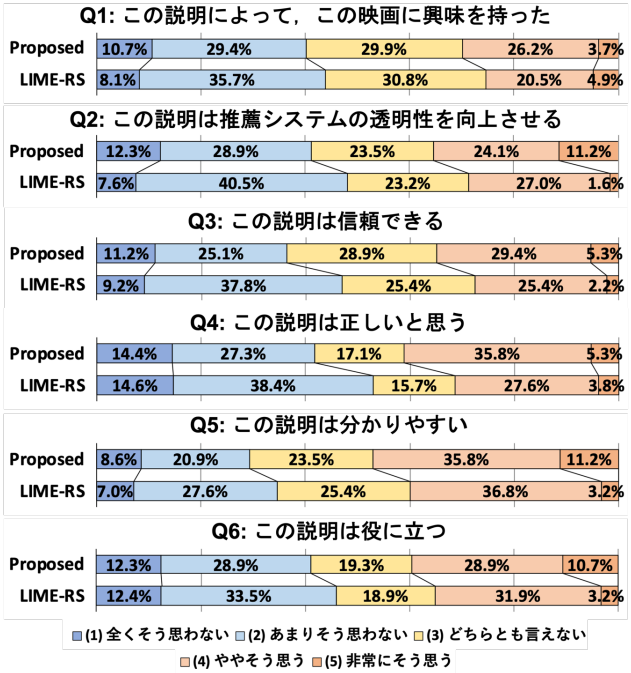

図 5.1 推薦モデルに FM を使用した場合における手法・質問別の 5 段階評価の割合

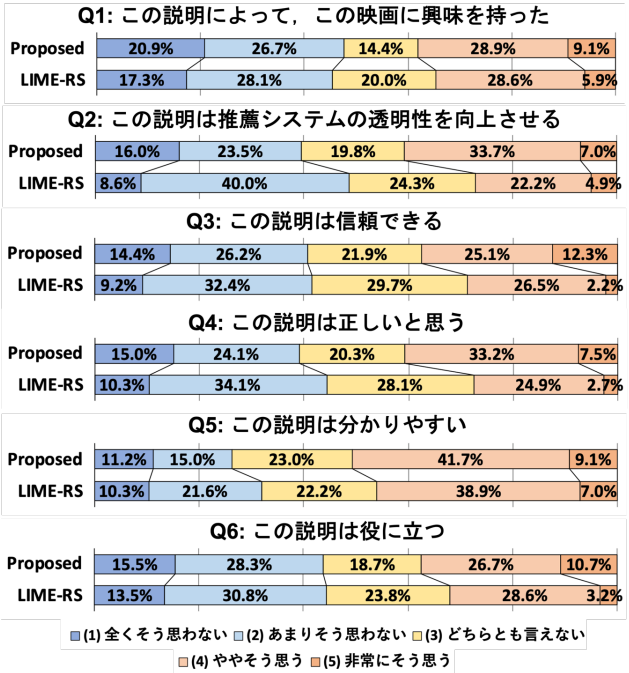

図 5.2 推薦モデルに NN を使用した場合における手法・質問別の 5 段階評価の割合

して，被験者の推薦システムに対する透明性，信頼性，満足度を改善したことが確認できた．一方で，提案手法によって生成された説明文が，推薦の効果を向上させることは確認することができなかった．説明によって推薦の効果を向上させるためには，説明に使用する特徴量の数の増加や説明文の改善といった，さらなる提案手法の改善の検討が必要である．

表 5.6 帰無仮説「提案手法に対する評価が LIME-RS に対する評価よりも優位とは言えない」としたときのマン・ホイットニーの U 検定の p 値
(帰無仮説の棄却範囲 p 値 ＜5%を太字で示す)

	推薦アルゴリズム	
	FM	**NN**
1. この説明によって，この映画に興味を持った	0.2048	0.4928
2. この説明は推薦システムの透明性を向上させる	**0.0217**	**0.0031**
3. この説明は信頼できる	**0.0057**	**0.0307**
4. この説明は正しいと思う	**0.0037**	**0.0024**
5. この説明は分かりやすい	**0.0080**	**0.0326**
6. この説明は役に立つ	**0.0346**	0.0824

5.6.3. model fidelity と被験者評価の関係の考察

本項では，Peake ら[7]によって提案された model fidelity の評価指標としての有用性を検証することを目的として，model fidelity@50 と被験者による説明文の評価値の相関関係を検証した．

n番目に表示された説明文に対する被験者uの質問qの 5 段階評価の値を$\mathrm{eval}_{(u,n,q)}$とする．このときの model fidelity@50 の算出方法を式 5.2 に示す．ここで，model fidelity@50 は $\mathrm{mf}_{(u,n)}$@50 と記す．また，$\mathrm{recommended_items}_{(u,n)}$は被験者$u$に対して$n$番目にアイテムを推薦したときの FM または NN を用いた推薦モデルによって算出される top-50 アイテムリスト，$\mathrm{explainable_items}_{(u,n)}$は被験者$u$に対して$n$番目の説明文を生成したときの提案手法または LIME-RS を用いた解釈モデルによって算出される top-50 アイテムリストである．

$$\mathrm{mf}_{(u,n)}@50 = \frac{\left|\mathrm{recommended\_items}_{(u,n)} \cap \mathrm{explainable\_items}_{(u,n)}\right|}{\left|\mathrm{recommended\_items}_{(u,n)}\right|} \quad (式\ 5.2)$$

このとき，質問qにおける$\mathrm{mf}_{(u,n)}$@50と$\mathrm{eval}_{(u,n,q)}$の間のピアソンの積率相関係数r_qを式 5.3 によって算出する．

$$r_q = \frac{\sum_{u,n}\left(\mathrm{mf}_{(u,n)}@50 - \overline{\mathrm{mf}_{(u,n)}@50}\right)\left(\mathrm{eval}_{(u,n,q)} - \overline{\mathrm{eval}_{(u,n,q)}}\right)}{\sqrt{\left(\mathrm{mf}_{(u,n)}@50 - \overline{\mathrm{mf}_{(u,n)}@50}\right)^2}\sqrt{\left(\mathrm{eval}_{(u,n,q)} - \overline{\mathrm{eval}_{(u,n,q)}}\right)^2}} \quad (式\ 5.3)$$

表 5.7 に，質問文別に算出した相関係数を示す．相関係数は 0.39 から 0.54 の間の値をとり，弱い正の相関関係となった．

次に，被験者ごとに 5 段階評価値と model fidelity@50 の間の相関係数を算出した．被験者別の相関係数の分布を図 5.3 に示す．図 5.3 より，各質問の相関係数は被験者に依存し，被験者によって相関係数の値に大きな違いが見られることが明らかとなった．

model fidelity はユーザの評価を集めずに，推薦の説明文における側面を評価するための指標として提案されたが，ユーザの評価との相関関係にはさらなる改善の余地があるといえる．ユーザの評価と相関の高い新たな指標を開発するためには，本実験において明らかとなった，被験者によって相関係数が大きく異なる要因を探ることが有用であると考えられる．

表 5.7 model fidelity と被験者による 5 段階評価値のピアソンの相関係数

質問文	相関係数
1.この説明によって,この映画に興味を持った	0.3945
2.この説明は推薦システムの透明性を向上させる	0.5050
3.この説明は信頼できる	0.5359
4.この説明は正しいと思う	0.4441
5.この説明は分かりやすい	0.4555
6.この説明は役に立つ	0.5179

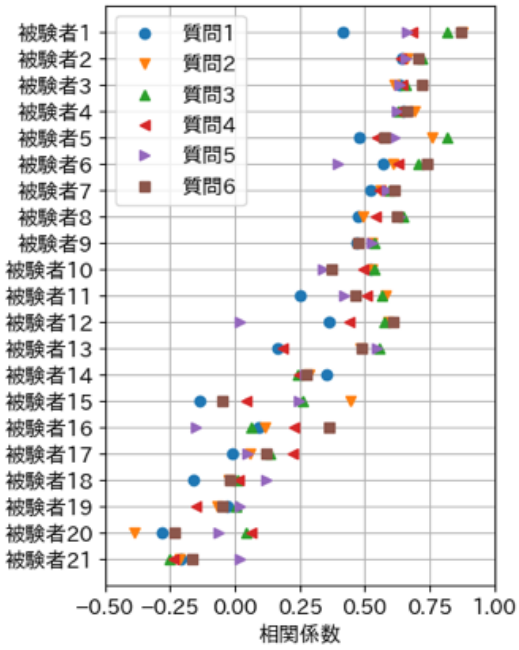

図 5.3 被験者別の相関係数の分布(全質問の相関係数の平均によって被験者のソートを行った)

6. おわりに

本論文では，筆者らが提案した，LIME の拡張によるユーザに対するアイテムの推薦理由の説明を生成する手法の検証を行った．提案手法における，解釈モデルの特徴量の選択手法は，forward selection が最も優れていることを示した．さらに，被験者による評価によって，提案手法によって生成された説明文は透明性，信頼性，満足度の面で LIME-RS に対して有効である結果が得られた．最後に，model fidelity と被験者の評価値の間に，弱い正の相関関係があることを明らかにした．一方で，負の相関を持つ被験者も存在し，model fidelity の評価指標としての改善の余地があることを明らかにした．

本研究における今後の課題は，1)複数の特徴量の関係に着目した説明文の生成方法の検討，2)model fidelity に代わるユーザの評価と高い相関を持つ新たな指標の開発，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，が挙げられる．

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] Y. Zhang and X. Chen, “Explainable Recommendation: A Survey and New Perspectives," Foundations and Trends in Information Retrieval, vol. 14, no. 1, pp. 1–101, 2020.

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

[4] 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.

[5] C. Nóbrega and L. Marinho, “Towards explaining recommendations through local surrogate models,” in Proc. of ACM SIGAPP, pp. 1671–1678, 2019.

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

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

[8] F. Zhu, M. Jiang, Y. Qiu, C. Sun and M. Wang, “RSLIME: An Efficient Feature Importance Analysis Approach for Industrial Recommendation Systems,” in Proc. of IJCNN, pp. 1–6, 2019.

[9] F. M. Harper and J. A. Konstan, “The MovieLens Datasets: History and Context,” ACM Transactions on Interactive Intelligent Systems, vol. 5, no. 4, pp. 1–19, 2016.

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

[11] I. Bayer, “fastFM: A Library for Factorization Machines,” Journal of Machine Learning Research, vol. 17, pp. 1–5, 2016.

[12] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” Advances in Neural Information Processing System, vol. 32, pp. 1-12, 2019.

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

[14] N. Tintarev and J. Masthoff, “Explaining Recommendations: Design and Evaluation,” Recommender Systems Handbook, Springer US, pp. 353–382, 2015.

[15] S. Morisawa and H. Yamana, “Faithful Post-hoc Explanation of Recommendation using Selected Features,” in Proc. of AAAI Spring Symposia, 2021. (採択済)