複数回クラスタリングによる協調フィルタリングを用いた推薦システム

FU LEI† 張 建偉†

† 岩手大学理工学研究科 〒 020–8551 岩手県盛岡市上田 4-3-5
E-mail: †{s3223006,zhang}@iwate-u.ac.jp

あらまし 協調フィルタリングは、過去数年間で過剰な情報の問題に対処するための効果的な手法として台頭している。しかし、シリング攻撃では、攻撃者が偽の評価を導入して推薦システムを影響することができる。本研究では、データセットに対する複数回クラスタリングを行って、攻撃の影響を軽減し、予測精度を維持または向上させる協調フィルタリング手法を提案している。複数回クラスタリングの効果は、映画に関する大規模なデータセットを用いて実験的に探求されている。提案手法は、実際のユーザ評価と予測評価の間の誤差を測定する方法を通じて評価する。また、攻撃前後の予測シフトや推薦リストの差分を比較することで、協調フィルタリングに対するロバスト性も調査する。研究結果は、複数回クラスタリングが協調フィルタリングに対して推薦精度とロバスト性が上がることを示している。

キーワード 推薦モデル、協調フィルタリング、シリング攻撃

1 はじめに

推薦システムは情報を処理する際に製品やサービスを候補から推薦することで人々をサポートする。協調フィルタリング（CF）は、推薦システムで使用される代表的な手法の一つである [1] [2]。協調フィルタリングがユーザ間の類似度を推定し、そのユーザに似たユーザの評価に基づいて評価しないアイテムの評価値を予測する。協調フィルタリングはインジェクション攻撃に対して脆弱性を抱えている [3]。インジェクション攻撃では、攻撃者が偽のユーザプロファイルを推薦システムに注入し、推薦結果に影響を与える。この欠点を取り除くことは、推薦システムの信頼性向上に重要である。協調フィルタリングは、ターゲットユーザに似たユーザを検索し、そのユーザの好みに基づいてアイテムを推薦する。類似したユーザをクラスタリングすることで、予測精度が向上すると期待される。本研究では、複数回クラスタリングを行い、各クラスター内で評価値を予測する手法を提案している。提案手法を評価するために、実際のユーザ評価と予測評価の誤差が測定される。攻撃に対するロバスト性は、予測シフトと TopN シフトの数値を比較している。

2 関連研究

研究に基づいて推薦システムに対する攻撃に関連する研究課題と手法の概要を提供している [4] [5] [6]。協調フィルタリングの攻撃は、特定の攻撃ユーザと推薦システムのロバスト性向上の、2 つの大きなカテゴリに分類できる [7]。攻撃ユーザを検出する手法には、統計的手法 [8]、攻撃の識別 [9]、およびクラスタリング手法の使用が含まれる。クラスタリングに基づく手法の 1 つは、定期的なクラスタリングを実施し [10]、クラスタの代表点に大きな変化がある場合に攻撃を検知する手法である。またクラスタリングに基づく別の手法としては、攻撃ユーザを検知する手法がある [11]。頑健な推薦システムを開発する手法には、攻撃検出 [12]、SVD ベースの協調フィルタリングアルゴリズム [13]、および段階的なヒントトレーニングなどが含まれる [14]。類似な研究では、クラスタリング手法を採用したモデルベースの協調フィルタリング を使用して攻撃を防御することを示している [15] [16] [17]。本研究の特徴は、2 回クラスタリング手法を導入し、その有効性を示すために比較実験を行っている。

3 提案手法

3.1 推薦システム

一般的な通販サイトなどでは多種多様かつ膨大な数のアイテムを取り扱っているため，サイトを利用するユーザが自力で求めるアイテムを探し出すのは非常に困難である．そのためにユーザが求めるアイテムを推薦するシステムはユーザに対してはもちろん，サイトの利便性を高める点において，サイト運営者に対しても有用である．

推薦システムに用いられる代表的な手法は 2 つあり，内容ベースフィルタリングと協調フィルタリングである．内容ベースフィルタリングは，あらかじめサイト内で扱うアイテムについての特徴をデータとして管理し，利用者から得た嗜好データに合うものを推薦する手法である．一方の協調フィルタリングは,「自分と似た嗜好のユーザが好むアイテムは，自分も好むだろう」という仮定を元に，アイテムに対しての評価から自分と似たユーザを探し出し，それらのユーザが好むアイテムを推薦する手法である．

3.2 協調フィルタリング

協調フィルタリング（CF）は、ユーザに推薦すべきアイテムを、そのユーザに似ているユーザがアイテムに行った評価を基に予測するシステムである。そのため予測の際にはあらかじめ似ているユーザ同士を同一クラスタに分類をしておくことで予

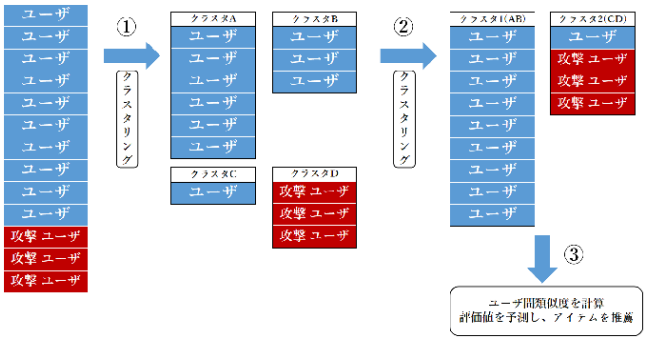

図 1　提案手法の流れ

測精度の向上が期待でき、また、攻撃ユーザが同一のクラスタに分類されれば、そのほかのデータは攻撃による影響を受けずに済む。しかし、攻撃ユーザが多くの一般のユーザに似ている場合、ユーザの分類を行うことでかえって攻撃の影響を増大させることも考えられる。

本研究の目標は、複数回クラスタリングを行い、1 つのクラスタ内のデータ数を増やすことで、推薦精度を維持または向上させ、攻撃の影響を軽減することを目指す。

3.3　提案手法の概要

提案手法は、データセットに対してクラスタリングを行い、ユーザをいくつかのクラスタに分けた後、それぞれのクラスタの中心点を計算し、再度クラスタリングを行うことで、各クラスタを結合する。その後、各クラスタ内のユーザ間の類似度を計算し、その類似度とユーザが付けた評価を用いて評価値を予測する。提案手法の流れを図 1 に示す。

3.4　クラスタリング

クラスタリングでは、データセット内のユーザを類似のクラスタにクラスタリングされる。まず、K-means クラスタリング [18] によって、ユーザが各クラスタに分けられ、次に各クラスタに属するユーザの中心点を使用して再度クラスタリングが実行される。データセット内のクラスタ C に属する各ユーザが $U_i(i = 1, \ldots, m)$ とマークされ、各アイテムが $I_j(j = 1, \ldots, n)$ とマークされ、ユーザ U_i がアイテム I_j に対して与えた各評価が R_{ij} とマークされる場合、クラスタ $C = (c_1, \ldots, c_j, \ldots, c_n)$ 内の各要素 c_j は次のように計算できる。第 1 回クラスタリングではクラスタ数を 2k、2.5k、および 3k に設定し、第 2 回クラスタリングでは k に設定する。

$$c_j = \frac{\sum_{i=1}^{m} R_{ij}}{m} \tag{1}$$

3.5　予測値の計算

協調フィルタリングは、2 つのステップで構成される。Step1 は、ユーザ間の類似度を計算し、似ているユーザを類似グループに形成する。ユーザ間の類似度は、ピアソンの積率相関係数を使用して計算する。ここで、$r(u,i)$ および $r(v,i)$ はユーザ u とユーザ v がアイテム i に付けた評価を表し、$\bar{r_u}$ および $\bar{r_v}$ はそれぞれユーザ u とユーザ v が付けた評価の平均値である。ユーザ u とユーザ v の類似度 $s(u,v)$ は次のようになる。

$$s(u,v) = \frac{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r_u})(r_{v,i} - \bar{r_v})}{\sqrt{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r_u})^2} \sqrt{\sum_{i \in I_u \cap I_v} (r_{v,i} - \bar{r_v})^2}} \tag{2}$$

Step2 は、ユーザ間の類似度と似ているユーザが付けた評価に基づいて、ターゲットユーザの評価値を予測する。ユーザ u が付けた評価の平均値 $\bar{r_u}$ を使用して、ユーザ u のアイテム i への予測評価 $P(u,i)$ が計算される。このとき、最も類似したユーザ $v(v = 1, \ldots, N)$ がアイテム i に付けた評価 r_v, i と、ユーザ u が付けた評価の平均値 $\bar{r_u}$ とユーザ v が付けた評価の平均値 $\bar{r_v}$ の差が、ユーザの類似度 $s(u,v)$ で重みづけされる。計算式は次のようになる。

$$P_{u,i} = \bar{r_u} + \frac{\sum_{v \in N} s(u,v)(r_{v,i} - \bar{r_v})}{\sum_{v \in N} |s(u,v)|} \tag{3}$$

予測値の計算は、クラスタリングが行わない場合はすべてのデータを使用し、クラスタリングが行う場合は同じクラスタ内のデータのみを使用して計算される。

4　評 価 手 法

4.1　評価手法の概要

研究目標は攻撃の影響を軽減しつつ、推薦精度を維持または向上させることである。ユーザが評価したアイテムに対して協調フィルタリングを適用し、ユーザが付けた実際の評価と協調フィルタリングによる予測評価との誤差を測定することで推薦精度を評価する。さらに、研究における攻撃に対するロバスト性は、予測シフトと TopN シフトを構成されている。攻撃に対するロバスト性は、攻撃前と攻撃後の予測値の差を測定することで、予測評価の変化に対応している。同様に、攻撃に対するロバスト性では、攻撃前と攻撃後のユーザの推薦結果の差を測定することで評価される。これは攻撃前と攻撃後の TopN 値の差を計算し、攻撃前と攻撃後の協調フィルタリングが推薦結果の変化を反映している。

4.2　誤 差 測 定

ユーザが付けた評価と予測評価との誤差は、協調フィルタ リングの推薦精度を評価するために使用される。ユーザが評価した各アイテムについて、それを未評価と仮定し、この仮定に基づいてユーザがアイテムの予測評価が計算される。その後、ユーザがつけた評価と予測評価との誤差を測定できる。まず、攻撃前の誤差が図 2 に示されている。攻撃が注入されていない場合で、協調フィルタリングが正確な予測を行うことが証明される必要があるため、攻撃前の誤差測定が重要である。次に、攻撃後の誤差が図 3 に示されている。攻撃前後の予測値は協調フィルタリングのロバスト性を評価する部分で、攻撃前後の予測評価の変化を調査するために使用される。誤差の指標としては MAE（Mean Absolute Error），RMSE（Root Mean Squard Error） を用いる。

$$MAE = \frac{\sum_{k=1}^{N} |P_k - R_k|}{N} \tag{4}$$

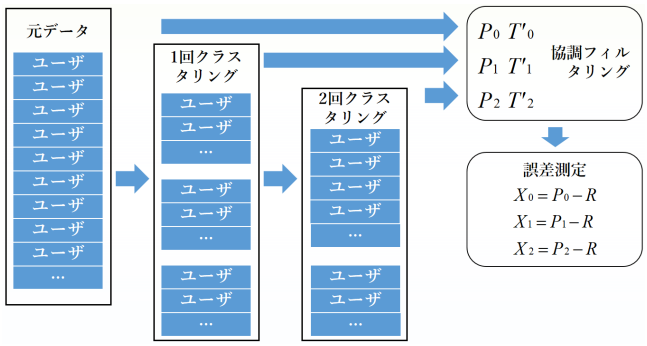

図 2　攻撃前の誤差測定

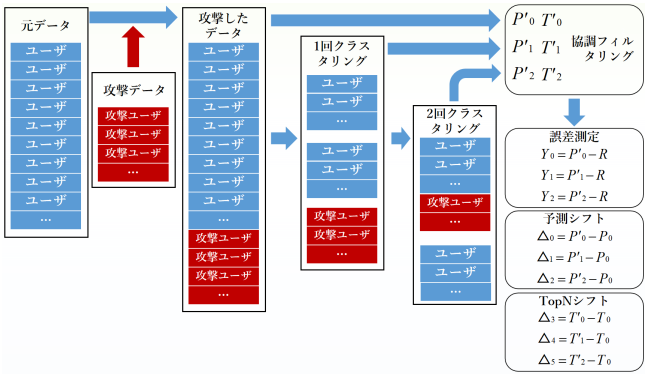

図 3　攻撃後の誤差測定とロバスト性評価

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N}(P_k - R_k)^2}{N}} \tag{5}$$

ここで、P_k は予測評価、R_k は実際の評価、N はユーザが評価したアイテムの数である。

提案方法では、第 1 回クラスタリングのクラスタ数を 2k、2.5k、および 3k に設定した場合、第 2 回クラスタリングのクラスタ数は k から始まりる。第 2 回クラスタリングでは、異なるクラスタ数を比較するために、2k → k 値、2.5k → k 値、および 3k → k 値の誤差が取られる。

4.3　攻撃に対するロバスト性

4.3.1　シリング攻撃

協調フィルタリングに対する攻撃にはいくつかの種類があり、実験ではランダム攻撃と平均攻撃が使用されている。ランダム攻撃では、各フィラーアイテムに対して評価が正規分布に基づいてランダムな評価値を与える。平均攻撃では、各フィラーアイテムに対して評価が各アイテムの評価の平均値を与える。攻撃の意図に応じて、特定のアイテムの評価を増加させることを目的とする攻撃を「プッシュ攻撃」と呼び、特定のアイテムの人気を減少させることを目的とする攻撃を「ニューク攻撃」と呼ぶ。本研究では、評価を増加させることを意図したプッシュ攻撃に焦点を当てている。プッシュ攻撃は、攻撃ユーザを元のデータに注入することで行われ、ターゲットアイテムがランダムに選択され、最高の評価が与えられる。ターゲットアイテム以外のアイテムはランダムに選択され、対応するアイテムの評価が与えられる。実験では、攻撃ユーザ数（攻撃サイズ）とフィラーアイテム数（フィラーサイズ）を異なる値で変化させ、攻撃の影響と協調フィルタリングのロバスト性を評価している。

4.3.2　予測シフト

攻撃に対するロバスト性は、攻撃前と攻撃後の予測値の差を計算することで評価される。図 3 は、ロバスト性評価（予測シフト）の流れを示している。元のデータに攻撃ユーザを生成して注入する。その後、協調フィルタリングが実行され、ユーザがアイテムの評価を予測する。攻撃後、各アイテムの予測評価は攻撃前の誤差測定と同じ方法で測定される。次に、攻撃前と攻撃後の予測値の差を計算する。攻撃前と攻撃後の予測値の差は、攻撃による予測のシフトを表すため、協調フィルタリングのロバスト性を評価するのに使用する。予測値の差が小さいほど、攻撃に対するロバスト性が高いことを示す。

4.3.3　TopN value

実際のアプリケーションでは、ユーザは推薦結果に注目している。この研究では、ユーザの推薦リストを測定するために TopN 値を導入し、予測値に基づいてユーザにアイテムを推薦する。まず、攻撃前の TopN 値を測定し、攻撃されたアイテムがユーザー推薦リストに入った回数を計算する。次に、攻撃後の TopN 値を測定し、攻撃されたアイテムがユーザー推薦リストに入った回数を計算する。攻撃前と攻撃後の TopN 値は、協調フィルタリングのロバスト性を評価する際に、攻撃前と攻撃後の推薦結果の変化を調査するためにも使用される。すべてのユーザは M でマークされる。各ターゲットアイテムを $j(j = 1, \ldots, N)$ とし、ターゲットアイテム j を評価したユーザの数を m_j、ターゲットアイテム j に未評価のユーザの数を l_j とする。すべてのターゲットアイテムがユーザの推薦リストに現れる回数は、以下のように計算できる。

$$TopN = \sum_{j=1}^{N} \frac{l_j}{M - m_j} \tag{6}$$

4.3.4　TopN shift

攻撃に対するロバスト性は、攻撃前と攻撃後の TopN 値の差を計算することで評価される。図 3 は、ロバスト性評価（TopN シフト）の流れを示している。攻撃前と攻撃後の TopN 値の差を測定し、その差分は攻撃による TopN の変化をしている。さらに、その差分は協調フィルタリングのロバスト性を評価するために使用される。TopN 値の差が小さいほど、攻撃に対するロバスト性が高いことを示す。

5　実　　験

実験では、MovieLens Latest Datasets small[1] を使用している。データセットには、610 人のユーザが 9,742 の映画アイテムに対して与えられた 100,836 件の評価が含まれており、評価は 0.5 から 5 の範囲である。クラスタリングには、クラスタリングなし、1 回クラスタリング、および 2 回クラスタリングの 3 つのメソッドが設定されている。1 回クラスタリングでは、クラスタの数 k が設定される。2 回クラスタリングでは、第 1 回

1：https://grouplens.org/datasets/movielens/latest

表 1　クラスタ数

1 回 クラスタリング	2 回クラスタリング			
	第 1 回			第 2 回
k	2k	2.5k	3k	k
2	4	5	6	2
6	12	15	18	6
10	20	25	30	10
15	30	37	45	15

表 2　誤 差 測 定

	0 回	1 回	2 回
攻撃なし	$X_0 = P_0 - R$	$X_1 = P_1 - R$	$X_2 = P_2 - R$
攻撃あり	$Y_0 = P'_0 - R$	$Y_1 = P'_1 - R$	$Y_2 = P'_2 - R$

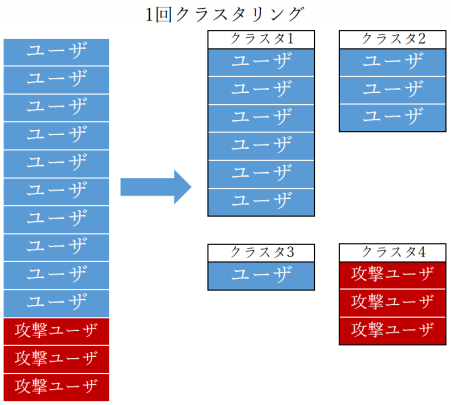

図 4　1 回クラスタリング傾向

クラスタリングには 2k、2.5k、および 3k のクラスタ数が設定され、第 2 回のクラスタリングには k が設定される。クラスタ数の設定は表 1 に示されている。実験結果を公平に比較するために、第 2 回クラスタリングのクラスタ数は第 1 回クラスタリングと同じである。実験ではクラスタ数の設定による 2k → k 値、2.5k → k 値、および 3k → k 値の結果を計算し、それぞれの結果が最小値、中央値、最大値の順に並べて統計する。

攻撃に対するロバスト性を測定する際には、全てのアイテムから 50 個のアイテムがランダムに選択され、これらを攻撃のターゲットアイテムとし、最高の評価を与える。攻撃ユーザは、ターゲットアイテム以外をランダムに選択し、各選択されたアイテムの評価をユーザが与えられた実際の評価に設定することで生成される。

5.1　クラスタリング傾向

5.1.1　1 回クラスタリング傾向

実験では、1 回クラスタリングを行い、クラスタ内データの傾向を収集して分析する。1 回クラスタリング傾向が図 4 に示されている。1 回のクラスタリングした、ユーザと攻撃ユーザは異なるクラスタに分割されて、クラスタ 1、2、3 は攻撃の影響を受けられない。1 回クラスタリングを行い、常に 1 つのクラスタ内のユーザ数量が全ユーザ数量の 60%を占めている。

5.1.2　2 回クラスタリング傾向 1

2 回クラスタリング傾向 1 は、主にクラスタ数が 6、10、15

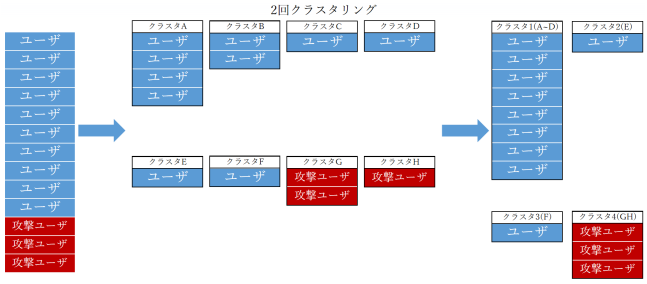

図 5　2 回クラスタリング傾向 1

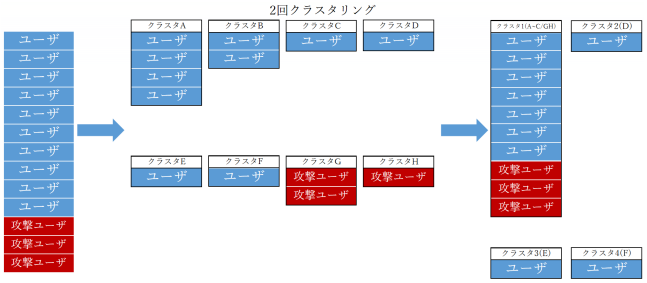

図 6　2 回クラスタリング傾向 2

の場合に発生する。2 回クラスタリング傾向 1 が図 5 に示されている。2 回のクラスタリングした、ユーザと攻撃ユーザは異なるクラスタに分割されて、クラスタ 1、2、3 は攻撃の影響を受けられない。傾向 1 は、常に 1 つのクラスタ内のユーザ数量が全ユーザ数量の 90%以上を占めている。

5.1.3　2 回クラスタリング傾向 2

2 回クラスタリング傾向 2 は、主にクラスタ数が 2 の場合に発生する。2 回クラスタリング傾向 2 が図 6 に示されている。ユーザと攻撃ユーザは同じクラスタに分割されて、クラスタ 1 は攻撃の影響を大きく受けられる。傾向 2 も同様に、常に 1 つのクラスタ内のユーザ数量が全ユーザ数量の 90%以上を占めている。

5.2　誤 差 測 定

攻撃前と攻撃後の誤差をクラスタリング回数（0 回、1 回、2 回）ごとに比較するため、実験の誤差測定は表 2 に示す 6 つのタイプに分類する。

5.2.1　攻撃前の誤差測定

異なる回数のクラスタリングに対して、攻撃前の予測誤差が測定され、それが表 2 の X_0 から X_2 に対応している。この実験では、クラスタリングなし (0 回)、1 回クラスタリング (1 回)、および 2 回クラスタリング (2 回) の各手法に対して協調フィルタリングの予測が行われる。予測された評価と実際の評価の誤差が各手法ごとに測定される。その後、各手法の誤差値が比較され、各行で最小のものを太字で示す。クラスタリングなし、1 回クラスタリング、および 2 回クラスタリングの比較における最大値を下線で示す。攻撃前、MAE に基づく結果が表 3 に示されている。結果から、クラスタリングなしに基づく誤差が最小であり、次に 2 回クラスタリングが続き、1 回クラスタリングが最も大きいことが示されている。さらに、2 回クラスタリングの最大誤差が 1 回クラスタリングの誤差よりも一

表 3　誤差測定 MAE (攻撃なし)

0 回	**1 回**		**2 回**		
MAE	クラスタ数	MAE	MAE		
			最小値	中央値	最大値
0.6068	2	0.6188	0.6176	0.6182	0.6182
0.6068	6	0.6514	0.6162	0.6201	0.6250
0.6068	10	0.6599	0.6224	0.6251	0.6353
0.6068	15	0.6689	0.6267	0.6353	0.6357

表 4　誤差測定 RMSE (攻撃なし)

0 回	**1 回**		**2 回**		
RMSE	クラスタ数	RMSE	RMSE		
			最小値	中央値	最大値
0.7992	2	0.8167	0.8137	0.8144	0.8144
0.7992	6	0.8563	0.8109	0.8145	0.8221
0.7992	10	0.8670	0.8174	0.8202	0.8352
0.7992	15	0.8761	0.8224	0.8350	0.8352

貫して小さいことがわかる。これは、2 回クラスタリングが 1 回クラスタリングよりも予測精度の向上に効果的であることが示されている。また、クラスタの数が増加するにつれて予測精度が低下する傾向がある。攻撃前、RMSE に基づく結果が表 4 に示されている。RMSE 結果が示す傾向は MAE 結果と類似している。

5. 2. 2　攻撃後の誤差測定

攻撃ユーザが攻撃を注入したことにより測定された、予測された評価と実際の評価との誤差を、表 2 の Y_0 から Y_2 にに示す。攻撃サイズとフィラーサイズは元のデータの 5 %と 10 %に設定されている。ランダム攻撃後、MAE 結果と RMSE 結果は表 5、6 に示されている。平均攻撃後、MAE 結果と RMSE 結果は表 7、8 に示されている。結果から、クラスタリングなしに基づく誤差が最小であり、次に 2 回クラスタリングが続き、1 回クラスタリングが最も大きいことが示されている。また、2 回クラスタリングの最大誤差が 1 回クラスタリングの誤差よりも一貫して小さいことがわかる。これは、2 回クラスタリングが 1 回クラスタリングよりも精度の向上に効果的であることが示されている。さらに、クラスタの数が増加すると誤差が増加する傾向がある。原因はクラスタ内ユーザの評価数量と関連している可能性がある。

5. 3　ロバスト性

5. 3. 1　予測シフト

攻撃前と攻撃後の予測値の差を測定し、クラスタリングなし、1 回クラスタリング、および 2 回クラスタリングの各手法について比較する。これは、表 9 の Δ_0 から Δ_2 に対応している。各手法の予測シフトの値を比較し、各行で各行で最小のものを太字で示す。クラスタリングなし、1 回クラスタリング、および 2 回クラスタリングの比較における最大値を下線で示す。ランダム攻撃前後の予測シフトの結果は表 10 に示されている。平均攻撃前後の予測シフトの結果は表 11 に示されている。ほとんどの結果から 2 回クラスタリングに基づく予測シフトが最小

表 5　誤差測定 MAE (ランダム攻撃)

攻撃サイズ	**フィラーサイズ**	**0 回**	**1 回**		**2 回**		
		MAE	クラスタ数	MAE	MAE		
					最小値	中央値	最大値
30	487	**0.6033**	2	0.6141	0.6034	0.6100	0.6109
30	487	**0.6033**	6	0.6345	0.6110	0.6202	0.6207
30	487	**0.6033**	10	0.6657	0.6217	0.6219	0.6234
30	487	**0.6033**	15	0.6603	0.6255	0.6343	0.6434
30	974	**0.6030**	2	0.6128	0.6091	0.6096	0.6136
30	974	**0.6030**	6	0.6451	0.6123	0.6159	0.6176
30	974	**0.6030**	10	0.6439	0.6156	0.6180	0.6184
30	974	**0.6030**	15	0.6594	0.6223	0.6238	0.6299
60	487	**0.5980**	2	0.6080	0.5977	0.6027	0.6083
60	487	**0.5980**	6	0.6415	0.6088	0.6110	0.6171
60	487	**0.5980**	10	0.6512	0.6181	0.6220	0.6243
60	487	**0.5980**	15	0.6613	0.6267	0.6336	0.6376
60	974	**0.5949**	2	0.6156	0.5951	0.5951	0.5953
60	974	**0.5949**	6	0.6322	0.6026	0.6048	0.6061
60	974	**0.5949**	10	0.6338	0.6173	0.6185	0.6220
60	974	**0.5949**	15	0.6506	0.6165	0.6192	0.6264

表 6　誤差測定 RMSE (ランダム攻撃)

攻撃サイズ	**フィラーサイズ**	**0 回**	**1 回**		**2 回**		
		RMSE	クラスタ数	RMSE	RMSE		
					最小値	中央値	最大値
30	487	**0.7951**	2	0.8107	0.7952	0.8027	0.8048
30	487	**0.7951**	6	0.8349	0.8042	0.8166	0.8174
30	487	**0.7951**	10	0.8738	0.8175	0.8180	0.8191
30	487	**0.7951**	15	0.8673	0.8210	0.8322	0.8443
30	974	**0.7944**	2	0.8088	0.8018	0.8034	0.8084
30	974	**0.7944**	6	0.8476	0.8051	0.8098	0.8114
30	974	**0.7944**	10	0.8461	0.8106	0.8120	0.8121
30	974	**0.7944**	15	0.8671	0.8181	0.8208	0.8283
60	487	**0.7875**	2	0.8024	0.7871	0.7923	0.8017
60	487	**0.7875**	6	0.8435	0.7995	0.8053	0.8098
60	487	**0.7875**	10	0.8560	0.8128	0.8170	0.8238
60	487	**0.7875**	15	0.8667	0.8223	0.8319	0.8375
60	974	**0.7834**	2	0.8123	0.7836	0.7836	0.7839
60	974	**0.7834**	6	0.8336	0.7941	0.7942	0.7956
60	974	**0.7834**	10	0.8324	0.8109	0.8124	0.8169
60	974	**0.7834**	15	0.8552	0.8105	0.8151	0.8235

であり、次にクラスタリングなしが続き、1 回クラスタリングが続くことが示されている。さらに、2 回クラスタリングに対する最大の予測シフト値は、他の手法よりも大きい。この結果は、クラスタリング傾向と関連している。

5. 3. 2　TopN シフト

攻撃前と攻撃後の TopN 値の差を測定し、クラスタリングなし、1 回クラスタリング、および 2 回クラスタリングの各手法について比較する。これは、表 9 の Δ_3 から Δ_5 に対応している。この実験では、TopN の N 値は 50 に設定されており、これはターゲットアイテムの数量と同じである。各手法の TopN シフトの値を比較し、各行で最小のものを太字で示す。クラスタリ

表 7　誤差測定 MAE (平均攻撃)

攻撃サイズ	フィラーサイズ	0 回	1 回		2 回		
		MAE	クラスタ数	MAE	MAE		
					最小値	中央値	最大値
30	487	**0.5870**	2	0.6086	0.5874	0.5926	0.5926
30	487	**0.5870**	6	0.6341	0.6020	0.6119	0.6146
30	487	**0.5870**	10	0.6549	0.6068	0.6095	0.6153
30	487	**0.5870**	15	0.6668	0.6185	0.6320	0.6344
30	974	**0.5818**	2	0.6060	0.5874	0.5961	0.5969
30	974	**0.5818**	6	0.6438	0.6027	0.6053	0.6186
30	974	**0.5818**	10	0.6497	0.5994	0.6073	0.6243
30	974	**0.5818**	15	0.6618	0.6187	0.6219	0.6224
60	487	**0.5788**	2	0.6022	0.5844	0.5913	0.5922
60	487	**0.5788**	6	0.6410	0.5922	0.5952	0.5970
60	487	**0.5788**	10	0.6624	0.6002	0.6029	0.6074
60	487	**0.5788**	15	0.6565	0.6113	0.6349	0.6359
60	974	**0.5757**	2	0.6042	0.5758	0.5814	0.5845
60	974	**0.5757**	6	0.6378	0.5919	0.5942	0.6154
60	974	**0.5757**	10	0.6457	0.5971	0.6144	0.6160
60	974	**0.5757**	15	0.6582	0.6217	0.6233	0.6281

表 8　誤差測定 RMSE (平均攻撃)

攻撃サイズ	フィラーサイズ	0 回	1 回		2 回		
		RMSE	クラスタ数	RMSE	RMSE		
					最小値	中央値	最大値
30	487	**0.7740**	2	0.8026	0.7745	0.7796	0.7796
30	487	**0.7740**	6	0.8344	0.7906	0.8057	0.8113
30	487	**0.7740**	10	0.8621	0.7975	0.8000	0.8079
30	487	**0.7740**	15	0.8751	0.8119	0.8302	0.8323
30	974	**0.7665**	2	0.7988	0.7719	0.7838	0.7852
30	974	**0.7665**	6	0.8473	0.7952	0.7959	0.8132
30	974	**0.7665**	10	0.8572	0.7857	0.7997	0.8213
30	974	**0.7665**	15	0.8685	0.8128	0.8168	0.8186
60	487	**0.7628**	2	0.7937	0.7683	0.7789	0.7794
60	487	**0.7628**	6	0.8433	0.7769	0.7837	0.7846
60	487	**0.7628**	10	0.8709	0.7872	0.7927	0.7994
60	487	**0.7628**	15	0.8632	0.8019	0.8333	0.8348
60	974	**0.7585**	2	0.7977	0.7587	0.7640	0.7697
60	974	**0.7585**	6	0.8401	0.7790	0.7818	0.8094
60	974	**0.7585**	10	0.8484	0.7837	0.8078	0.8095
60	974	**0.7585**	15	0.8655	0.8165	0.8182	0.8260

表 9　ロバスト性

	0 回	1 回	2 回
攻撃なし	P_0 , T_0	P_1 , T_1	P_2 , T_2
攻撃あり	P'_0 , T'_0	P'_1 , T'_1	P'_2 , T'_2
予測シフト	$\Delta_0 = P'_0 - P_0$	$\Delta_1 = P'_1 - P_0$	$\Delta_2 = P'_2 - P_0$
TopN シフト	$\Delta_3 = T'_0 - T_0$	$\Delta_4 = T'_1 - T_0$	$\Delta_5 = T'_2 - T_0$

ングなし、1 回クラスタリング、および 2 回クラスタリングの比較における最大値を下線で示す。ランダム攻撃前後の TopN シフトの結果は表 12 に示されている。平均攻撃前後の TopN シフトの結果は表 13 に示されている。結果から、1 回クラスタリングに基づく TopN シフトが最小であり、次に 2 回クラスタリング、クラスタリングなしとなっている（クラスタ数が 2 の場合)。ほとんどの場合、2 回クラスタリングに基づく TopN シフトが最小であり、次に 1 回クラスタリング、クラスタリングなしとなっている（クラスタ数が 6、10、15 の場合)。さらに、2 回クラスタリングに対する最大の TopN シフト値は、1 回クラスタリングの TopN シフトよりも大きい。この結果は、クラスタリング傾向と関連している。

表 10　予測シフト (ランダム攻撃)

攻撃サイズ	フィラーサイズ	0 回	1 回		2 回		
		予測シフト	クラスタ数	予測シフト	予測シフト		
					最小値	中央値	最大値
30	487	0.0013	2	**0.0006**	0.0008	0.0013	0.0018
30	487	0.0013	6	0.0036	**0.0006**	0.0053	0.0073
30	487	0.0013	10	0.0150	**0.0005**	0.0020	0.0036
30	487	0.0013	15	0.0073	**0.0001**	0.0094	0.0105
30	974	**0.0004**	2	0.0010	0.0036	0.0040	0.0040
30	974	**0.0004**	6	0.0132	0.0008	0.0029	0.0075
30	974	0.0004	10	0.0006	**0.0002**	0.0006	0.0029
30	974	**0.0004**	15	0.0072	0.0020	0.0035	0.0070
60	487	0.0035	2	0.0028	**0.0006**	0.0014	0.0040
60	487	0.0035	6	0.0069	**0.0009**	0.0016	0.0041
60	487	0.0035	10	0.0208	**0.0009**	0.0013	0.0045
60	487	**0.0035**	15	0.0192	0.0059	0.0071	0.0133
60	974	0.0028	2	**0.0019**	0.0022	0.0029	0.0033
60	974	0.0028	6	0.0041	**0.0000**	0.0008	0.0072
60	974	0.0028	10	0.0026	**0.0011**	0.0028	0.0049
60	974	0.0028	15	0.0023	**0.0007**	0.0042	0.0050

表 11　予測シフト (平均攻撃)

攻撃サイズ	フィラーサイズ	0 回	1 回		2 回		
		予測シフト	クラスタ数	予測シフト	予測シフト		
					最小値	中央値	最大値
30	487	0.0021	2	0.0046	**0.0018**	0.0049	0.0049
30	487	0.0021	6	0.0053	**0.0020**	0.0036	0.0043
30	487	0.0021	10	0.0093	**0.0005**	0.0022	0.0054
30	487	**0.0021**	15	0.0080	0.0022	0.0024	0.0059
30	974	**0.0050**	2	0.0058	0.0064	0.0074	0.0078
30	974	0.0050	6	0.0135	**0.0003**	0.0066	0.0080
30	974	0.0050	10	0.0026	**0.0003**	0.0024	0.0062
30	974	0.0050	15	0.0021	**0.0000**	0.0051	0.0063
60	487	**0.0027**	2	0.0051	0.0052	0.0065	0.0083
60	487	0.0027	6	0.0054	**0.0009**	0.0009	0.0076
60	487	0.0027	10	0.0070	**0.0011**	0.0023	0.0030
60	487	0.0027	15	0.0025	**0.0010**	0.0066	0.0094
60	974	0.0147	2	**0.0037**	0.0145	0.0166	0.0171
60	974	0.0147	6	0.0089	**0.0032**	0.0128	0.0171
60	974	0.0147	10	0.0158	**0.0034**	0.0144	0.0205
60	974	0.0147	15	0.0066	**0.0014**	0.0032	0.0033

表 12　TopN シフト (ランダム攻撃)

攻撃サイズ	フィラーサイズ	0 回	1 回		2 回		
		TopN	クラス	TopN	TopN シフト		
		シフト	タ数	シフト	最小値	中央値	最大値
30	487	2.4747	2	**0.1836**	0.2051	0.9621	2.4846
30	487	2.4747	6	0.0665	**0.0391**	0.2543	0.3647
30	487	2.4747	10	**0.0132**	0.0232	0.2461	0.2493
30	487	2.4747	15	0.0251	**0.0095**	0.0717	0.1882
30	974	5.7368	2	**0.1044**	2.6409	3.9595	4.4177
30	974	5.7368	6	0.0750	**0.0066**	0.9741	2.7886
30	974	5.7368	10	0.0902	**0.0067**	0.0408	0.4320
30	974	5.7368	15	0.0688	**0.0258**	0.1290	0.1934
60	487	5.3052	2	**0.1225**	1.0826	4.7498	4.9241
60	487	5.3052	6	0.0633	**0.0232**	2.3195	2.3881
60	487	5.3052	10	0.0634	**0.0112**	0.1186	0.1671
60	487	5.3052	15	**0.0015**	0.0670	0.1357	0.1484
60	974	17.430	2	**1.2529**	16.911	16.970	17.041
60	974	17.430	6	**0.0090**	13.717	13.764	14.726
60	974	17.430	10	0.0582	**0.0129**	0.0248	0.0326
60	974	17.430	15	0.0172	**0.0060**	0.0131	3.7134

表 13　TopN シフト (平均攻撃)

攻撃サイズ	フィラーサイズ	0 回	1 回		2 回		
		TopN	クラス	TopN	TopN シフト		
		シフト	タ数	シフト	最小値	中央値	最大値
30	487	0.4975	2	**0.0188**	0.5057	0.5583	0.5583
30	487	0.4975	6	0.1125	**0.0763**	0.1509	0.6687
30	487	0.4975	10	**0.0779**	0.1728	0.3602	0.4405
30	487	0.4975	15	0.0895	**0.0111**	0.0417	0.0758
30	974	0.7537	2	**0.1044**	0.6749	0.7145	0.7718
30	974	0.7537	6	0.1588	**0.0210**	0.2133	0.5767
30	974	0.7537	10	**0.0372**	0.0883	0.3531	1.0087
30	974	0.7537	15	0.0238	**0.0047**	0.0753	0.1026
60	487	0.0540	2	0.0632	**0.0526**	0.0672	0.0837
60	487	0.0540	6	**0.0476**	0.0671	0.1198	0.1936
60	487	0.0540	10	**0.0049**	0.1131	0.1504	0.1919
60	487	0.0540	15	**0.0284**	0.0666	0.0719	0.1752
60	974	0.5041	2	**0.0032**	0.5008	0.5846	0.6650
60	974	0.5041	6	0.0452	**0.0049**	0.5567	0.6492
60	974	0.5041	10	0.0896	**0.0230**	0.1421	0.5999
60	974	0.5041	15	0.0311	**0.0013**	0.0215	0.0448

6　結　　論

本研究では 2 回クラスタリングを行い、クラスタ内でアイテムを推薦することで、頑健な協調フィルタリング手法を提案した。攻撃前後の予測評価の誤差を測定することで、予測評価の手法を紹介した。さらに、攻撃前後の予測値と TopN 値の差を測定することで、ロバスト性評価の手法を導入した。

実験結果によれば、2 回クラスタリングを行う推薦方法は、攻撃の影響を抑制することで一定の効果を表している。2 回クラスタリングの誤差は 1 回クラスタリングよりも小さく、予測精度が向上していることを示している。ロバスト性では、2 回クラスタリングのロバスト性が他のクラスタリング手法よりも優れていることがある。ただし、実験結果では、2 回クラスタリングの安定性は 1 回クラスタリングほど良くないことも示されている。平均攻撃と比較して、協調フィルタリングはランダム攻撃に大きく影響を受けられる。さらに、シリング攻撃が予測シフトに小さい影響を与えるが、実際の推薦結果に大きな影響を与える可能性があることを示す。

今後の課題として、クラスタリング手法については DBSCAN 法を使用し、データセットについては MovieLens 1B Synthetic Dataset を追加し、評価に用いた指標の吟味が挙げられる。

謝　　辞

本研究は JSPS 科研費 JP22K12271 の助成を受けたものである.

文　　献

[1] J. B. Schafer, D. Frankowski, J. L. Herlocker, and S. Sen. Collaborative filtering recommender systems. *Adaptive Web 2007*, pages 291–324, 2007.

[2] Rui Chen, Qingyi Hua, Yanshuo Chang, Bo Wang, Lei Zhang, and Xiangjie Kong. A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks. *IEEE Access*, 6:64301–64320, 2018.

[3] M. P. O'Mahony, N. J. Hurley, and G. C. M. Silvestre. Promoting recommendations: an attack on collaborative filtering. *DEXA 2002*, pages 494–503, 2002.

[4] I. Gunes, C. Kaleli, A. Bilge, and H. Polat. Shilling attacks against recommender systems: a comprehensive survey. *Artif. Intell. Rev.*, 42(4), 2014.

[5] M. Si and Q. Li. Shilling attacks against collaborative recommender systems: a review. *Artif. Intell. Rev.*, pages 1–29, 2018.

[6] Agnideven Palanisamy Sundar, Feng Li, Xukai Zou, Tianchong Gao, and Evan D. Russomanno. Understanding shilling attacks and their detection traits: A comprehensive survey. *IEEE Access*, 8:171703–171715, 2020.

[7] Reda A. Zayed, Lamiaa Fattouh Ibrahim, Hesham A. Hefny, Hesham Abou El Fetouh Salman, and Abdulaziz Almohimeed. Experimental and theoretical study for the popular shilling attacks detection methods in collaborative recommender system. *IEEE Access*, 11:79358–79369, 2023.

[8] Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Felice Antonio Merra. A regression framework to interpret the robustness of recommender systems against shilling attacks. *IIR*, 2021.

[9] Siyu Wang, Yuanjiang Cao, Xiaocong Chen, Lina Yao, Xianzhi Wang, and Quan Z. Sheng. Adversarial robustness of deep reinforcement learning based dynamic recommender systems. *CoRR*, abs/2112-00973, 2021.

[10] M. P. O'Mahony, N. J. Hurley, and G. C. M. Silvestre. Collaborative filtering-safe and sound? *ISMIS 2003*, pages 506–510, 2003.

[11] Carlos Pedroso and Aldri Santos. Dissemination control in dynamic data clustering for dense iiot against false data injection attack. *Int. J. Netw. Manag*, 32(5), 2022.

[12] Hongtao Yu, Lijun Sun, and Fuzhi Zhang. A robust bayesian probabilistic matrix factorization model for collaborative filtering recommender systems based on user anomaly rating behavior detection. *KSII Trans. Internet Inf. Syst*, 13(9):4684–4705, 2019.

[13] F. Zhang and S. Xu. Analysis of trust-based e-commerce recommender systems under recommendation attacks. *ISDPE 2007*, pages 385–390, 2007.
[14] Y. Du, M. Fang, C. Xu, J. Cheng, and D. Tao. Enhancing the robustness of neural collaborative filtering systems under malicious attacks. *IEEE Trans. Multimedia*, 21(3):555–565, 2019.
[15] B. Mobasher, R. D. Burke, and J. J. Sandvig. Model-based collaborative filtering as a defense against profile injection attacks. *AAAI 2006*, pages 1388–1393, 2006.
[16] Lam SK and Riedl JT. Shilling recommender systems for fun and profit. *WWW 2004*, pages 393–402, 2004.
[17] Zhang S, Ouyang Y, Ford J, and Makedon F. Analysis of a low-dimensional linear model under recommendation attacks. *SIGIR 2006*, pages 517–524, 2006.
[18] Renato Cordeiro De Amorim. A survey on feature weighting based k-means algorithms. *CoRR*, abs/1601-03483, 2016.