EasyChair Preprint

№ 247

Shilling Attack Detection Based on Data Tracking

Lingtao Qi, Haiping Huang, Peng Wang and Ruchuan Wang

June 9, 2018

Shilling Attack Detection Based on Data Tracking

Lingtao Qi
Dept. of Computer Science and Technology
Nanjing University of Posts and Telecommunications
Nanjing, China
851388561@qq.com

*Haiping Huang
Dept. of Computer Science and Technology
Nanjing University of Posts and Telecommunications
Nanjing, China
hhp@njupt.edu.cn

Peng Wang
Dept. of Computer Science and Technology
Nanjing University of Posts and Telecommunications
Nanjing, China
1256037701@qq.com

Ruchuan Wang
Dept. of Computer Science and Technology
Nanjing University of Posts and Telecommunications
Nanjing, China
wangrc@njupt.edu.cn

***Abstract*—Collaborative filtering recommender system is one of the most widely used recommender systems, while it is vulnerable to shilling attack because of its openness. In recent years, many shilling attack detection methods have been proposed and achieved some results. However, with the rapid growth of data, the detection efficiency of existing detection methods can not meet the requirements. To solve the above problem, a detection algorithm based on data tracking adapted to Big Data environment is proposed. Based on two new data features, the algorithm uses extended Kalman filter to track and predict the item's rating, and detects the abnormal item in real-time efficiently. Experimental comparison shows that this algorithm has high detection rate and small time overhead.**

***Keywords*—collaborative filtering recommender system, shilling attack detection, extended Kalman filter, data tracking.**

I. INTRODUCTION

Collaborative filtering recommender system has been widely used and has brought huge economic benefits for many companies, such as Netflix, Amazon and so on. Since the collaborative filtering recommender system makes recommendation based on item's ratings given by users, it is vulnerable to shilling attack [1-3]. Shilling attack manipulates the recommender system to promote the recommendation of certain items, or derogate the recommendation of certain items. This problem will affect the authenticity of the recommender system and lead to the users and potential users' distrust of the system.

For the problems caused by the shilling attack on the collaborative filtering recommender system, Burke et al. [4] Trained several supervised classifiers to detect the attacks by extracting features of user profiles. These classifiers can detect several kinds of attacks,but they suffer from low accuracy. Zhang et al. [5] proposed a supervised learning detector based on Hilbert-Huang transform (HHT) and Support Vector Machine (SVM). This method extracts and verifies the user profile features based on the new detection attributes, which improves the detection accuracy of the SVM method. However, when a new attack is formed, the SVM classifier needs to be trained offline. On the other hand, an unsupervised method UnRAP [6] was proposed to detect shilling profiles by analyzing user profiles' rating deviation on the target item. It overcomes the limitation of supervised learning detectors and has higher detection precision and recall rate, while it can only detect the attack users profiles of individual items. For the problem of UnRAP, Qing et al. [7] proposed the AP-UnRAP detection algorithm. In their method, every single attack behavior is analyzed and then clustered into a attack group. In most circumstances, its detection performance is better than UnRAP algorithm.

However, the advent of Big Data era makes the data volume of recommender system become great, and the existing detection algorithms are time-consuming and inefficient. In this paper, a shilling attack detection algorithm based on data tracking (DT) is proposed, which is adapted to Big Data processing. The extended Kalman filter (EKF) is first used in this field to quickly track and accurately predict the rating status of the item. And then the detector determines the abnormal item according to the comparison of predicted ratings and actual ones.

The rest of the paper is structured as follows: we present a brief introduction on shilling attack and extended Kalman filter in Sect. II. Sect. III illustrates the proposed method in detail. The experimental results are discussed in Sect. IV. Finally, the conclusions are provided in Sect. V.

II. BACKGROUND

A. Shilling Attack

Shilling attacks try to manipulate the recommendation list by injecting deliberately constructed false ratings, which push target items into more user's recommendation list (for push attack) or prevent target items entering the recommendation list (for nuke attack). In this paper, it only takes push attack as an example.

The basic framework of shilling attack is proposed by researchers based on the purpose of attack, the scale of attack and the preliminary knowledge [3, 8]. I_T is a set of target items, I_S is a set of items deliberately selected based on the purpose of attack, I_F is a set of filler items randomly selected, and $I_\emptyset$ is a set of items those are not rated.

Table I lists two kinds of popular attack models (where r_{max} is the highest rating): random attack and bandwagon attack. For random attack, I_F is randomly selected and the ratings of I_F are subject to the normal distribution of mean and standard deviation of all the ratings in the data set. I_S is null and the target items are rated with the highest score. For bandwagon attack, I_S is the set of items widely rated. I_F is randomly selected and the ratings of I_F are the the average mean of the corresponding items.

TABLE I. TWO KINDS OF SHILLING ATTACK MODELS

Attack Models	I_T	I_S	I_F	$I_\emptyset$
Random Attack	$r_{\max}$	Ø	random	Ø
Bandwagon Attack	$r_{\max}$	widely rated items	item mean	Ø

The intensity of attack is usually measured in terms of attack size and filler size. Attack size refers to the proportion of attack user profiles in the recommender system. Filler size refers to the ratio of the number of items rated by attacker to the total number of items in the recommender system, which describes the sparseness of the item ratings.

B. Extended Kalman Filter

The extended Kalman filter method (EKF) is an extension of the standard Kalman filter method in nonlinear systems. The core idea of EKF is to obtain an approximate linear model by expanding the nonlinear function into the Taylor series around the filter values and omitting the second-order and above items in the Taylor expansion, and then apply Kalman filter to complete the state estimation [9]. At present, EKF is widely used in edge detection, image tracking, target recognition and so on. In practical, it is difficult to solve the nonlinear filtering problem by obtaining the posterior probability density function of the target state. Therefore, in view of the changing preferences of interest reflected by item rating matrix, EKF uses the non-linear item rating data as a model to make Taylor series expand near the state estimation value. The first-order approximation term obtained by truncation is taken as the system state equation and the observation state equation to achieve the linear calculation. The state estimation is compared with the actual status as the predicted status and the items with normal status will be excluded from abnormal items for the next phase of continuous observation.

III. DETECTION METHOD BASED ON DATA TRACKING

In this paper, two new attributes SACA and SVCA are proposed to describe the data matrix. Data tracking detection algorithm tracks and predicts these two characteristics in real time. Meanwhile, the abnormal items are detected based on the continuous comparison between the predicted state and the actual state.

Several definitions used in our proposal are shown as follows.

1) User set (the set of m users in the system): $U=\{U_1, U_2, \ldots, U_m\}$.

2) Item set (the set of n items in the system): $J=\{j_1, j_2, \ldots, j_n\}$.

3) Short-term average change activity (SACA). It reflects that the attackers constantly push up the target item's ratings, which leads to a short period of rapid ascension of the average score of target item. The SACA is shown in Eq. 1:

$$SACA_{j,t} = \left| \left(\frac{Over\_avg_{j,t+1}}{|F_{j,t+1}|} \right) - \left(\frac{Over\_avg_{j,t}}{|F_{j,t}|} \right) \right|, \quad Over\_avg_{j,t} = \begin{cases} 1, avg_{j,t+1} > avg_{j,t} + \tau \\ 0, otherwise \end{cases} \tag{1}$$

where $avg_{j,t}$ represents the average score of item j at time t, τ is the corrected value of SACA. In the practical scenario, the average score of item at time t+1 may be slightly larger than that at time t however if the increment exceeds τ, this item will be considered as an abnormal one. $F_{j,t}$ represents the set of all ratings to item j except those abnormal ratings (i.e. the highest score) at time t, and meanwhile $|F_{j,t}|$ represents the number of ratings in $F_{j,t}$. All in all, when the target item in the recommender system encounters the profile injection group attack, its SACA value will quickly increase and maintain a high value in the short term.

4) Short-term variance change activity (SVCA). It reflects that the attackers consecutively push up the target item's rating which causes the score variance of target item decreases rapidly during a short period. This attribute can be used to show the tempestuous change of ratings of target item in the short term, which is expressed by Eq. 2 :

$$SVCA_{j,t+1} = \left| var_{j,t+1} - var_{j,t} - \upsilon |F_{j,t}| \right| \tag{2}$$

where $var_{j,t}$ represents the score variance of item j at time t, υ (determined by the amount of data) is the corrected value of SVCA, which is used to fix the difference value of variance between time t+1 and time t. And only when the score variance of item j changes sharply and exceeds υ, item j can be regarded as an abnormal one.

The system state of item j at time t is $\mathbf{x}(t) = \begin{pmatrix} SACA_t \\ SVCA_t \end{pmatrix}$. The state transition equation and observation equation of item j at time t are shown in Eq. 3 and 4, respectively:

$$\mathbf{X}_t = f(\mathbf{X}_{t-1}, \mathbf{u}_t) + \mathbf{W}_t \tag{3}$$

$$\mathbf{Y}_t = h(\mathbf{X}_t) + \mathbf{V}_t \tag{4}$$

where $\mathbf{u}_t$ is the control matrix, $\mathbf{X}_t$ is the true state at time t, $\mathbf{X}_{t-1}$ is the true state at time t-1, and $\mathbf{W}_t$ is the process noise and assumed to be subject to an independent multivariate normal distribution with the mean of 0 and the covariance matrix of $\mathbf{Q}_t$. $\mathbf{V}_t$ is observation noise and assumed to be subject to an independent multivariate normal distribution with the mean of 0 and the covariance matrix of $\mathbf{R}_t$. $f()$ and $h()$ are nonlinear functions obtained by the first-order Taylor expansion of the system.

The mathematical description of prediction is show as Eq. 5 ~ 8 :

$$\hat{\mathbf{X}}_{t|t-1} = f(\hat{\mathbf{X}}_{t-1|t-1}, \mathbf{u}_t) \tag{5}$$

$$\mathbf{P}_{t|t-1} = \mathbf{F}_t \mathbf{P}_{t-1|t-1} \mathbf{F}_t^T + \mathbf{Q}_t \tag{6}$$

$$\tilde{\mathbf{y}}_t = \mathbf{Y}_t - h(\hat{\mathbf{X}}_{t|t-1}) \tag{7}$$

$$\mathbf{S}_t = \mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^T + \mathbf{R}_t \tag{8}$$

where $\hat{\mathbf{X}}_{t|t-1}$ is the predict value of time t based on the value of time t-1, $\hat{\mathbf{X}}_{t-1|t-1}$ is the estimated state at time t-1, and $\mathbf{H}_t$ is the observation matrix that maps implicit real state space to the observed space. $\mathbf{P}_{t|t-1}$ is the covariance matrix of posterior estimation error used to measure the accuracy of the prediction. $\mathbf{F}_t$ is the transformation matrix acting on the state $\mathbf{X}_{t-1}$ at time t-1. $\tilde{\mathbf{y}}_t$ is the difference between the actual value and the estimated output. $\mathbf{S}_t$ is the covariance matrix of $\tilde{\mathbf{y}}_t$.

The process of update is show as Eq. 9 ~ 11 :

$$\hat{\mathbf{X}}_{t|t} = \hat{\mathbf{X}}_{t|t-1} + \mathbf{K}_t \tilde{\mathbf{y}}_t \tag{9}$$

$$\mathbf{K}_t = \mathbf{P}_{t|t-1} \mathbf{H}_t^T \mathbf{S}_t^{-1} \tag{10}$$

$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t \mathbf{H}_t) \mathbf{P}_{t|t-1} \tag{11}$$

where $\hat{\mathbf{X}}_{t|t}$ is the estimated state at time t and $\mathbf{K}_t$ is the Kalman gain. Eq. 10 finds the optimal Kalman gain and Eq. 11 simplifies the posteriori prediction error covariance matrix at the optimal Kalman gain.

The judgement formula based on the tracking result is show as Eq. 12 and 13, where $Total_j$ denotes the total number of times the item j tracked in a short-term T. $CONT_{j,t}$ denotes the proximity number regarding item j when the solution difference value between the state transition equation and the observation equation is less than the given minimum threshold ζ at time t, and meanwhile $|\mathbf{X}_{j,t}|$ must be greater than the abnormal state transition threshold ρ and $|\mathbf{Y}_{j,t}|$ must be greater than the observation outlier threshold ω. CAL_j is the probability of item j being a abnormal item. The closer is CAL_j to 1, the more likely item j is to be target item.

$$CONT_{j,t} = \begin{cases} 1, & \left| \mathbf{X}_{j,t} - \mathbf{Y}_{j,t} \right| \le \xi, |\mathbf{X}_{j,t}| \ge \rho, |\mathbf{Y}_{j,t}| \ge \omega \\ 0, & otherwise \end{cases} \tag{12}$$

$$CAL_j = \frac{\sum_{t \in T} CONT_{j,t}}{Total_j} \tag{13}$$

IV. Simulation and Experiments

The benchmark data set (MovieLens 20M) is used in our experiments, which includes 20 million ratings (1-5 marks) on 27,000 movies by 138,000 users. HHT-SVM [5] and AP-UnRAP [7] are the two detection methods used to compare with DT. Both of them are the typical detection methods in current research.

Random attacks and bandwagon attacks are employed to inject fake profiles into MovieLens data set. Filler size is 5% and attack sizes are set to 3% and 15%, respectively. SACA average correction value τ and SVCA variance correction value υ is determined by the number of data set. In this experiment, τ is set as 0.0001 and υ is set as 0.00025. The difference ζ of the state transition equation and observation equation is determined by the number of single item rating and it is set as 0.0001 in our experiments. The abnormal threshold ρ of the state transition is affected by the number of times of tracking prediction and it is set as 0.001. The threshold ω of the observation equation is affected by the variance of the normal distribution formed by the item ratings, and it is set as 0.025. In addition, T is set as 0.01 seconds.

In order to evaluate the detection performance of the detection methods, we used the three evaluation indexes: precision, recall and running time, and Eq. 14 and 15 respectively shows the mathematical expression of precision and recall.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

where TP indicates the number of abnormal items detected, FP indicates that the number of normal items are incorrectly detected as abnormal items, and FN indicates the number of abnormal items incorrectly detected as normal items. The greater the precision is, the less the number of normal items misjudged as abnormal items is. The higher the recall is, the less the number of undetected abnormal items is.

Figure 1 shows the comparisons of precision of the three detection algorithms. As the attack size increases, the precision of the three algorithms is also increasing. Furthermore, bandwagon attacks are more difficult to be detected than random attacks in this experimental scenario. The detection rate of DT algorithm is slightly higher than the other two algorithms, since the continous tracking of the status of item ratings can better discover its potential law, which helps the detector make decision. Figure 2 shows the comparisons of recall of the three detection methods. Faced with different attack size and attack model, the recall of DT is higher than the

two others because the use of continuous tracking and predicting in DT can reduce the possibility of misjudgment. Figure 3 shows the comparisons of running time of the three methods. Under the same operating condition, the running time of DT is far less than the two other methods. Since the DT algorithm only tracks suspicious data in real time, the amount of data needs to be processed is greatly reduced. While the other two algorithms cannot detect shilling attacks in real time.

According to the above experiments, it can be found that DT detection algorithm has the best performance in face of different attack sizes of random attack and bandwagon attack, especially in running time for mass data processing.

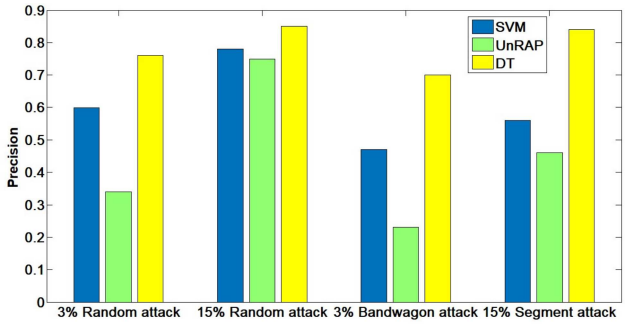

Fig. 1. Comparisons of precision of detection methods

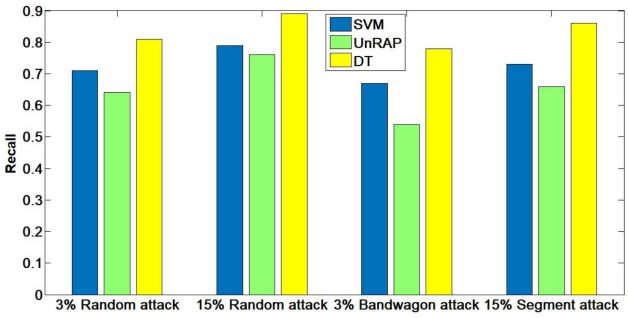

Fig. 2. Comparisons of recall of detection methods

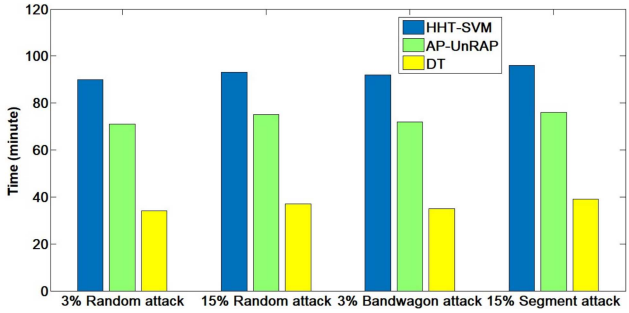

Fig. 3. Comparisons of the running time of detection methods

V. Conclusion

Shilling attack detection based on data tracking (DT) first adopts extended Kalman filter to continuously track and predict the item's rating status based on two new detection attributes SACA and SVCA, and meanwhile reduces the detection of a large number of unrelated data, thus improving the detection efficiency. Experiments on the MovieLens 20M data set show that this proposal greatly reduces the running time of detection while ensuring a high precision and recall. In the future, we will consider applying this method to distributed systems to handle larger data and further improve detection efficiency.

Acknowledgment

This work was supported by the National Natural Science Foundation of P. R. China (No. 61672297), the Key Research and Development Program of Jiangsu Province (Social Development Program, No.BE2017742), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. SJCX17_0235) and the Sixth Talent Peaks Project of Jiangsu Province (No. DZXX-017).

References

[1] M. O'Mahony, N. Hurley, N. Kushmerick and G. Silvestre. " Collaborative recommendation: a robustness analysis, " ACM Transactions on Internet Technology, 2004, 4(4):344-377.

[2] F. Cacheda, V. Carneiro, and D. Fernandez. "Comparison of collaborative filtering algorithms:limitations of current techniques and proposals for scalable, high-performance recommender systems," ACM Transactions on the Web, 2001, 5(1):3–34.

[3] I. Gunes, C. Kaleli, A. Bilge, and H. Polat. "Shilling attacks against recommender systems: a comprehensive survey," Artificial Intelligence Review, 2014, 42(4):767-799.

[4] R. Burke, B. Mobasher, C. Williams, and R. Bhaumik. "Classification features for attack detection in collaborative recommender systems," The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006: 542–547.

[5] F. Zhang, and Q. Zhou. "HHT–SVM: An online method for detecting profile injection attacks in collaborative recommender systems," Knowledge-Based Systems, 2014, 65(4):96-105.

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

[7] Q. X. Wang, Y. Ren, N. Q. He, M. Wan, and G. B. Lu. "A group attack detecter for collaborative filtering recommendation," International Computer Conference on Wavelet Active Media Technology and Information Processing, 2015:454-457.

[8] S. K. Lam, and J. Riedl. "Shilling recommender systems for fun and profit," International Conference on World Wide Web, 2004:393–402.

[9] S. Sabatelli, M. Galgani, L. Fanucci, and A. Rocchi. "A double-stage kalman filter for orientation tracking with an integrated processor in 9-d imu," IEEE Transactions on Instrumentation & Measurement, 2013, 62(3):590-598.