A Data Mining based Approach to Detect Attacks in Information System Filtering

Anshu Sharma
Department of Computer Science and Engineering
Lovely Professional University
Phagwara

Shilpa Sharma
Department of Computer Science and Engineering
Lovely Professional University
Phagwara

Chirag Sharma
Department of Computer Science and Engineering
Lovely Professional University
Phagwara

ABSTRACT

Securing information system filtering from malicious attacks has become an important issue with increasing popularity of information system filtering. Data mining is the analysis of observational data sets to find unsuspected relationships and to summarize the data novel ways that are both understandable and useful to data owners. Information systems are entirely based on the input provided by users or customers, they tend to become highly prone to attacks. To prevent such attacks several mechanisms can be used. In this paper, we show that the unsupervised clustering one of the data mining technique can be used for attack detection for all types of attacks. This method is based on computing detection attributes modeled on basic descriptive statistics. Our study showed that attribute based unsupervised clustering algorithm can detect spam users with high degree of accuracy and fewer misclassified genuine users regardless of attack strategies.

General Terms

Information system filtering.

Keywords

Data mining, clustering, information system filtering.

1. INTRODUCTION

Information systems are common targets for malicious attackers. The product sellers who are interested in promoting their own product to generate more revenue might be interested in biasing these systems which have the influence on customers. Such attackers can use automated tools to create and throw fake profiles in the information system database, which may rate their items high and may rate the opponent’s items low. These information systems must be open to users, in order to get the opinions of the users. This is the reason why designing an attack proof system is a complicated task.

In recent years research has shown that personalization based on explicit user feedback typically in the form of ratings. Those are vulnerable to profile injection or shilling attacks [1], [2] [3].

Early detection algorithms exploited signatures of attack profiles and were moderately accurate. However, these detection algorithms suffered form low accuracy in detecting shilling profiles, since they looked at individual users and mostly ignored the combined effects of such malicious users. Moreover, these algorithms did not perform well when the spam profiles are obfuscated.

Unsupervised anomaly detection approaches address these issues by training on an unlabelled data set. These methods involve much lesser computational effort as compared to supervised approaches, especially if the training data has to be generated. It also facilitates online learning and improves detection accuracy.

Mehta et al. [4] showed that clustering based on principal component analysis (PCA) performed very well against standard attacks when evaluated on Movielens dataset. The motivation behind this approach is that attacks consist of multiple profiles which are highly correlated with each other, as well as hiving high similarity with a large number of authentic profiles. However, while other attacks can be detected with high accuracy and fewer misclassified authentic users, performance of AOP attack detection is not satisfactory.

Bryan et al. [5] observed that the task of identifying attack\k profiles is information system is similar to the task of identifying bi-clusters in micro array expression data. Variance adjusted score was used to find the anomalous profiles which are correlated across the subset of items. They conducted an extensive evaluation on this metric performed well in separating attack profiles from genuine profiles for most of the attack strategies.

Hurley et al. [6] proposed to use Neuman-Pearson statistical attack detection to identify attack profiles. They developed a statistical model of standard attacks and introduced a new strategy to obfuscate average attack.

This paper describes an attribute based K-means clustering approach to identify attack profiles regardless of attack types. This approach involves the clustering of neighborhoods of two clusters, where user profiles in smaller cluster have been given low preferences while generating recommendation, therefore be less likely to influence prediction behavior. This approach assumes that normal and anomalous profiles from different clusters in the feature space.

2. ATTACK TYPES

We will focus on following attack types:

2.1 Standard Attacks

Profile injection attacks can be categorized based on the knowledge required by the attacker to mount the attack, the intent of a particular attack, and the size of the attack. The attack types are characterized by how they identify the selected items and what proportion of the remaining items they choose as filler items, and how the rating are assigned to items. All attack types include a target item which the attacker wants recommended more highly or wants prevented from being recommended.

Random Attack: This attack generates profiles in which the items and their ratings are chosen randomly based on the overall distribution of user ratings in the database, except for the target item. This attack is very simple to implement, but it has limited effectiveness.

Average Attack: In the average attack, each assigned rating for a filler item corresponds to the mean rating for that item, across the users in the database who have rated it. This attack requires knowledge about the system.

Segment Attack: An attacker may be interested in a particular set of users-likely buyers of the product. A segment attack attempts to target a specific group of users who may already be predisposed towards the target item. Increased recommendation of the target item to these users may be just as effective as one that raises the recommendation rate across all the users.

2.2 Obfuscated Attacks

Attacks that closely follow one of the attack types mentioned above can be detected and their impact can be significantly reduced. As a result, an attacker would need to deviate from these known types to avoid detection.

Noise Injection: it involves adding a noise to ratings according to a standard normal distribution multiplied by a constant, which governs the degree of noise to be added. This can be used to blur the profile signatures that are often associated with known attack types.

User Shifting: It involves incrementing or decrementing all ratings for a subset of items per attack profile by a constant amount in order to reduce the similarity between attack users. Shifts can take the positive or negative form, where the amount of each shift for each profile is governed by a standard normally distributed random number.

Target Shifting: For a push attack, it is simply shifting the rating given to the target item from the maximum rating to a rating one step lower, or increase the target rating to one step above the lowest rating.

3. ATTACK DETECTION VIA CLUSTERING

3.1 Detection Attributes

For the detection algorithm's dataset, the number of generic attributes and a residue-based attributes are used to capture the distribution differences. The various attributes used are:

Weighted Degree of Agreement: It is introduced to capture the sum of differences of the profile's rating form the item's average rating divided by the item's rating frequency.

Weighted Deviation from Mean Agreement: it is designed to help identify anomalies, places a high weight on rating deviations for sparse items.

Length Variance: it is introduced to capture how much the length of a given profile varies from the average length in database. If there are a larger number of possible items, it is unlikely that very large profiles come from real users, who would have to enter them all manually, as opposed soft-bot implementing a profile injection attack.

Residue Based Metrics: Residue based metrics have their origins within bioinformatics, particularly the gene expression analysis domain. It is used in an attempt to better model the gene functional modules within the expression data, that correlate over subset of experimental conditions.

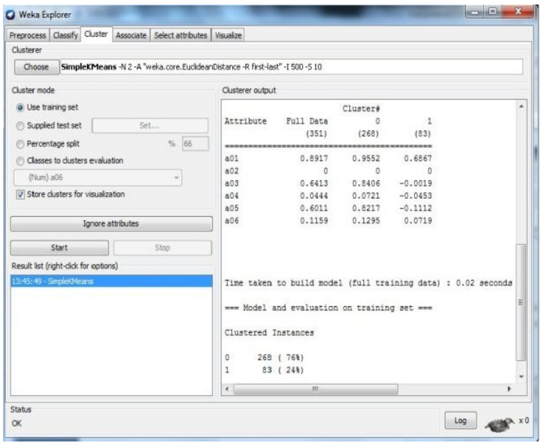
3.2 Identifying Anomalous Clusters

Attack profiles tend to be highly correlated, which is a result of the colluded nature of shilling attacks. It is assumed that the attack profiles are smaller in number and dominate one cluster due to their similarity. To identify an attack, the profiles for every user in the database are created. The representation of profiles consists of feature based on the detection attributes described above. The profiles are then partitioned into two groups of similar users. Assuming that the smaller cluster typically corresponds to attack profiles, the smaller cluster is marked as "anomalous" and gives low preference to all the profiles in this cluster when generating recommendations.

4. CLUSTERING PERFORMANCE

In this section we analyze how well our clustering algorithm build on the attributes described above performs on detecting profile attacks. The aim is to correctly identify the anomalous cluster with fewer genuine user profiles. It is observed that the detection performance using K-means is as good as UnRAP algorithm for all other push attacks except segment attack. The first reason for taking this algorithm is that the UnRAP algorithm could not detect any attack profiles for segment attack. So this is the reason why this approach is taken in designing segment attack. A typical segment attack profile consists of a number of selected items that are likely to be favored by the target user segment, in addition to the random filler items. Selected items are assigned to the maximum rating value along with the target item and the random filler items are assigned to the minimum rating. So user's mean rating becomes low as compared to the other attacks.

The second reason is the, in case of the segment attack each user's rating deviation from their mean rating becomes very low due to the low rating of the filler items, compared to the other attacks where the rating for the filler items are high.



Clusterer: SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Cluster mode:
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) a06
☑ Store clusters for visualization

Result list (right-click for options): 13:45:49 - SimpleKMeans

Clusterer output:

```
                        Cluster#
Attribute    Full Data        0         1
                (351)     (268)      (83)
==========================================
a01             0.8917    0.9552    0.6867
a02                  0         0         0
a03             0.6413    0.8406   -0.0019
a04             0.0444    0.0721   -0.0453
a05             0.6011    0.8217   -0.1112
a06             0.1159    0.1295    0.0719



Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      268 ( 76%)
1       83 ( 24%)
```

Status: OK

Fig 1: Clustering usinh K-means Algorithm

As our results shows the number of real users detected as an attack is very low in UnRAP algorithm so the specificity is higher than the k-means algorithm.

In fig1, the two clusters are formed and everything is forced into these clusters and can potentially result in clusters that are non cohesive.

The dataset consist of the 100,000 ratings and all ratings are integer values between one and five where one is the lowest and five is the highest. Our data includes all users who have rated atleast 20 items. The set of attacked items consist of 50 items whose rating distribution matches the overall rating distribution of all the sets.

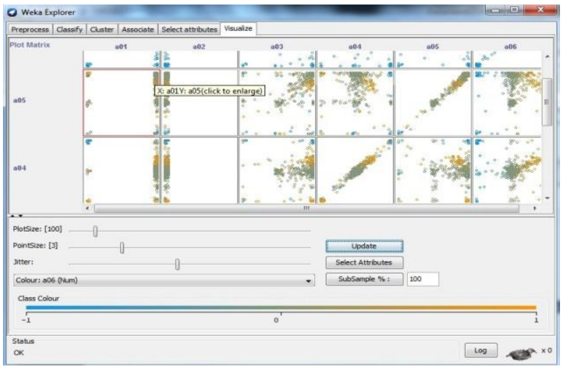


Fig 2: Rating Distribution Using K-Means Clustering

5. CLUSTERING EVALUATION

To evaluate the obfuscation method first we apply noise injection, user shifting and target shifting approach to average and random attacks.

Considering two clusters only, everything is forced into these clusters and can potentially result in clusters that are not cohesive. Spreading user profiles into more than two clusters may reduce the number of misclassified real user profiles.

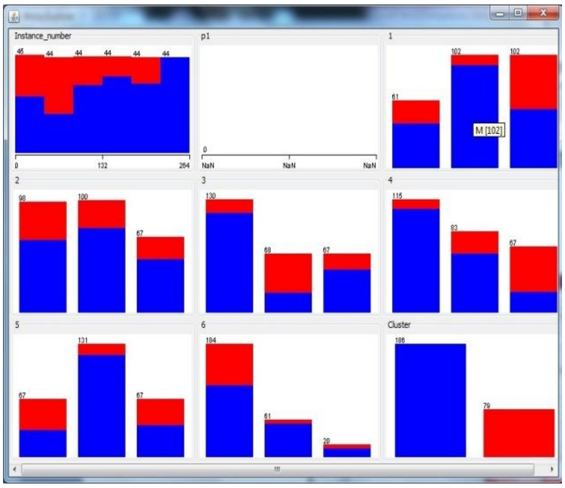
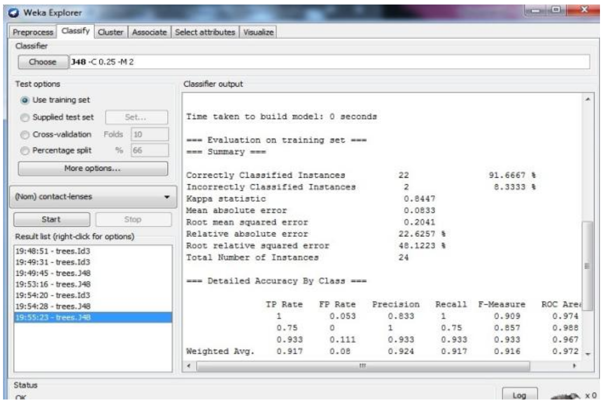


Fig 3: Classified Clusters

Overall our experimental results showed that the attribute based K-Means clustering approach can be a good detection technique regardless of attacks strategies. The detection performance of the attack is significantly better than the other approaches. It is also observed that by dividing the user profiles into different clusters, the attack profiles are always in one or two clusters of small size. So we need only one or two clusters to mark as anomalous and disregard while generating predictions.



```
Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          22               91.6667 %
Incorrectly Classified Instances         2                8.3333 %
Kappa statistic                          0.8447
Mean absolute error                      0.0833
Root mean squared error                  0.2041
Relative absolute error                 22.6257 %
Root relative squared error             48.1223 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Are
                 1         0.053      0.833     1         0.909      0.974
                 0.75      0          1         0.75      0.857      0.988
                 0.933     0.111      0.933     0.933     0.933      0.967
Weighted Avg.    0.917     0.08       0.924     0.917     0.916      0.972
```

Fig 4: Correctly Clustered Instances

6. CONCLUSION

The issue of security and robustness in information system is a major concern. In this paper we presented an unsupervised anomaly detection algorithm using K-Means clustering for detecting shilling attacks. Our results showed that segment attack which is designed to target a specific group of likely buyers can be easily detected with high accuracy using K-Means clustering. It is also proved that unsupervised clustering may achieve reasonably good performance against the attack types discussed above.

7. REFERENCES

[1] M. O'Mahony , N. Hurley, N. Kushmerick, and G. Silvestre, "Collaborative Recommendation: A Robust Analysis," ACM Transactions on Internet Technology, Vol 4, No. 4, pp-344-377, 2004.

[2] S. Lam and J. Reidl, "Shilling recommender systems for fun and profit," in Proceedings of the 13th International WWW Conference, New York, May 2004

[3] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams, "Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness," ACM Transactions on Internet Technology(TOIT), Volume 7 , Issue 4 (October 2007), 2007.

[4] B. Mehta, "Unsupervised shilling detection for collaborative filtering,"AAAI, 1402-1407, 2007.Sannella.

[5] M. O. K. Bryan and P. Cunningham, "Unsupervised retrieval of attack profiles in collaborative recommender systems," in *Technical Report, University College Dublin*, 2008.

[6] N. Hurley, Z. Cheng, and M. Zhang, "Statistical attack detection," *Proceedings of the third ACM conference on Recommender systems*, 2009.