Evaluating the Robustness of Collaborative Filtering Recommender Systems against Attacks

Andrew Zhou
Department of Data Science
University of Washington
Seattle, WA 98195
ajz55@uw.edu

Kenny Zhang
Department of Statistics
University of Washington
Seattle, WA 98195
zhehaoz@uw.edu

Shruthi Kundapura
Department of Computer Science
University of Washington
Seattle, WA 98195
skunda@uw.edu

1 Introduction

E-commerce has existed as early as the 1970's, but it is only in the recent 20-30 years when E-commerce has taken off, and one of the biggest contributions to the success of E-commerce is the use of Recommender Systems. One of the most widely used paradigms in recommender systems used today is Collaborative Filtering. However, collaborative filtering has some major downfalls such as cold start, scalability, vulnerability to attacks, etc. This paper will mainly focus on the vulnerability of attacks on collaborative filtering.

2 Related Works

There are various studies related to the vulnerability of attack on recommender systems, either from an attacker perspective or a defender perspective. In this section, we provide details on different studies done to create an effective attack on recommender system (RS). The goal of these attacks are to either reduce accuracy of the RS or to manipulate the RS such that the attacker chosen target item gets recommended. With respect to this, we discuss the study done so far to identify different attack models, identify metrics to determine the effectiveness of the attack, impact of attack on different kinds of RS and studies to increase robustness of RS.

Attacks on RS are mainly categorized as data poisoning attacks and profile pollution attacks. Data poisoning attacks uses fake user profiles with carefully crafted ratings injected to RS with a goal to promote target item (Deng et al. 2016)[1]. Profile pollution injects information into users' profiles to perturb results obtained from services using personalization algorithms Xing et al. (2013)[12]. We will primarily focus on the data poisoning attacks (also called shilling attacks). Shilling attack models are mainly of 5 types as per prior work [9]: Random attack, Average attack, Bandwagon attack, Segmented attack, Sampling attack. This categorization of attack models is done on the basis of way in which filler items are chosen and what ratings are updated for these items on a fake profile. Every attack aims to create fake profiles containing items and ratings. These items will include target item and some filler items with appropriate rating, such that target item gets recommended. When these fake profiles are fed into RS systems, its behavior can be altered.

Lam et al. (2004)[6] in their paper describe an approach to create Random and Average attack model on kNN user-user and user-item collaborative filtering RS model. This paper also gives an account of how effective the attacks were on collaborative filtering neighborhood-based RS by using metrics like Prediction shift, Expected TopN Occupancy and Mean Absolute Error(MAE). With experiments performed on MovieLens dataset, the paper concluded that user-user kNN is more susceptible to attacks than item-item kNN. Similarly, Sandvig et al.(2007)[11] concludes that neighbourhood based RS is more vulnerable to attacks. This paper also describes RS based on association rule and Probabilistic Latent semantic Analysis(PLSA), and shows that they are robust to attacks by evaluating them against Hit Ratio metric. Though association rule based RS is more robust, it doesn't have good coverage on items to be recommended as it chooses only items which meet support threshold.

With deep learning based RS gaining popularity, some studies have been done to build a custom attack model different from traditional model. (Huang et al. 2021)[4] gives an account of one such attack model built relying on the complete knowledge of what RS algorithm is used. This paper provides a systematic study of data poisoning attacks on deep learning based RS where the attack is formulated as a non-convex optimization problem under the restriction of a small amount of fake user injections. The attack is performed on both MovieLens datasets and we see a 5% fake user profiles can make unpopular target items more popular or achieve a high hit ratio. The paper also touches upon partial knowledge settings and the attack performance under a basic detection algorithm. This paper could serve as a baseline for studying shilling attacks on Neural Matrix Factorization model.

Lin et al. (2020)[10] provides another way for shilling attack on deep learning based RS. The paper provided an Augmented Shilling Attack framework (AUSH) based on the idea of Generative Adversarial Network and claimed this attack is very hard to detect. This is a more complicated example that uses a neural network to attack the deep learning based RS. The model architecture is hard to implement and computationally expensive but we can borrow some idea from this paper.

Li et al. (2016)[8] focuses on data poisoning attacks on factorization-based collaborative filtering. The paper explores two popular factorization-based collaborative filtering algorithms: the alternative minimization formulation and the nuclear norm minimization method. Full knowledge of the learner is assumed and two goals of the attacker is formulated: (i) to maximize the error of the collaborative filtering system, and eventually render the system useless. (ii) to boost (or reduce) the popularity of (a subset) of items. Both of the attacks are formulated as optimization problems and MovieLens is used as a real world dataset for experiment.

However, all these works don't give a clear comparison of how these traditional (Random, Average, Bandwagon attack), easy to launch attacks models have an impact on traditional neighborhood-based RS model versus matrix-factorization-based RS model versus deep-learning based RS model. So, in this work we want to create some of the traditional shilling attack and evaluate its impact on different Collaborative Filtering (CF) based RS. Lam et al.(2004)[6] provides a summary of attack on traditional kNearest Neighbour(kNN) based CF algorithms. We would want to extend this and check the impact of traditional shilling attack model (Random, Average, Bandwagon) on the following 4 categories of collaborative filtering algorithms: User-User kNN based CF, Item-Item kNN based CF, Matrix-fatorization based CF (Koren et al. 2009)[5] and Neural Network Matrix Factorization(NMF) (Dziugaite et al. 2015)[2]. With this study on these different CF algorithms, we mainly want to answer the following questions:

- Which of the systems are more prone to attack?
- Has these systems over the time become more vulnerable or robust?

3 Proposal

3.1 Introduction

Recommender Systems which have become essential part of many online applications are known to be vulnerable to attacks. Among the different attacks studied, we are focusing on understanding the impact of shilling attacks on different domains. Shilling attacks are initiated by injecting carefully crafted fake user profiles with chosen items and ratings. Choosing the right item and an appropriate ratings for it is the key to creating a successful attack. We have chosen a couple of different implementations of the collaborative filtering type recommender systems and plan to attack the recommender systems with different types or combinations of shilling attacks. For our project, we propose to analyze the effectiveness of certain types of attacks on recommender systems, find the robustness of certain types of recommender systems, and discover any patterns among the different types of attacks.

3.2 Objectives

Some of the goals/objectives we want to accomplish in this project:

- Find the effectiveness of a chosen type of attack on a recommender system.

- Find out how the number of injections (fake user profiles) impact different recommender systems.
- Find the overall robustness of a chosen recommender system.
- Discover any patterns or findings in the different types of attacks.

3.3 Data set

To accomplish our project, we will be using the MovieLens dataset from the GroupLens Research Project at the University of Minnesota. The dataset contains data on different movies, movie information, user ratings, and user information.

This data set is available at https://grouplens.org/datasets/movielens/, and is available for research uses only. (Harper et al. 2015)[3]

3.4 Algorithms and Techniques

We will be focusing on four different types of collaborative filtering recommender systems to experiment on.

- Traditional User-User kNN based collaborative filtering
- Item-Item kNN based collaborative filtering
- Matrix Factorization based collaborative filtering (Koren et al. 2009)[5]
- Neural Network Matrix Factorization (NMF) (Dziugaite et al. 2015)[2]

In terms of attacking the data set, we will be mainly focusing on shilling attacks also known as data poisoning attacks. We will assume to have complete knowledge of the recommender system algorithm used and some user-item data that is stored to make more powerful attacks. There are mainly five categories of shilling attacks, and we will be trying some of these attacks((Li et al. 2017)[9]) on the different recommender systems from above.

- Random Attack - Generates random values within a global distribution
- Average Attack - Generates random values within an individual's distribution
- Bandwagon Attack - Frequently rate the targeted items
- Segmented Attack - Frequently rate items that are similar to the targeted items
- Sampling Attack - Sample attacks from the whole model to simulate real users

Our process would involve building these four different collaborative filtering recommender systems on the MovieLens dataset and obtain a baseline metric for evaluation. Our task is to generate fake users and reviews based on one or more of the types of shilling attacks mentioned above and inject the data into the recommender system. Then we obtain a second set of metrics for evaluation and evaluate the effectiveness of shilling attacks and robustness of recommender systems.

3.5 Evaluation

Our project will compose of mainly two portions that need to be evaluated:

- the effectiveness of the shilling attacks
- the robustness of the recommender system

The first metric we would like to use is the average **prediction shift** to estimate the effectiveness of the shilling attack (Lee et al. 2012)[7]. This is used to estimate how much the data has been shifted through the injection of fake user profiles and reviews. The second metric is the **mean absolute error (MAE)**, which can be used to evaluate both the effectiveness of the shilling attack and the robustness of the recommender system as we can compare this value with the baseline and after the attack. The third metric we are considering is the **root mean squared error (RMSE)**, which is a very similar metric as MAE.

We would also like to keep track of the number of injected data profiles and reviews into the recommender system. We would like to determine how many rows of data it takes to significantly distort a recommender system's ability to recommend the correct products.

3.6 Expectations

We expect by the end of this project to have a full analysis on the effectiveness of different types of shilling attacks and the robustness of different collaborative filtering recommender systems. We should be able to find out how much injected data is required to generate "bad" recommendations. We also hope to find similar patterns among the different types of shilling attacks that can be helpful to counter against shilling attacks as a whole.

References

[1] Zi-Jun Deng, Fei Zhang, and Sandra PS Wang. "Shilling attack detection in collaborative filtering recommender system by PCA detection and perturbation". In: *2016 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*. IEEE. 2016, pp. 213–218.

[2] Gintare Karolina Dziugaite and Daniel M. Roy. *Neural Network Matrix Factorization*. 2015. arXiv: 1511.06443 [cs.LG].

[3] F. Maxwell Harper and Joseph A. Konstan. "The MovieLens Datasets". In: 4 (2015). URL: http://dx.doi.org/10.1145/2827872.

[4] Hai Huang et al. "Data Poisoning Attacks to Deep Learning Based Recommender Systems". In: *arXiv preprint arXiv:2101.02644* (2021).

[5] Yehuda Koren, Robert Bell, and Chris Volinsky. "Matrix Factorization Techniques for Recommender Systems". In: 42.8 (2009), pp. 30–37. ISSN: 0018-9162. URL: https://doi.org/10.1109/MC.2009.263.

[6] Shyong K. Lam and John Riedl. "Shilling Recommender Systems for Fun and Profit". In: *Proceedings of the 13th International Conference on World Wide Web*. WWW '04. New York, NY, USA: Association for Computing Machinery, 2004, pp. 393–402. ISBN: 158113844X. DOI: 10.1145/988672.988726.

[7] Jong-Seok Lee and Dan Zhu. "Shilling attack detection–a new approach for a trustworthy recommender system". In: *INFORMS Journal on Computing* (2012).

[8] Bo Li et al. "Data poisoning attacks on factorization-based collaborative filtering". In: *arXiv preprint arXiv:1608.08182* (2016).

[9] Xinxin Niu Li Yang Wei Huang. "Defending shilling attacks in recommender systems using soft co-clustering". In: *IET Information Security* (2017).

[10] Chen Lin et al. "Attacking Recommender Systems with Augmented User Profiles". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 855–864.

[11] J. J. Sandvig, Bamshad Mobasher, and Robin Burke. "Robustness of Collaborative Recommendation Based on Association Rule Mining". In: *Proceedings of the 2007 ACM Conference on Recommender Systems*. RecSys '07. New York, NY, USA: Association for Computing Machinery, 2007, pp. 105–112. ISBN: 9781595937308. DOI: 10.1145/1297231.1297249. URL: https://doi.org/10.1145/1297231.1297249.

[12] Xingyu Xing et al. "Take this personally: Pollution attacks on personalized services". In: *22nd {USENIX} Security Symposium ({USENIX} Security 13)*. 2013, pp. 671–686.