Proceedings of the 2018 International Conference on Machine Learning and Cybernetics, Chengdu, China, 15-18 July, 2018

DETECTION OF SHILLING ATTACK IN COLLABORATIVE FILTERING RECOMMENDER SYSTEM BY PCA AND DATA COMPLEXITY

FEI ZHANG[1], ZI-JUN DENG[2], ZHI-MIN HE[3], XIAO-CHUAN LIN[1], LI-LI SUN[1]

[1]College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China
[2]School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
[3]School of Electronic and Information Engineering, Foshan University, Foshan 528000, China
E-MAIL: zhangfei@htu.edu.cn

Abstract:

Collaborative filtering (CF) recommender system has been widely used for its well performing in personalized recommendation, but CF recommender system is vulnerable to shilling attacks in which shilling attack profiles are injected into the system by attackers to affect recommendations. Design robust recommender system and propose attack detection methods are the main research direction to handle shilling attacks, among which unsupervised PCA is particularly effective in experiment, but if we have no information about the number of shilling attack profiles, the unsupervised PCA will be suffered. In this paper, a new unsupervised detection method which combine PCA and data complexity has been proposed to detect shilling attacks. In the proposed method, PCA is used to select suspected attack profiles, and data complexity is used to pick out the authentic profiles from suspected attack profiles. Compared with the traditional PCA, the proposed method could perform well and there is no need to determine the number of shilling attack profiles in advance.

Keywords:

Collaborative filtering; Recommender system; Shilling attack detection; PCA; Data complexity

1. Introduction

With the development of society, a large amount of data can be obtained, but it is easy for us to be suffered by information overload, and it is more difficult to effectively get the information we need [1]. Personalized recommender system has been proposed to solve the problem of information overload, which could provide the users interested information actively [2, 1, 20]. Collaborative filtering (CF) recommender system [3] is one of the effective technologies in personalized recommender system, which has been widely applied in electronic commerce, such as Amazon, Last.fm, Netflix Prize etc. Based on the concept that the target users may have the same tastes as their neighbors, CF recommender system could provide recommendations to the target users according to their nearest neighbors' preferences [4]. But if the preferences of the nearest neighborhoods' are fabricated, the system may provide inaccurate recommendations to the target users. That is, CF recommender system could effectively provide the information we need, but it is vulnerable to shilling attacks in which shilling attack profiles are injected into the system by attackers to alter recommendations [5].

To reduce the impact of shilling attacks in CF recommender system, robust recommender system design and attack detection methods have been proposed [6, 9]. One of the effective and simplest detection method is PCA, PCA could provide an accurate experimental result if we could get prior information of the amount of attack profiles (attack size of the shilling attack profiles). But in reality, we could not know the attack size in advance, so the validity of PCA is suffered. To address this problem, in this paper, PCA and data complexity are combined to detect shilling attack profiles. Data complexity is skilled in measuring the intrinsic characteristics of the given data set, and data complexity is verified that it could distinguish attack samples from normal ones [7]. In the proposed method, shilling attack profiles are detected by traditional PCA firstly, then the label of each profile is flipped and data complexity is calculated before and after the label flips, the change of the data complexity measures could illustrate whether the target profile is attacked or not. Experimental result confirms that the proposed method gets rid of the limitation of the traditional PCA and even performs better.

The paper is organized as follows. Sect. 2 gives a brief p-

978-1-5386-5214-5/18/$31.00 ©2018 IEEE

resentation on shilling attacks, PCA detection, supervised detection features and data complexity. Sect. 3 illustrate the proposed method in detail. Experimental analysis is shown in Sect. 4. Finally, the conclusions are provided in Sect. 5.

2. Related Work

In this section, shilling attacks, unsupervised PCA detection, supervised detection features and data complexity are introduced briefly.

2.1 Shilling Attack

Because of relying on the rating information of users, CF recommender system are easy to be manipulated by attackers to reach the purposes of profit. Shilling attacks can be separated into push attacks and nuke attacks according to attackers' purpose. Push (nuke) attacks increase (decrease) the recommended frequency of some particular items to target users. To distinguish shilling attack profiles and authentic profiles, shilling attack profiles are assumed to have the following structure, as is shown in table 1.

TABLE 1. General structure of a shilling attack profiles

Item	i^S	i^F	$i^\emptyset$	i^T
Rating	$\delta(i^S)$	$\sigma(i^F)$	$null$	$\gamma(i^T)$

i^F refers to filler items, filler items are selected such that the attack profiles may be appear similar to authentic profiles, which could reduce the probability of detection. i^S refers to the selected items, which could make the attack more effective. i^T refers to the target items, increase or decrease recommended frequency of target items to obtain illegal profit is the ultimate purpose of the attacker.

TABLE 2. Commonly used push attack models

Attack Model	$\delta(i^S)$	$\sigma(i^F)$	$\gamma(i^T)$
Random	$null$	N($\mu_{total}, \delta^2_{total}$)	max rating
Average	$null$	N($\mu_{i^F}, \delta^2_{i^F}$)	max rating
Bandwagon	max rating	N($\mu_{total}, \delta^2_{total}$)	max rating

Table 2 shows construction of shilling attack profiles of three commonly used push attack models, all the target items are given maximum rating values. Random attack randomly chooses some items as filler items and generate random ratings based gaussian distribution of total rating situation (N($\mu_{total}, \delta^2_{total}$). It is the simplest model that easy to carry out by attackers, but its attack effects is not fine. For average attack, the attacker need more information to do the attack, so the randomly selected filler items are filled with gaussian distribution of the rating situation of its corresponding filler item (N($\mu_{i^F}, \delta^2_{i^F}$), so average attack is highly effective than random attack. Bandwagon attack not only chooses some filler items, but also selects small amount of popular items which are easy to obtain, popular items which get maximum ratings could narrow the gap of shilling profiles and normal profiles, so the attack effect of bandwagon is acceptable. Other attack models such as segment attack, reverse bandwagon attack, AOP attack, obfuscated attack and even hybrid attack can be found in [13] for detail information.

To describe characteristics of an attack, attack size and filler size are used to evaluate the attack strength of the shilling attack model. Attack size refers to how many shilling profiles are injected into a recommender system , and filler size is the ratio between rated items and total items of each profile. Experiment in [9] shows that higher filler size leads to stronger effect of attack but instead will weaken the attack effect if it increases up to very high. when attack size is not so high, the effect of attack will becomes better as attack size rises. However, too high attack size doesn't better the attack effect and will increases the risk of detection. So in this paper, the attack size is given for 0.01, 0.05,0.1 and 0.15, and filler size is given as 0.01, 0.03, 0.05 and 0.1, if the attack profiles can be detected in these scenarios, then the validation of detection method is verified.

Because of shilling attacks, recommender system may recommend unnecessary or harmful information to users, and recommender system may lose the users' trust over time. To reduce the impact of shilling attacks, robust recommender system design method and attack detection method are proposed to defend against shilling attacks [9, 19]. For supervised methods, Chirita and Burke *et al.* [8] put forward several generic features and some model-specific features to distinguish shilling attack profiles, these detection features combines C4.5 have been discussed by Williams *et al.*, and feature selection method are used to probe the attack profiles [18], but these method could not copy with novel attacks and hybrid attacks. For unsupervised methods, Mehta [10] proposed PCA detection method and Bryan *et al.* [17] presents UnRAP detection method, skillfully making use of the characteristics of shilling users, but these methods need some prior information in advance. For semi-supervised methods, Wu *et al.* [16] proposed HySAD to handle hybrid attack, but this method more or less depends on some man-made features.

2.2 PCA Detection

PCA (Primary Component Analysis) is designed to reduce dimensionality of a data set. The original data is projected into another low dimensional space while the information of the original data almost reserved, that is, PCA could discard redundant dimensionality which carry only few information. Mehta found that covariance among shilling users is much lower than covariance among authentic users [15]. In addition, covariance between shilling users and authentic users is still very low [14]. From the concept of PCA, we can obtain that shilling users are redundant since they provide few new information in user rating matrix, and PCA can discard these shilling users.

Algorithm 1 PCA Detection

Input:

R: users' rating matrix ;

k: cutoff parameter

Output:

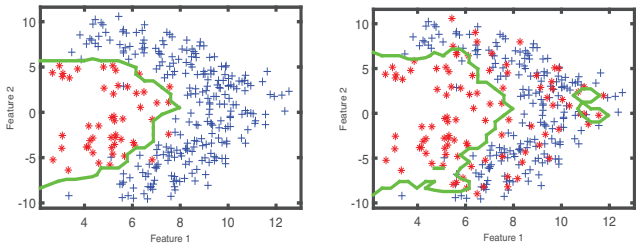
A list of k suspected shilling users;

1: Calculate user-user correlation matrix $Corr_U$ on matrix R;
2: Do Eigenvalue Decomposition for $Corr_U$;
3: Select the first two eigenvectors as primary components (PCA_1 ,PCA_2) ;
4: Calculate distance of each user u after dimension reduction $Distance(u) = PCA_1(u)^2 + PCA_2(u)^2$;
5: **return** k users with smallest Distance

Algorithm 1 provides the procedure for how to select shilling attack profiles by PCA. The first step is to calculate the correlations between profiles (line 1). The second step is to do eigenvalue decomposition, and select the eigenvectors which have the first k largest eigenvalue (line 2-3), k is always selected as 2 or 3. Finally, calculate the distance of each user profile on the projection space (line 4). The detection method PCA choose top k users which have the smallest distances as shilling attack users, PCA performs well when k equals to the attack size. However, in most cases, we have no prior information of the attack size. If k is smaller than the attack size, there must exist some shilling users escaping from detection. If k is larger than the attack size, some authentic users will be mistakenly regarded as shilling users. Therefore, an improved detection method based on PCA is proposed, the proposed method could achieve good performance without the information k in advance.

2.3 Data Complexity

Data complexity puts forward by Tin Kam Ho [11] to evaluate the complexity of a classification problem on a data set. But the applications of data complexity are more than that. Data complexity measures a data set by characterizing its geometrical characteristics [12]. We can see a small example in figure 1, where there are 2 classification problem classified by the k-nn classifier. Observed visually, sample distribution in the first case appears more "neat" and its classification boundary is simple. However, sample distribution in the second case appears more interleaved and its classification boundary is very complicated. Data complexity will provide a good measure to evaluate the classification complexity.



(a) Less complex case (b) More complex case

FIGURE 1. Data complexity of two classification problem

Since the details of data complexity are not the discussion emphasis of this paper, we will just give a brief introduction to them. Although there are many extensions features of data complexity, here we only consider the main features, which are F1, F1v, F2, F3, F4, L1, L2, L3, N1, N2, N3, N4 and T1. They respectively describe characteristics of a data set from the perspective of features discrimination, overlapping region of classes, error situation of classification, class boundary details, interleaved situation of classes and so on. Table 3 shows the correlation between features and classification complexity.

TABLE 3. Correlation between features and classification complexity.

Data Complexity Features	Correlation
F1, F1v, F3, F4	Negative (Higher value indicates lower classification complexity)
F2, N1, N2, N3, N4, L1, L2, L3, T1	Positive (Higher value indicates higher classification complexity)

Data complexity will be applied to evaluate the classification complexity of the 2-class data set. Here we put forward $CC_{Measure}$ to evaluate that.

$$CC_{Measure} = \sum_{i=1}^{13} (-1)^m DataComplexityFeature(i) \quad (1)$$

where m = 1 if $DataComplexityFeature(i)$ is negatively correlated to classification complexity and m = 2 if

$DataComplexityFeature(i)$ is positively correlated to classification complexity.

2.4 Detection Features

In this paper, it is unreasonable to use data complexity directly, for the user rating matrix is so sparse that the data complexity measure could not provide an accurate result. So in the proposed method, after the shilling attack profiles are selected by PCA, a two-class problem can be constructed, and then supervised detection features will be employed to calculate data complexity measures. Four commonly used generic attributes are selected to distinguish attack profiles and normal profiles, such as RDMA, WDMA, WDA and LengthVar. Besides, some model-specified features are suitable to detect specified attack. More details can be seen in [1]

The motivation behind attacker and authentic user is different, and the statistical characteristics of attack profiles and authentic profiles. RDMA estimate the inconformity of user's rated item and the item's average score. Derived from RDMA, WDMA place a higher weight on the sparse rated items, which could provide more information. Based on RDMA and WDMA, WDA ignores the user's rated item number. LengthVar captures the deviation between user's rated item number and the system's rated item number, LengthVar is designed to detect profiles that containing small or large amount of rated items. Features mentioned above are frequently applied in supervised and semi-supervised detection method. So the proposed method also use them as tools to measure the data complexity characteristics of each user.

Algorithm 2 Proposed Detection Method

Input:
R: users' rating matrix ;
k: cutoff parameter of PCA;
Output:
A list of shilling users;
1: $DetFea$ = { }; U = all users in R;
2: **for** each user's rating r in R **do**
3: $DetFea$ = [$DetFea$; DectFeatures(r)];
4: **end for**
5: $AttUsers$ = PCA(R,k);
6: $NorUsers$ = U $-$ $AttUsers$;
7: $DetFea(AttUsers)$ = [$DetFea(AttUsers)$, +1];
8: $DetFea(NorUsers)$ = [$DetFea(NorUsers)$, -1];
9: $BaseDc$ = DataComplexity($DetFea$);
10: CC_{base} = $CC_{Measure}(BaseDc)$;
11: CC_{Change} = zeros(k,1);
12: **for** $i = 1$ to k **do**
13: $TempUserFeatMat$ = $DetFea$;
14: $TempUserFeatMat(AttUsers(i), end)$ = -1;
15: $TempDc$ = DataComplexity($TempUserFeatMat$);
16: $CC_{Change}(i)$ = $CC_{Measure}(TempDc) - CC_{base}$;
17: **end for**
18: **for** $i = 1$ to k **do**
19: **if** $CC_{Change}(i) < 0$ **then**
20: $DetFea(AttUsers(i), end)$ = -1;
21: **end if**
22: **end for**
23: **return** users in $DetFea$ who have $positive$ label

3. Proposed Detection Method

PCA could achieve sufficiently good performance when k is equal to the number of shilling attack profiles, but we have no information about the amount of shilling attack users in advance. From algorithm 1, it is obvious that when k is large enough, almost all of the shilling attack users are included in top k users. Since k is larger than the real number of attack profiles, there must be authentic users among the top k users. The main problem we need to concern about is how to detect and then remove the authentic users from the top k users, algorithm 2 presents a detail illustration of how to detect shilling attack users with few authentic users misclassified.

In algorithm 2, detection features RDMA, WDMA, WDA and LengthVar are calculated for each user which are reserved for data complexity measurements (line 1-4). PCA is then applied to detect shilling attack profiles, and the cutoff parameter k will be given a larger value, which means that a number of actually authentic users are mistakenly regarded as shilling users. After PCA detection, we can separate the users into two classes in the space spanned by detection features, and the detected profiles will given positive labels, while other profiles obtain negative labels (line 5-8). From figure 1, we could obtain that mislabeled sample will increase the data complexity of the classification problem, $CC_Measure$ (equation 1) is used to evaluate the classification complexity. But if the label of an actually authentic user is flipped from positive to negative, the classification complexity maybe decrease. We cannot give strict guarantee for this statement, but in most of cases it's correct. Based on the change of data complexity before and after label flips, whether a user profile given a correct label can be determinate (line 9-22). As described above, the proposed method could ingeniously utilize the advantage of PCA detection and data complexity method, and there is no need to obtain the amount of shilling attack profiles in advance.

4. Experiment result analysis

Experimental result of the proposed method will be evaluated and analyzed in this section. MovieLens-100k data set, which contains 943 authentic users and 1682 movie items, is used in our experiment. Push shilling attacks, which are frequently occurring in reality, are considered in this paper, and shilling attack profiles are constructed by three commonly used attack model mentioned in sec. 2.1.

To verify the availability of the proposed method, PCA detection method is used as the comparative method, and F-measure is applied to evaluate the detection result. Equation 2 presents the formula of F-measure, higher F-measure indicates better detection result. In our experiment, filler size, attack size and PCA cutoff parameter k are all the variables we need to consider. The experiments have been implemented for different setting of attack size, filler size and k, where each case is repeated independently for 5 times. To have a obvious comparison of detection result of PCA and proposed method, average result of different filler size for the same attack size and k is exhibit in table 4, 5, 6.

$$\begin{aligned} Precision &= \frac{\#TruePositive}{\#TruePositive + \#FalsePositive} \\ Recall &= \frac{\#TruePositive}{\#TruePositive + \#FalseNegative} \\ F-measure &= \frac{2 * Precision * Recall}{Precision + Recall} \end{aligned} \tag{2}$$

TABLE 4. F-measure of detection for random attack

Attack Size	k					Proposed method
	1%	5%	10%	15%	20%	
1%	0.91	0.32	0.17	0.12	0.09	0.24
5%	0.35	0.96	0.64	0.48	0.38	0.76
10%	0.21	0.71	0.95	0.75	0.62	0.94
15%	0.14	0.56	0.85	0.94	0.80	0.96

TABLE 5. F-measure of detection for average attack

Attack Size	k					Proposed method
	3%	5%	10%	15%	20%	
1%	0.88	0.32	0.17	0.12	0.09	0.14
5%	0.35	0.94	0.64	0.48	0.38	0.66
10%	0.21	0.70	0.95	0.75	0.62	0.89
15%	0.14	0.56	0.86	0.93	0.79	0.93

As the results shown in the tables above, we could obtain that if k equals to the attack size, PCA detection method can achieve best detection effects, but in actual circumstance, it is hard to choose a suitable parameter k, which means it is hard for PCA to reach such a good detection result all the time. For the proposed method, k is set to a fixed value, i.e. 20% of the total number of collected profiles, include authentic profiles and attacked profiles. It is obvious that as the increase of attack size, the performance of the proposed method is increased for the three attack models, the main reason is that as the increase of attack size, less misclassified authentic profiles affect the accurate of the base data complexity measures (as is shown in algorithm 2 line 10). The proposed method performs well than PCA almost all the cases if PCA has no information of the attack size in advance, no matter in which attack model. If the attack size is larger, the proposed method also performs well than PCA. Hence, it is possible and feasible for us to select a larger k for the proposed method to reach excellent results. What is more, these results confirm the effective of data complexity to detect label flips attacks once again.

TABLE 6. F-measure of detection for bandwagon attack

Attack Size	k					Proposed method
	3%	5%	10%	15%	20%	
1%	0.65	0.32	0.17	0.12	0.09	0.19
5%	0.31	0.90	0.64	0.48	0.38	0.75
10%	0.19	0.68	0.94	0.75	0.62	0.94
15%	0.13	0.54	0.84	0.93	0.79	0.96

5. Conclusions

Recommender system may provide unnecessary or harmful recommendations for users because of shilling attacks, PCA is a simple and effective method to detect shilling attacks if we have prior information of attack size, otherwise, PCA will be suffered. In this paper, we demonstrate a new method combining PCA and data complexity. Because data complexity is designed to solve a 2-class classification problem, and data complexity can be applied to detect label flips attack, so PCA is used to pick out some suspicious shilling attack profiles firstly, then data complexity is calculated before and after label flips of each suspicious attack profile. Experimental results show that in most of cases, the proposed method can better recognize the shilling attack users in recommender system than PCA detection method.

Acknowledgements

This work is supported by Scientific Research Foundation of Henan Normal University (5101119170132), Research Grants for Universities and Colleges in Henan (17A520037), National Natural Science Foundation of China (Grant Nos.

61802061) and Project of Department of Education of Guangdong Province.(No. 2017KQNCX216).

References

[1] Burke R, O Mahony M P, Hurley N J. Robust Collaborative Recommendation[J]. Recommender Systems Handbook, 2015:805-835.

[2] Caamares R, Castells P. Should I Follow the Crowd?: A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems[C]// The, International ACM SIGIR Conference. ACM, 2018:415-424.

[3] Chung C Y, Hsu P Y, Huang S H. A novel approach to filter out malicious rating profiles from recommender systems[J]. Decision Support Systems, 2013, 55(1):314-325.

[4] Schafer J.B., Frankowski D., Herlocker J., Sen S. (2007) Collaborative Filtering Recommender Systems. In: Brusilovsky P., Kobsa A., Nejdl W. (eds) The Adaptive Web. Lecture Notes in Computer Science, vol 4321.

[5] Yang Z, Cai Z. Detecting abnormal profiles in collaborative filtering recommender systems[J]. Journal of Intelligent Information Systems, 2017, 48(3):499-518.

[6] Gunes I, Kaleli C, Bilge A, et al. Shilling attacks against recommender systems: a comprehensive survey[J]. Artificial Intelligence Review, 2014, 42(4):767-799.

[7] Chan Patrick P. K., He Z M, Li H, et al. Data sanitization against adversarial label contamination based on data complexity[J]. International Journal of Machine Learning & Cybernetics, 2018, 9(6):1039-1052.

[8] Burke R, Mobasher B, Williams C, et al. Classification features for attack detection in collaborative recommender systems[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 542-547.

[9] Burke R, Mobasher B, Bhaumik R. Limited knowledge shilling attacks in collaborative filtering systems[C]//Proceedings of 3rd International Workshop on Intelligent Techniques for Web Personalization (ITWP 2005), 19th International Joint Conference on Artificial Intelligence (IJCAI 2005). 2005: 17-24.

[10] Mehta B. Unsupervised shilling detection for collaborative filtering[C]//AAAI. 2007: 1402-1407.

[11] Ho T K, Basu M. Complexity measures of supervised classification problems[J]. IEEE transactions on pattern analysis and machine intelligence, 2002, 24(3): 289-300.

[12] Ho T K, Basu M, Law M H C. Measures of geometrical complexity in classification problems[M]//Data complexity in pattern recognition. Springer London, 2006: 1-23.

[13] Cheng Z, Hurley N. Effective diverse and obfuscated attacks on model-based recommender systems[C]//Proceedings of the third ACM conference on Recommender systems. ACM, 2009: 141-148.

[14] Mehta B, Hofmann T, Fankhauser P. Lies and propaganda: detecting spam users in collaborative filtering[C]//Proceedings of the 12th international conference on Intelligent user interfaces. ACM, 2007: 14-21.

[15] Mehta B, Nejdl W. Unsupervised strategies for shilling detection and robust collaborative filtering[J]. User Modeling and User-Adapted Interaction, 2009, 19(1-2): 65-97.

[16] Wu Z, Wu J, Cao J, et al. HySAD:a semi-supervised hybrid shilling attack detector for trustworthy product recommendation[C]// Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012:985-993.

[17] Bryan K, O'Mahony M. Unsupervised retrieval of attack profiles in collaborative recommender systems[C]// ACM Conference on Recommender Systems. ACM, 2008:155-162.

[18] WU Z,Zhang Y, Wang Y, et al. Shilling Attack Detection Based on Feature Selection for Recommendation Systems,ACTA ELECTRONICA SINICA, Vol.40 No. 8,1687-1693.

[19] Kaur P, Goel S G. Shilling Attack Detection in Recommender Systems[D]. , 2016.

[20] Logesh R, Subramaniyaswamy V, Vijayakumar V. A personalised travel recommender system utilising social network profile and accurate GPS data[J]. Electronic Government, an International Journal, 2018, 14(1): 90-113.