Robustness Analysis of Naïve Bayesian Classifier-Based Collaborative Filtering

Cihan Kaleli and Huseyin Polat

Computer Engineering Department, Anadolu University,
Eskisehir, 26470, Turkey
ckaleli,polath@anadolu.edu.tr

Abstract. In this study, binary forms of previously defined basic shilling attack models are proposed and the robustness of naïve Bayesian classifier-based collaborative filtering algorithm is examined. Real data-based experiments are conducted and each attack type's performance is explicated. Since existing measures, which are used to assess the success of shilling attacks, do not work on binary data, a new evaluation metric is proposed. Empirical outcomes show that it is possible to manipulate binary rating-based recommender systems' predictions by inserting malicious user profiles. Hence, it is shown that naïve Bayesian classifier-based collaborative filtering scheme is not robust against shilling attacks.

Keywords: Shilling, Naïve Bayesian classifier, Robustness, Prediction.

1 Introduction

In e-commerce applications, one of the most popular method for producing predictions is collaborative filtering (CF). By employing CF services, online vendors provide personalized referrals to their customers to boost their sales. Online vendors need to collect users' preferences about several products that they previously purchased or showed interest. Such preferences can be expressed in binary form in which ratings must strictly belong to one of two classes, like or dislike. Naïve Bayesian classifier (NBC)-based CF is widely used algorithm to produce binary recommendations, which is proposed by [1]. NBC-based CF considers all users' data for estimating a prediction for a target item (q) for an active user (a).

Malicious users can insert bogus profiles, referred to as shilling attacks, in a very straightforward way into recommender systems' databases to manipulate the estimated predictions on behalf of their advantages. The advantage helping people be part of recommender systems easily then becomes a vulnerability for the systems. Consequently, CF algorithms can be faced with various profile injection attacks [2,3]. In a traditional example of attacking scenario, any product producer may want to increase its product's popularity. To do so, it tries to insert fake user profiles into the system in which the target product is extremely liked. In another scenario, the same producer might intend to decrease the popularity of one of its competitor's product by creating and inserting bogus profiles.

C. Huemer and P. Lops (Eds.): EC-Web 2013, LNBIP 152, pp. 202–209, 2013.

CF algorithms suffer from shilling attacks. Thus, researchers introduce several studies examining the robustness of CF algorithms against them [4,5]. However, previous works examine numerical ratings-based CF algorithms and there is no work covering the case when the ratings are in binary form. Hence, we primarily focus on how the common basic attack models can be applied to NBC-based CF. All users having rating for the target item participate in recommendation process in NBC-based scheme. Thus, vulnerability of NBC-based CF algorithm might increase against profile injection attacks. We particularly introduce binary forms of six mostly implemented attack types, i.e., segmented attack intends to push a product, reverse bandwagon and love/hate attacks are employed as nuke attacks, while random, average, and bandwagon attacks can be considered for achieving both goals. We investigate how robust NBC-based CF algorithm under such attacks. For the purpose of measuring success of binary attacks, we propose a new metric. We perform real data-based experiments and their results clearly show that the proposed binary forms of shilling attacks are capable of biasing prediction results of NBC-based CF algorithm in the direction of their aims.

2 Related Work

Dellacoras [6] discusses negative effects of fraudulent behaviors of users on online reputation systems inspiring shilling attacks concept. O'Mahony et al. [2,3] introduce the first works about shilling attacks, where the authors analyze vulnerabilities of CF systems against biasing prediction results. Initially, shilling attack strategies are discussed by O'Mahony [7]. The proposed attacks are performed by inserting fake user data to the CF systems. Later, Lam and Riedl [8,9] introduce four open questions related to effectiveness of shilling attacks. Mobasher et al. [10,11] determine attack strategies and present the basic attack types such as random, average, bandwagon, and love/hate attacks. Burke et al. [4] examine bandwagon and popular item attacks. Burke et al. [5] propose a different attack type called segmented attack targeting a set of particular users. Cheng and Hurley [12] propose diverse and obfuscated attack models to be effective on model-based CF schemes. To bias users' reputation, copied-item injection attack is presented by Oostendorp and Sami [13]. Gunes et al. [14] present a comprehensive survey about shilling attack studies explaining attack descriptions, detection methods, designing robust recommender algorithms, and evaluation metrics and data sets.

The studies presented above study various attack models and investigate the robustness of numerical ratings-based CF schemes against such attacks. However, CF systems might employ binary ratings rather than numerical ratings; and there is no work analyzing robustness of CF systems with binary preferences. Therefore, in this study, we distinctively focus on robustness analysis of NBC-based CF algorithm, which is proposed to estimate binary ratings-based recommendations. We also propose a new metric to measure the effects of shilling attacks on binary systems.

Miyahara and Pazzani [1] utilize NBC to provide binary predictions. The "naïve" assumption states that features are independent given the class label.

Table 1. Generic Attack Profiles

Attack Type	I_F		I_S		$I_\emptyset$	i_t
	Items	Ratings	Items	Ratings		
Random	Random	System mean	N/A		$I - I_F$	r_{max}
Average	Random	Item mean	N/A		$I - I_F$	r_{max}
Bandwagon	Random	System mean	Popular	r_{max}	$I - \{I_F U I_S\}$	r_{max}
Segmented	Random	r_{min}	Segmented	r_{max}	$I - \{I_F U I_S\}$	r_{max}
Reverse Bandwagon	Random	System mean	Unpopular	r_{min}	$I - \{I_F U I_S\}$	r_{min}
Love/Hate	Random	r_{max}	N/A		$I - \{I_F U I_S\}$	r_{min}

Hence, the probability of an item belonging to $class_j$, where $j \in \{like, dislike\}$, given its n feature values, can be written, as follows:

$$p(class_j|f_1, f_2, \ldots f_n) \propto p(class_j) \prod_u^n p(f_u|class_j), \tag{1}$$

where f_u corresponds the feature value of q for user u. The probability of each class is computed and q is assigned to the class with the highest probability.

Generic attack profile is depicted in Fig. 1 [11]. It consists of filler items (I_F), selected items (I_S), and the rest of the items, which are left unrated but attacked item (i_t). Items in I_F are selected randomly with a function θ. I_S is determined according to items' popularity by using the rating function δ and i_t is chosen with a function Υ. We analyze random, average, bandwagon, segmented, reverse bandwagon, and love/hate [14]. Their profile details are given in Table 1.

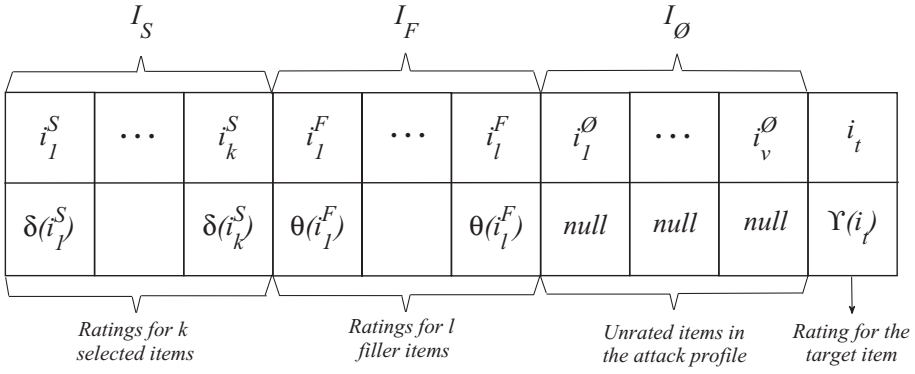

Fig. 1. General form of an attack profile

3 Designing Binary Attacks Against NBC-Based CF

We propose new attack design strategies for attacking databases including binary ratings. We first discuss binary forms of attacks without I_S. Random attack can

be considered as the base model for other attack models [4]. If an attacker knows the system rating mean, she can perform random attack. However, an attacker cannot choose a rating value around the system mean for binary votes. Moreover, if the attacker chooses one as the rating values, it must fill all I_F values with the same value, which ends up with a detectable attack profile. The same scenario is valid for item means. Thus, in binary ratings-based systems, ratings of I_F cannot be selected around item means in average attack, as well. However, since all items do not have the same mode value, it is possible to employ item modes directly in binary attack profiles. Unlike random and average attack strategies, since I_F is filled with the maximum rating in numerical ratings-based love/hate attack, its methodology can be directly applied in the case of binary ratings. Thus, binary forms of random, average, and love/hate attacks have common steps but filling I_F items. Methodology of generating such attacks using binary ratings can be characterized, as follows:

1. Number of filler items (I_F) is determined with respect to a predefined range.
2. Filler items in I_F are uniformly randomly chosen among all items except i_t.
3. All other items (I_S and $I_\emptyset$) are left unrated.
4. For each attack type, do the followings:
   - If the attack type is random, in order to fill each item i in I_F, the attacker generates a random number α_i between 0 and 1. If α_i is greater than 0.5, then item i is filled with 1, otherwise it is filled with 0.
   - If the attack type is average, items in I_F is filled with their mode values.
   - If the attack type is love/hate, items in I_F is filled with maximum rating value, which is 1.
5. The attacker can employ random and average attacks for both pushing and nuking i_t. If the goal is pushing, i_t is filled with 1, otherwise it is filled with 0. Since love/hate is a nuke attack, i_t is assigned to 0 only.
6. Finally, the attacker determines number of all bogus user profiles and generates them by following the above steps.

We then discuss binary forms of attacks with I_S. To perform effective attacks, bandwagon, segmented, and reverse bandwagon attack models utilize I_S item set to increase correlations between fake and genuine users. Bandwagon and reverse bandwagon attacks utilize popular and unpopular items in CF systems, respectively. In segmented attack, the attacker selects a subset of users having an interest to a certain kind of products as target. Such segment of users is constituted by users who have maximum rating value for selected items. In the binary form of these attacks, items in I_S can be filled with either maximum or minimum rating value. On the other hand, strategy of filling items in I_F is changed and the methodology in binary attacks without I_S is employed. The overall methodology of creating binary forms of bandwagon, segmented, and reverse bandwagon attacks can be described, as follows:

1. For each attack type, do the followings:
   - In case of binary bandwagon attack, k of the most popular items are selected as I_S and they are filled with rating value 1.

   - For binary segmented attack, m of the most popular items in the selected segment of users are chosen as I_S and they are filled with rating value 1.
   - If the attack is binary reverse bandwagon attack, k of the most unpopular items are selected as I_S and they are filled with rating value 0.
2. Number of filler items (I_F) is determined with respect to a predefined range.
3. Filler items are uniformly randomly selected among all items except $\{I_S \cup i_t\}$.
4. All other items ($I_\emptyset$) are left unrated.
5. For all attack types, in order to fill each item i in I_F, the attacker generates a random number α_i between 0 and 1. If α_i is greater than 0.5, then item i is filled with 1, otherwise it is filled with 0.
6. Since binary bandwagon attack can be performed for pushing and nuking purposes, i_t gets either 1 or 0 value according to aim of the attack. Unlike binary bandwagon attack, binary segmented and reverse bandwagon attacks are applied only for one purpose. Thus, i_t is filled with 1 in binary segmented attack, while it is assigned to 0 in binary reverse bandwagon attack.
7. Lastly, the attacker determines number of all bogus user profiles and generates them by following the above steps.

Prediction shift is defined as the average changes in the predicted rating before and after the attack for an attacked item. It is utilized to measure success of an attack [14]. However, it works on only numerical ratings. Thus, we propose a new metric, called *ratio shift*, which measures the ratio of 1's in prediction results before and after attack. The metric can be formulated, as follows:

$$Ratio\ Shift = Ratio\ of\ 1s\ after\ attack - Ratio\ of\ 1s\ before\ attack, \quad (2)$$

where $ratio\ of\ 1s\ after\ attack$ represents the percentage of 1's in the produced predictions for an attacked item after an attack while $ratio\ of\ 1s\ before\ attack$ indicates the same percentage in predictions of the corresponding item before it is attacked. We computed only 1's ratio. If the target item is aimed to be pushed, *ratio shift* for that item is a positive value for a successful attack. Conversely, if the attacker's goal is to nuke an item, then *ratio shift* becomes a negative value.

4 Experimental Evaluation

In our trials, we used MovieLens Public (MLP) data set collected by GroupLens research team at the University of Minnesota (http://www.grouplens.org). It includes 100,000 discrete ratings of 943 users about 1,682 movies. We performed experiments for varying *attack size* and *filler size* values. *Attack size* is the percentage of shilling attack profiles in the system while *filler size* represents the percentage of items to be filled with fake ratings in bogus profiles to form I_F.

We labeled items as 1 if the numerical rating for the item was bigger than 3, or 0 otherwise in MLP. We followed all-but-one experimentation methodology in which one of the users acts as an active user and the rest of them forms training set during all iterations. To select two distinct target items, we analyzed 1's and 0's ratio. We finally randomly selected 50 movies for push and nuke attacks.

We chose target items for push attack from the items having zero ratings more than ones, conversely, target items for nuke attack were chosen from items having mostly one ratings. To constitute I_S sets in binary forms of bandwagon and reverse bandwagon attacks, we determined 10 most popular and unpopular items to set k. We targeted a segment of users and selected five of the popular items for that segment for binary segmented attack to set m.

For all items in both target item sets, we produced predictions for all users who do not have a rating for those items. We computed 1's ratio values for each of the target items. All target items were attacked individually for all users in the system. We estimated *ratio shift* values for each item and and overall averages for all target items were presented for each binary attack type. We first varied *attack size* from 1% to 15% and kept *filler size* fixed at 15%. Secondly, *attack size* was kept fixed at 15% and *filler size* values were varied from 1% to 15%. We displayed the outcomes in Fig. 2 and Fig. 3. As shown in Fig. 2 and Fig. 3, average and bandwagon binary push attacks are significantly effective. Binary forms of such push attacks achieve 56.75% and 39% *ratio shift* values when attack and filler sizes are set to 15%. When an attacker aims to push a target item whose 1's ratio is 30%, she can successfully push the item and 1's ratio can be increased to 86.75% if binary average push attack is employed. The result will be 69% if binary bandwagon push attack is chosen. Binary segmented and random push attacks can push a target item; however, their impacts are comparably smaller. At some attack sizes, random attack is not successful. With increasing attack and filler sizes, the success of binary attacks also increases. Improvements in *ratio shift* are noticeable for smaller attack and filler sizes. They then become stable as attack and filler sizes become larger.

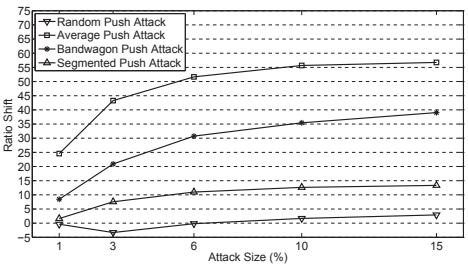

Fig. 2. *Ratio shift* values for varying *attack size* in binary push attack

We performed another set of experiments for binary nuke attacks. All attack types but segmented attack were employed. We followed the same methodology used in previous trials. Since *ratio shift* values are negative numbers for successful nuke attacks, we displayed absolute values of the outcomes. Since obtained similar outcomes, we showed outcomes for varying attack sizes only in Fig. 4. As seen from Fig. 4, we observed similar outcomes as in push attacks. For smaller attack and filler sizes, improvements in *ratio shift* are significant. However, as they

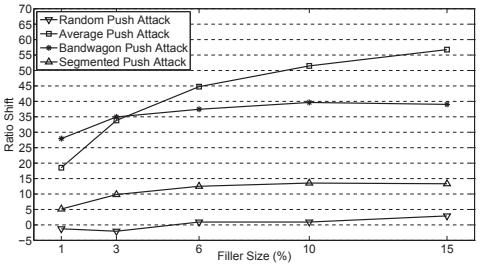

Fig. 3. *Ratio shift* values for varying *filler size* in binary push attack

become larger, such improvements become stable. Binary average nuke attack is the most successful attack type for nuking an item. Bandwagon, love/hate, and reverse bandwagon nuke attacks can be employed for nuking. However, their success ratio is smaller. Although the results for varying attack sizes are similar with binary push attacks' results, the outcomes for varying filler sizes for binary nuke attacks differentiate from the values in push attacks. With increasing filler sizes, *ratio shift* decreases for random and bandwagon binary nuke attacks. However, they can still manipulate an item's popularity for nuking purposes if controlling parameters are set to smaller values.

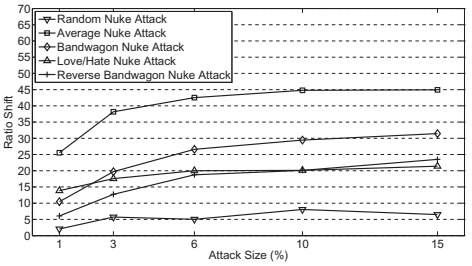

Fig. 4. *Ratio shift* values for varying *attack size* in binary nuke attack

5 Conclusions and Future Work

We proposed binary form of six shilling attack types and assessed their effects on NBC-based CF. We proposed a new metric, *ratio shift*, to measure the success of the proposed attacks. We performed real data-based trials and the outcomes indicated that it is possible to manipulate predictions. Thus, NBC-based CF is not a robust algorithm. We also pointed out that average and bandwagon attacks are the most successful attack for pushing. Average attacks achieves the best outcomes for nuking. As future studies, we plan to introduce new binary attack types and examine their effects on predictions. Also, we plan to analyze robustness of privacy-preserving NBC-based CF against profile injection attacks.

Acknowledgements. This work is supported by the Grant 111E218 from TUBITAK.

References

1. Miyahara, K., Pazzani, M.J.: Collaborative filtering with the simple Bayesian classifier. In: The 6th Pacific Rim International Conference on Artificial Intelligence, Melbourne, Australia, pp. 679–689 (2000)
2. O'Mahony, M.P., Hurley, N.J., Silvestre, G.C.M.: Towards robust collaborative filtering. In: O'Neill, M., Sutcliffe, R.F.E., Ryan, C., Eaton, M., Griffith, N.J.L. (eds.) AICS 2002. LNCS (LNAI), vol. 2464, pp. 87–94. Springer, Heidelberg (2002)
3. O'Mahony, M.P., Hurley, N.J., Silvestre, G.C.M.: Promoting recommendations: An attack on collaborative filtering. In: Hameurlain, A., Cicchetti, R., Traunmüller, R. (eds.) DEXA 2002. LNCS, vol. 2453, pp. 494–503. Springer, Heidelberg (2002)
4. Burke, R.D., Mobasher, B., Bhaumik, R.: Limited knowledge shilling attacks in collaborative filtering systems. In: Workshop on Intelligent Techniques for Web Personalization, Edinburgh, UK (2005)
5. Burke, R.D., Mobasher, B., Bhaumik, R., Williams, C.A.: Segment-based injection attacks against collaborative filtering recommender systems. In: The 5th IEEE International Conference on Data Mining, Houston, TX, USA, pp. 577–580 (2005)
6. Dellarocas, C.: Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In: The 2nd ACM Conference on Electronic Commerce, Minneapolis, MN, USA, pp. 150–157 (2000)
7. O'Mahony, M.P.: Towards robust and efficient automated collaborative filtering. PhD Dissertation, University College Dublin (2004)
8. Lam, S.K., Riedl, J.T.: Shilling recommender systems for fun and profit. In: The 13th International Conference on WWW, New York, NY, USA, pp. 393–402 (2004)
9. Lam, S.K., Riedl, J.T.: Privacy, shilling, and the value of information in recommender systems. In: User Modeling Workshop on Privacy-Enhanced Personalization, Edinburgh, UK, pp. 85–92 (2005)
10. Mobasher, B., Burke, R.D., Bhaumik, R., Sandvig, J.J.: Attacks and remedies in collaborative recommendation. IEEE Intelligent Systems 22(3), 56–63 (2007)
11. Mobasher, B., Burke, R.D., Bhaumik, R., Williams, C.A.: Towards trustworthy recommender systems: An analysis of attack models and algorithm robustness. ACM Transactions on Internet Technology 7(4), 23–60 (2007)
12. Cheng, Z., Hurley, N.J.: Effective diverse and obfuscated attacks on model-based recommender systems. In: The 3rd ACM International Conference on Recommender Systems, New York, NY, USA, pp. 141–148 (2009)
13. Oostendorp, N., Sami, R.: The copied-item injection attack. In: Workshop on Recommender Systems and the Social Web, New York, NY, USA, pp. 63–70 (2009)
14. Gunes, I., Kaleli, C., Bilge, A., Polat, H.: Shilling attacks against recommender systems: A comprehensive survey. Artificial Intelligence Review (2012), doi: http://dx.doi.org/10.1007/s10462-012-9364-9