Detection of Shilling Attacks in Collaborative Recommender Systems

PROJECT REPORT

Submitted by

SIBY RAJU

REG NO : TKM20CSCE15

In partial fulfillment for the award of the degree of

MASTER OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Under the guidance of
Prof. Jini Raju

Thangal Kunju Musaliar College of Engineering
Kerala

SEPTEMBER 2022

Thangal Kunju Musaliar College of Engineering
Dept. of Computer Science & Engineering

C E R T I F I C A T E

This is to certify that this report titled ***Detection of Shilling Attacks in Collaborative Recommender Systems*** is a bonafide record of the **project** presented by **SIBY RAJU (TKM20CSCE15)** under our guidance and supervision in partial fulfillment of the requirements for the award of the degree, **M.Tech in Computer Science & Engineering** in **APJ Abdul Kalam Technological University** .

Coordinator	Supervisor	Head of the Department
Dr. Ansamma John	Prof. Jini Raju	Dr. Dimple A Shajahan
Professor	Assistant Professor	Associate Professor
Dept. of CSE	Dept. of CSE	Dept. of CSE
TKMCE	TKMCE	TKMCE

ACKNOWLEDGEMENT

A successful project is a fruitful culmination of efforts by many people, some directly involved and some others indirectly, by providing support and encouragement. Firstly I would like to thank the almighty for giving me the wisdom and grace for making my project a memorable one. I thank him for steering me to the shore of fulfillment under his protective wings.

I express my sincere gratitude to **Dr. T A Shahul Hameed**, Principal, T.K.M College of Engineering for giving me an opportunity to present my project. I would like to thank **Dr. Dimple A Shajahan**, Professor and Head of the Department, CSE, TKMCE, for her constant support and encouragement throughout the work.

With a profound sense of gratitude, I would like to express my heartfelt thanks to my guide **Prof. Jini Raju**, Assistant Professor, CSE, TKMCE, and **Dr. Ansamma John**, Professor, CSE, TKMCE, for their expert guidance, cooperation and immense encouragement. I also extend my thanks to the entire faculty members and staffs of the Department of Computer Science and Engineering, TKMCE, who has encouraged me throughout this work.

I also express my thanks to my loving parents, brother and friends, for their support and encouragement in the successful completion of this project work.

SIBY RAJU

Abstract

Recommender systems are backbone for ecommerce website today. With the rapid growth in the ecommerce web sites, recommendation systems plays very important role to provide personalized service to the user. A good recommender system determines the quality of service provided by ecommerce. Ecommerce websites like amazon.com and ebay.com are widely popular because of their recommender systems. Collaborative filtering is one of the most widely used recommendation system. Recommendations made using collaborative filtering depend on relationship between the user and items. Unfortunately, due to its openness and dependency on user ratings, Collaborative filtering is prone to shilling attack and problem is with their security. Such attacks alter the recommendation process to promote or demote a particular product. Attacker who cannot be separated with distinguished from genuine user may inject biased profiles in the system to affect the service of system. It may leads to degradation of recommender system's objective. It is therefore essential to detect the shilling attacks in such a way that there are in-depth analysis of user behaviors and uses two key mechanisms (i.e., behavior features extraction and detection) to distinguish shilling profiles from genuine ones. In the stage of detection, a classifier is then built to distinguish attack profiles from genuine user profiles by constructing training data from authentic profiles and attacks generated by attack models. The combined effectiveness of this approach is then evaluated with the supervised classification algorithm Support Vector Machine. The experimental results demonstrate that proposed supervised detection model achieve a better detection performance is about 85.41% in shilling detection.

Contents

List of Figures

List of Tables

List of Abbreviations

CF ..... Collaborative Filtering
SVM ..... Support Vector Machine
KNN ..... K-Nearest Neighbour
PCA ..... Principle Component Analysis

Chapter 1

Introduction

From the beginning of internet services, the information on the internet happens to be increasing at quick rate. Retrieval of of good use information is a significant dilemma. A recommender system[1] is just a system that shows items towards the individual by which he could suggests items to the user in which he may be interested. It is often needed for individuals to check with other people and have about their previous experiences whenever making alternatives. To recommend a product to user a recommender system centers around the similarity regarding the individual along with other users into the system and description of item. A recommender system is quite very theraputic for a person who may have lack of knowledge for finding appropriate item on the net. In social recommendation behavior, individuals share their with one another and determine whether or otherwise not to behave on such basis as whatever they hear from other people.

Utilizing the quick growth in the ecommerce web sites, recommendation systems plays extremely important role to offer service that is personalized the consumer. A recommender system that determines the quality of service supplied by ecommerce. Ecommerce internet sites like amazon.com and ebay.com are widely popular due to their recommender systems. Collaborative filtering[2] is one of the most widely used recommendation system. But there is however a problem with utilizing the publicly accessible collaborative systems, problem is with their safety. Attacker whom cannot be separated with distinguished from genuine individual may inject biased profiles in the system to influence the solution of system. It might leads to degradation of recommender system's objective and accuracy of the system device. As formerly stated also a small quantity of assault can degrade the accuracy regarding the system. Collaborative recommender system is very susceptible towards the assault because it is situated regarding the feedback from individual, so an attacker can directly influence the ratings of the system. Outcomes of a collaborative recommender system very influenced by

the truthful feedback of its users. Recommendations are determined on the basis of the genuine feedback will be of good quality. As in collaborative recommender system identities of other users is not understood, it is therefore feasible to introduce the fake individual profiles identical towards the genuine users contained in the system. In this method attacker make the device create recommendation behavior which he desires. For these type of weaknesses shilling attack term is used. Recent studies has revealed that the behavior of all regarding the recommendation algorithms can be manipulated by even modest attacks.

Collaborative filtering systems will be the most usual variety of internet type that is common of personalization system and tend to be very susceptible to attacks. Attackers inject a huge amount of biased profiles in to the system, leading to system recommendation that favor or disfavor a offered item. Shilling attack term is generally utilized for such means of attacks. Promoting and demoting a certain item is the actual only real cause for making use of such attacks. Within the shilling attacks, attacker with make an effort to bias the system's outcomes interacts with using the collaborative recommender system to create a large numbers of fake profiles identical with the actual users in the system. Inserting fake profiles in the system can be carried out by lots of aspects, attacker can insert profiles either manually or they can develop a automatic means to place profiles. In the event that system is simple in term of users and ratings they incorporate the manually insertion approach. Because manually inserting profiles in the system simply take a lot of commitment. In simple system yet few fake profiles can promote and demote a particular item. But, if the system is large containing lots of users and billions ratings, in this condition few fake users usually do not cause much impact in the system. In this setting, he must acquire the automation tools to embed the biased profiles in the system generating recommendations that favors or disfavors a certain item. A collaborative recommendation system must certainly be available to user inputs, so it is therefore extremely tough to develop a collaborative recommender system that may not be attacked. Therefore it is mostly concentrated on the defending against such attacks. A number of studies have indicated that collaborative recommender systems are prone to shilling attacks which can be illustrated by an example.

Consider a recommender system that identifies movies that user might like to watch using user based collaborative approach. A user profile in the

Table 1.1: An Example of a Shilling Attack for Promoting Movie6.

users	item1	item2	item3	item4	item5	item6	correlation
Siby	5	2	3	3		?	
u1	2		4		4	1	-1.00
u2	3	1	3		4	1	0.76
u3	4	2	3	1		1	0.72
u4	3	3	2	1	3	1	0.21
u5		3		1	2		-1.00
u6	4	3		3	3	2	0.94
u7		5		1	5	1	-1.00
a1	5		3		2	5	1.00
a2	5	1	4		2	5	0.89
a3	5	2	2	2		5	0.93
Correlation with item6	0.85	-0.55	0.00	0.48	-0.59		

operational system consists of user's ratings in the scale of 1 to 5 with 1 as the lowest and 5 as the highest rating on the various movies. In the visits that are previous has built his profile and he return to the system for new recommendations. It consists of eight users that are genuine including Siby. It also contains 3 attack profiles (Attack 1-3) inserted by the attacker. All the attack profiles have offered ratings to the item 6. Assume that our system is making use of user based collaborative filtering approach. Prediction rating can be obtained by finding the closest neighbor for Siby on item 6 can be obtained by finding the neighbor that is closest to Siby. In this case, prediction related to item 6 for Siby would be 2. Or simply it can state that item 6 is will be disliked by Siby. Now consider the scenario that our system is assaulted, now attack 1 profile would be the most similar one that is similar towards Siby and also it would yield a predicted rating of 5 of a rating that is predicted for item 6 and it is entirely opposing from what she has predicted without the attack. So that in this instance, assault is successful and Siby are certain to get item 6 as being a recommendation, it generally does not matter whether this is actually the most effective recommendation for him. Now assume if the system is using item based collaborative filtering approach. The prediction rating for item 6 would be calculated by comparing the rating vector of item 6 with those of other items. This method does not lend itself to an assault as previous one because attacker doesn't have control over the other users for their ratings to a specific given item. If attacker obtains some knowledge of rating distributions then this can make an effective assault . In

the example of Table 1.1, attacker knows that item 1 is a popular item among a significant group of users to which Siby also belongs. Designing an attack profile in a way that high rating is related to both items 1 and 6. Attacker can try to boost the similarity of those two items resulting in producing a higher possibility that Siby will get item 6 as being recommendation.

The main contribution in this project is just a description of a procedure for detecting shilling attacks with supervised classification. The method is based on determining attributes of profiles by a user behavior analysis. This is certainly achieved by way of a three pronged strategy to creating characteristics to facilitate attack classification. In the stage of detection, a classifier is then built to distinguish attack profiles from genuine user profiles by constructing training data from authentic profiles and attacks generated by attack models. The combined effectiveness of this approach is then evaluated with the supervised classification algorithm Support Vector Machine (SVM).

Chapter 2

Background

2.1 Attack Models

A shilling attack against a recommender system is made up of set of profiles inserted towards the system by the attacker.

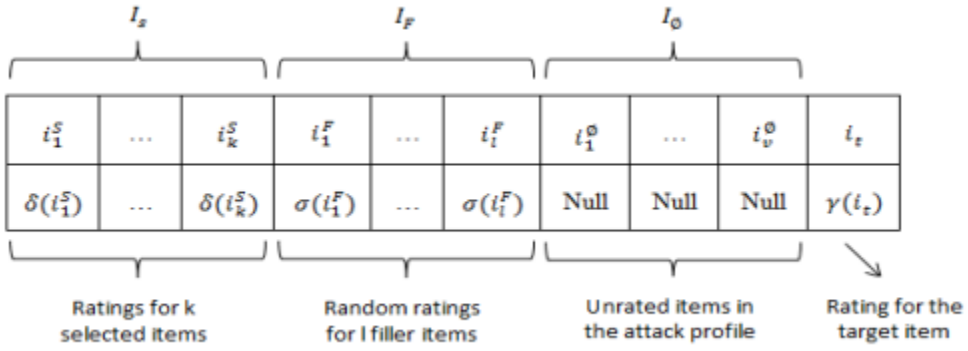


Figure 2.1: General structure of a profile in a shilling attack.

As depicted in Figure 2.1, items are divided in to four groups and each profile comprise each category[3].

i. **Target item**(i_t): There will likely be exactly one target product in the items set.

ii. **Selected items**(i_s): It is a collection of chosen products that represents a group that is small of items which have been selected because of their relationship utilizing the target item. For a number of attacks this set may be empty. These are usually chosen randomly.

iii. **Filler item**(i_F): The group of filler items represents a combined group

of arbitrarily chosen items in the database that are assigned ratings based on the proportion associated with assault. Size of the selected item set is small, so size of each and every profile is set by the dimensions of filler item set.

iv. **Unrated item set(i_ϕ)**: This is actually the group of items that maybe not rated by attack profile. A shilling attack against a collaborative filtering system consists a significant number of attack profiles of same kind based on same attack model is included with the database of real individual profiles. The motive behind is always to either increase (in case of push attack) or decrease (in case there is nuke attack) the system's predicted rating for a target product for an offered user.

2.1.1 Characterising Attacks

A shilling attack on a recommender system consists of the attacker injecting a set of profiles into the system. A profile is made up of a set of rating/item pairs with a rating value for each item and a null value for unrated items. For the attacks, there will always be a target item that the attacker wants to promote or demote. In most cases, there will also be a set of filler items chosen at random from the available options. This set will be denoted by the letter I_F. Some attack models also make use of a set of items drawn from a database. The small set is usually associated with the target item (or a targeted segment of users).

Online rating systems are crucial in recommender systems. In reality, collaborative filtering recommender systems are extremely vulnerable to "shilling" attacks. Shilling attacks can be classified as either a push or a nuke attack, depending on whether a product is promoted or demoted to gain an economic advantage over competitors. Several times over the years, Attack profiles and models have been developed. Many detections occur concurrently. To counter such attacks, techniques and algorithms have emerged. Almost all of them while generating malicious users, the attack models use the same attack profile.

2.1.2 Standard Attacks

Standard attacks are attack models that do not focus solely on evading detection in a recommender system. Many detection algorithms are more likely

to detect the shilling attack profiles injected by these attacks.

Random Attack

In random attack model, average rating of system is assigned to the filler items and pre specified ratings are assigned towards the target item e.g. rating 5 is given to the goal item in the full instance of push attack and rating 1 is given towards the target product into the instance of nuke attack. In random attack model illustrated in Figure 2.2, group of chosen products is empty. Knowledge needed for this attack is minimal because the overall mean associated with the system can be simply measured by outsider. However some expense of execution is included, but it is really not that much because this assault involves assigning mean rating to every filler product. The Random Attack, is the most fundamental type of shilling attack. Except for the target item, the items rated by the attack profile are randomly selected in this model. The ratings for these items are close to the system mean. Depending on whether it is a push or a nuke attack, the target item receives a maximum or minimum rating. It is the simplest attack, but it is also the least effective. A random attack's goal is usually to disrupt the efficiency of a recommender system rather than to promote the target item. Random attacks are simple to execute because they require little knowledge. All the attacker requires is the overall system mean, which can be easily calculated empirically. It is not particularly effective because it is the simplest attack. In terms of item selection, the average attack is similar to the random attack. The randomly selected items are rated using the individual item rating distribution. The mean rating of each filler item is assigned. This attack is only possible if the attacker has extensive knowledge of the dataset on which the recommender system is based. This model's effectiveness is proportional to the attacker's knowledge. The only difference between a random attack and an average attack is the filler ratings; the average attack is far more effective.

Average Attack

Each filler item is assigned score i.e. mean rating for that item throughout the users in the database who have rated it. Similar to attack that is random, set of filler products is opted for randomly and set of selected items ie.I_s can be combined into the set of unrated items i.e. I_ϕ and I_s is not rated in this attack. This attack can additionally be utilized as nuke attack. The basic distinction

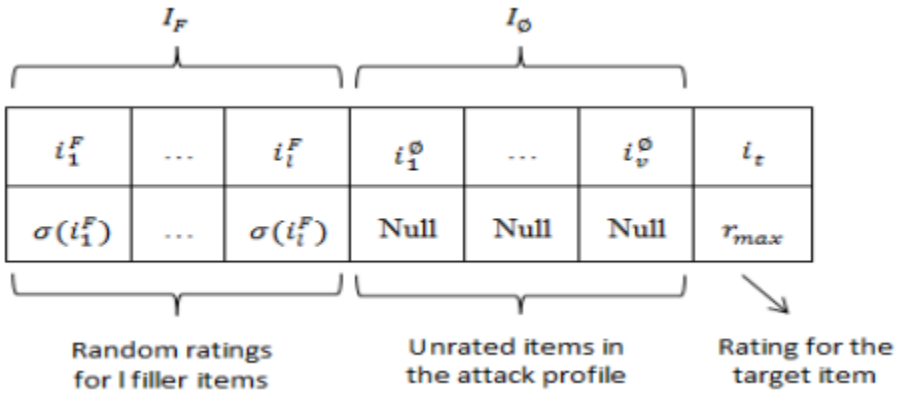


Figure 2.2: An attack profile based on random or average attack model.

between random and average attack is just how in which ratings are assigned towards the filler products in the profile. Figure 2.2 shows the general as a type of average attack. Some effort can be taking part in associated with this attack for producing the ratings. Average attack has knowledge expense of order i.e. quantity of filler items products in the assault profile. This attack is prosperous in user based collaborative technique whenever size of filler item set is small. Thus the knowledge cost included in this attack may be substantially reduced. Nevertheless this assault model is not effective against item based collaborative algorithms.

Bandwagon Attack

Bandwagon Attack is a type of attack in which attackers' profiles are filled with popular items with high ratings. The attack profiles are, by definition, closer to a large number of users. The highest rating is given to the target item. Bandwagon attacks are also low-knowledge attacks because the attacker only requires publicly available data. This attack could be seen as the expansion of random attack, where set of chosen products are additionally included in assault profiles. It can be seen in term of both ratings assigned to them by genuine users and because these items are generally liked. For example these popular items could be a Box-office hit movie or even a best-selling book. These forms of items have actually huge probability of being rated by large numbers of genuine profiles.

In the attack profiles, these chosen items get high ratings to make sure high. These selected items are given high ratings to ensure high amount of similarity between assault and genuine profiles. This assault model is known as low knowledge cost. It does not require any system particular data because it is not extremely tough to determine the group of most popular items in any item space. Figure 2.3 shows the attack profile for the bandwagon attack. Items set is selected because this group of items has been rated by a large numbers of users into the database and items into the set are offered maximum rating values together with the target product. Ratings for the group of products is set in the similar way in random attack.

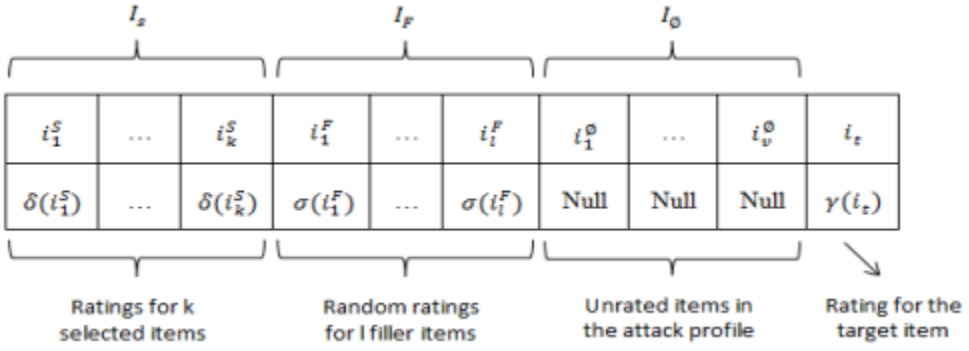


Figure 2.3: A bandwagon attack profile.

Chapter 3

Related Works

The vulnerabilities of collaborative recommender systems to shilling assaults have actually led to a wide range of studies concentrating on detecting shilling assaults over the past decade.

3.1 Supervised Approaches

The shilling attack problem was treated as being a category problem by Chirita et.al. [4], utilized the function etrics for detecting harmful profiles. The method had been developed to detect random and bandwagon assaults. Burke et al. [5] to improve the classifier's performance. KNN, and C4.5 were the most commonly utilized classifiers for the detection of fake injected profiles. The issue with using the attributes that are generic that numerous authentic users who had extreme behaviors were misclassified as shilling profiles. To overcome this problem, along with to improve the precision regarding the classifications, attack attributes that are specific developed by [6]. Different assault specific attributes were formed for average, and bandwagon attacks.

Williams et.al. [7] utilized three strategies to increase the precision of detection in the approaches that are supervised similarity to reverse-engineered assaults, target concentration, and rating anomaly detection. This detection strategy is effective due to the added robustness to system, however it is highly reliant regarding the classifier's choice. Their study shows that combining different attributes improves the classifier's performance, particularly the support vector machine, and significantly decreases the effect of the most extremely powerful assault models. The features found in their method are was incorporated in [8] to diminish the perturbation brought on by the imbalance. The writers first ease the hard category task by making use of well designed features for an individual profiles. Hao et.al. [9] employed an ensemble detection method on features extracted from ratings, item appeal,

and user-user graph. The function removal is carried out by making use of PCA. It automatically extracts individual features with different corruption rates. It utilized a three-stage process involving information preprocessing, feature extraction, and detection using poor classifiers.

3.2 Unsupervised Approaches

Unlike supervised or semi-supervised detection methods, unsupervised approaches do not require training classifier on labeled profiles. The initial unsupervised approach introduced by Mehta et.al. [10] applied Principle Component Analysis to the profile detection problem. Four facets led to generated this issue being appropriate PCA: spam users are highly correlated, low deviation from mean rating value, a high similarity by having a wide range of users, and the assumption that spam users work together. All the consumer profiles in the recommender system had been projected onto a hyperplane formed from the user-item matrix. An individual profiles which were clustered closer to the beginning of the hyperplane were the attack profiles. The sparsity of the user-item it is harder for these predictions to be dependable. Bryan et.al. [11] formulated a generic characteristic aiding in the detection of assault profiles within an unsupervised manner. Their approach treats the attack profiles detection problem being an anomalous framework detection problem. The metric used is just a variation regarding the Hv-score metric which was initially used in gene information analysis to aid in locating biclusters. This algorithm, called the UnRAP, appears useful in detecting both standard and obfuscated attacks. Their approach has better possibilities of catching assault methods which will escape supervised methods. Yang et.al. [12] proposed a unsupervised detection method, which detected the shilling profiles by using the similarity of topological structure of attack users in the graph. Recently, Yang et.al. [13] exploited a density-based clustering method to detect suspicious users, which clustered suspected users based on features extracted from item distribution and detected suspicious users using target item analysis. Chai et.al.[14] detected an unsupervised approach for detecting shilling profiles, which does not need to know the attack size or to label the candidate spammers. In that case, many genuine profiles are misclassified as attack ones.

Chapter 4

Proposed System

One of the most significant strengths of collaborative recommender systems is the ability for users with uncommon preferences to get meaningful recommendations by the system users that are identifying comparable peculiarities. This strength normally one of many challenges in securing recommender systems. Specifically the variability of opinion helps it be hard to state with certainty whether a specific profile is an attack profile. It is impractical to expect all profiles to be categorized correctly. The objectives for detection will consequently be: minimize the impact of attack, reduce steadily the chance of a effective assault, and reduce any negative impact resulting from the addition of the detection scheme. The approach to attack detection will consequently focus on recognizing attacks based on these reverse engineered attack models. An ideal outcome would be one by which a system could possibly be rendered secure by simply making attacks against it no longer cost effective, where expense is calculated in the attacker's knowledge, effort, and time. There are other techniques have been studied for defending against shilling attacks as well, however this project mainly focus on a profile classification approach. To increase the performance of shilling detection, a novel behavior-based supervised classification based approach for detecting shilling profiles. As depicted in Figure 4.1 , first extract detection features to characterize the behavior differences of genuine and shilling profiles, and then distinguish shilling profiles from genuine people by presenting a classification based shilling detection method.

4.1 Feature Extraction

Human individual behavior is just a complex procedure included with intention. The indepth understanding of human habits is truly important in just about all human-related application scenarios, which will help to describe a large amount of socio-economic phenomena. In recommender systems, the

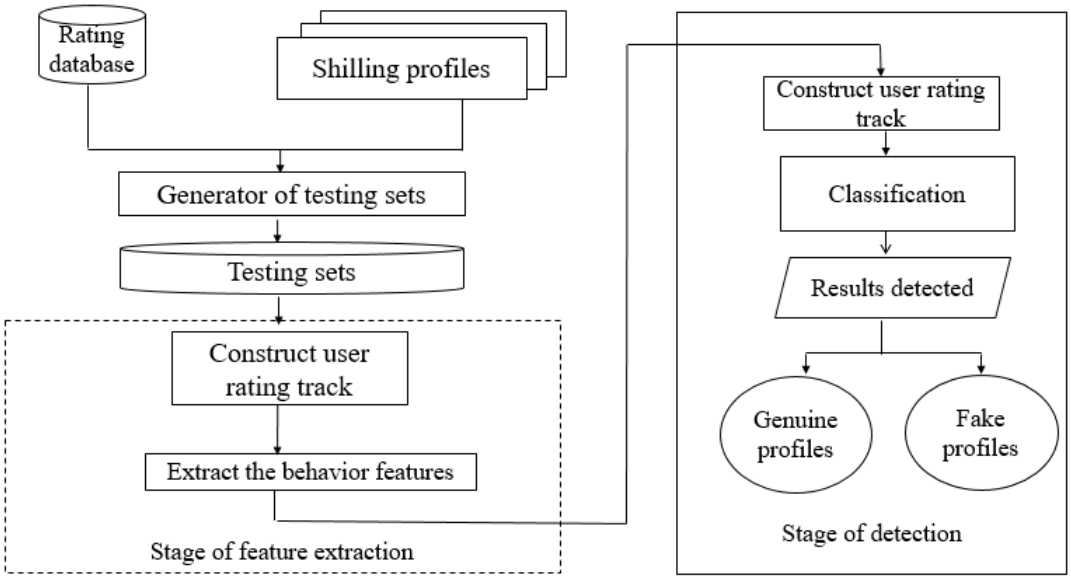


Figure 4.1: Framework of Supervised Approach

behaviors of genuine users usually are driven by their interests, thus development of interest behaviors has the characteristics of long-range correlation. By contrast, the attack users usually rate filler items at random from those items that are non-target with lack of real consumption motivation, which results in a significant difference in many aspects contrasted with genuine users. Accordingly, it measure these characteristics and extract behavior features which could reflect apparent distinction between genuine and shilling profiles. In recommender frameworks, the ways of behaving of authentic users are generally determined by their interests, consequently development of interest ways of behaving has the attributes of long-range correlation. By contrast, the attack users rate filler items from those non-target items for lack of real consumption motivation, which prompts a tremendous distinction in numerous perspectives contrasted with genuine users. Likewise, measure these attributes and extract three features which can reflect clear contrast between genuine and shilling profiles.

4.1.1 Measurement of Long-Range Correlation on Interest

Individual interests are shown by the rated items in recommender systems, therefore coherence between interests is concordant with the relationship involving the corresponding rated items. In this section, Jaccard similarity coefficient[14] is employed to determine the relationship between products, then track correlative degree is used to describe the correlation that is long-range interest.

Definition 1 (User rating track).

For any user $u_k \in U$, user rating track of u_k is an order sequence of I_{uk}, which is sorted based on the rating time of items rated by uk in ascending order and shown in Equation 4.1.

$$Track_k = p_{v1}, p_{v2}, ...p_{v_s} \tag{4.1}$$

where s is the length of the order sequence s.t. s = $\mid Iuk \mid$ and $p_{vi}(1 \leq i \leq s)$ is the ith item on $Track_k$ s.t. $t_{k,v1} \leq t_{k,v2}$....

Definition 2 (Correlation coefficient between items).

Jaccard similarity coefficient is utilized to compute the value of user rated items that is pertinent to the quantity of users who have co-evaluated the items.

Definition 3 (Track correlation degree).

$$Track_{corr_k} = \frac{1}{|Track_k|} \sum_{l=1}^{|Track|} Jac_{vl,vl+1} \tag{4.2}$$

In Equation 4.2 , $\mid Track_k \mid$ is the cardinality of $Track_k$, $Jac_{vl,vl+1}$is the correlation coefficient between the l_{th} and $(l+1)_{th}$ item on $Track_k$.

4.1.2 Variety proportions of users interest

Variety of individual interest in fascination with recommender systems means the users varied tastes, suggested by the rated products. Various sources for determining variety, such as for example taxonomies of items, genres of products , and latent topics of items. In view of the idle subjects of things,

there are two new measurements (i.e., track similitude degree and interest vulnerability degree) to catch the variety of users interest by utilizing the relationships between's vectors of idle subjects.

Definition 4 (Track similitude degree).

For any user, u_k the trace similitude level of user gauges the normal interest comparability between things, which is characterized as in Equation 4.3:

$$Track_{sim_k} = \frac{\sum_{p_{vx}} \in Track_k \sum_{p_{vy}} \in Track_k, p_{vx} \neq p_{vy} f_o(pv_x, pv_y)}{|Track_k| \times (|Track| - 1)} \tag{4.3}$$

where $fo(p_{vx}, p_{vy})$ denotes the similarity between vectors of latent topic distributions corresponding to items p_{vx} and p_{vy} .

Definition 5 (Interest uncertainty degree).

$$Track_u nck = \frac{1}{q} \sum_{x=1}^{q} \sqrt{\frac{1}{|Track_k|} \sum_{v=1}^{|Track_u nck|} (dp_{v,x} - \mu_x)^2} \tag{4.4}$$

In Equation 4.4, q means the total number of latent topics, $Track_k$ denotes the number of items rated by user u_k, $dp_{v,x}$ means the probability that the v_{th} item on $Track_k$ belongs to the x_{th} topic, and μ_x is the mean of probabilities corresponding to the x_{th} topic. That is to say, the interest uncertainty degree of user u_k is calculated by the average of standard deviation of all latent topic distribution vectors corresponding to the items on $Track_k$.

4.2 Shilling Detection

An individual shilling profile has little impact on the target item while a group of shilling profiles can bring remarkable effect. To produce better attack effect, a group of shilling profiles may be injected into the rating database simultaneously, which manifests high similarity on their behavior features for the same intention. So, classify shilling profiles together due to their high behavior similarity

Attack profile classification

Since some knowledge of assaults are effective, attack detection as being a conventional pattern category issue by which classify profiles as known attack models. It might be that some genuine users will be categorized as attackers. Therefore, classification characteristics are constructed with characteristics associated to detecting traits of the particular attack model and target levels across profiles.

4.2.1 Classifier model

Using supervised learning methods for classification, user profiles presents some significant challenges. The exponential amount of combinations of attack types, feasible attack objectives, and filler items helps it be infeasible to enumerate a training set making use of the ratings profiles alone. Because of this some strategies must be employed to generalize the notion of an authentic or attack profile beyond the ratings information. To achieve this, detection characteristics are accustomed to capture analytical top features of a profile. In an attempt to train the classifier, a training set first needs to generally be prepared. This is accomplished by firmly taking a collection of profiles from the profile database, these profiles are suspected to get from non-malicious users and tend to be labeled authentic. Into this training information a mixture of attack types at different attack sizes and filler sizes is inserted and labeled attack. The detection characteristics are then created for every single rating profile, and just the detection characteristics and label of profile is kept within the training set. Training the classifier then follows supervised learning techniques. Another challenge that separates attack classification from conventional classification could be the competitive nature associated with an issue. In conventional classification dilemmas, it is always have the challenge of attempting to take into account information conditions into the unseen information that are not contained in training data. For attack classification, this issue is compounded by the exact fact that there is an adversary, the attacker strengths and therefore thought to definitely try to find approaches to simply take benefit of these conditions. Hence to allow a classification scheme for being robust against attacks; it does not just requires the detection features to remain adaptable sufficiently to capture deviations, moreover it ought a classification algorithm that is ro-

bust to harmful noise. To determine such type of classification algorithm, it is usually worthwhile considering conceptually exactly how classifications or the classification model is established and its vulnerabilities.

Conceptually a learning scheme that combined findings all over the the complete training set all together would better made than a technique according to an even more localized approach . Hence a SVM classifier is proposed and probably will be better made than other models since its classifier primarily includes all training examples simultaneously in analyzing certain profile. The SVM algorithm which illustrated in Algorithm 1 remains analyzed generally, in aspect due to its theoretical perspective and properties of its determination boundary. Particularly it mathematically finds optimal decision hyperplane utilizing the prominent margin per characteristic. This simply means for the adversarial classification dilemma is that most features are viewed as and weighted in a way that each of them can meaningfully influence the classification. Conceptually this method has the good function that it would definitely be more complicated for some kind of attacker to disguise the complete signature yet still need an effective attack.

Algorithm 1 Supervised Detection Algorithm

Input: SVM,$\mathrm{X}_{train}, X_{test}$
Output:X_{result}
1: Classifier $\leftarrow$) train SVM using X_{train}
2: $\mathrm{X}_{result} \leftarrow null$
3: $X_t \leftarrow (label_v)$ or v' is user v after feature extraction of X_{test}
4: for v $\leftarrow$ 1 to V do
5: $\mathrm{X}_{result} \leftarrow \mathrm{X}_{result} \bigcup Classifer(v')$
6: end for
7: return X_{result}

Chapter 5

Experimental Results and Discussion

Performance of collaborative recommender systems is based on the reliable feedback of its users. Through the instant increase on the internet gains shopping, news and plenty of more providers it become very essential that recommender system suggest right item to its users. In the previous several years great deal of research was done in the area of recommender systems. Recommender systems are remarkably susceptible to the attacks, in this type of instance people are interested not merely the performance of recommender system but also improvement in the performance of system soon after the attack.

5.1 Dataset

For this experiments, publicly available MovieLens 100K dataset[14] is utilized. It is usually utilized dataset for experiments on recommender system. This dataset contains 100,000 ratings by 943 users on 1682 movies. Ratings are given in integer type ranging between 1 to 5 shown in Figure 5.1 where rating 1 is least (dislike) and 5 will be the maximum (most liked) ratings. This dataset contains almost all the users who possess rated to at the very least 20 movies. Mean rating is 3.6 for all users in this dataset. To conduct the experiments , divide the dataset into two parts, training and testing set and kept 70% data for training and 30% data for testing purposes. For every attacks, should created a wide range of attack profiles and inserted into the system. There are around 1000 users in the system, therefore 10% attack size implies that 100 attack profiles included with the system. As there are actually 1682 unique items in the system therefore 1% filler size means one individual user gives ratings to the 168 items.

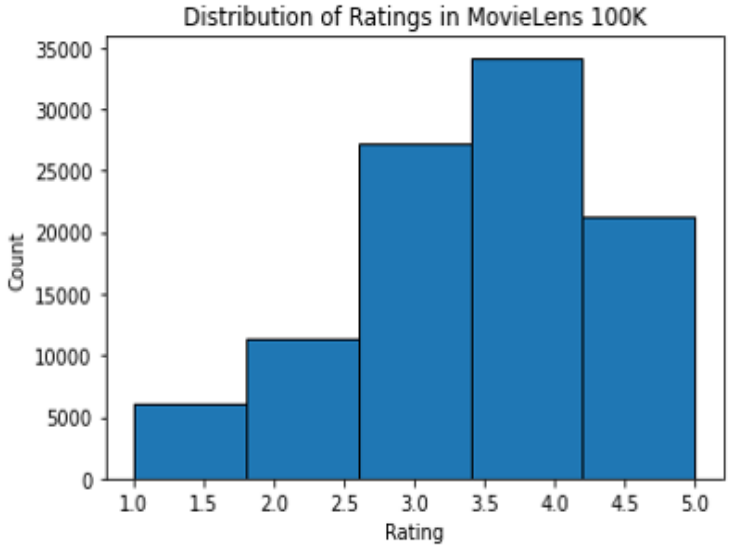


Figure 5.1: Distribution of Ratings in MovieLens 100K

5.2 Evaluation Metrics

There are considerable research in the area of recommender systems evaluation. Several of those concepts can additionally be applied to placed on the evaluation regarding the security of recommender systems. To compare various classification algorithms, people are interested mainly in measures of classification performance. An accurate classifier will avoid attack profiles from having a direct impact. One additional component that identified in previous research is mistakes induced by false positives. Most algorithms classify genuine profiles as attackers, thus possibly impacting the accuracy regarding the recommendations produced. Therefore it is essential to measure the impact of attack detection on recommendation accuracy.

$$Precision = \frac{TP}{TP + FN} \tag{5.1}$$

$$Recall = \frac{TP}{TN + FP} \tag{5.2}$$

where TP is the number of shilling profiles correctly detected and FP is the number of genuine ones misclassified as shilling profile. Thus true positives is the number of correctly classified attack profiles, false positives is the number

of authentic profiles misclassified as attack profiles, and false negatives is the number of attack profiles misclassified as authentic profiles. Thus sensitivity measures the proportion of attack profiles correctly identified, and specificity measures the proportion of authentic profiles correctly identified.

5.3 Experimental results and analysis

In this stage, collect the dataset for the attack methods. For this, load the dataset and then findout the average rating for each movie. This is needed because to get a baseline on how all the users have rated for each movie. Shilling profiles are generated and injected under various attack models, random attack, average attack and bandwagon attack on these movies. On performing these attack models, the first step is to set the attack size for that calculate the average rating, maximum rating and maximum users. Then set the target size and filler size that corresponds to the target items are selected randomly out of 1682 items. In random attack model, average rating of system is assigned to the filler items and ratings are assigned to the target item e.g. rating 5 is given to the target item in the case of push attack and rating 1 is given to the target item in the case of nuke attack. In random attack model, set of selected items is empty. Knowledge required for this attack is minimal because the overall mean of the system can be easily measured by outsider. Average attack model uses the individual mean for each item rather than the global mean except for the target item. Each filler item is assigned rating i.e. mean rating for that item across the users in the database who have rated it. Similar to random attack, set of filler items is chosen randomly. Bandwagon attack can be seen as the extension of random attack, where set of selected items are also included in attack profiles. These selected items are the small number of frequently rated items i.e. items that are popular in the domain. It can be viewed in term of both ratings assigned to them by genuine users and because these items are generally liked.

5.3.1 Feature extraction

In the stage of feature extraction, first construct rating track of each user and obtain latent topics of each item, then extract behavior features from the view of human behavior.

1. Track correlative degree

User rating track is an order sequence which is sorted according to the timestamp of items rated by in ascending order. Jaccard similarity coefficient is used to measure the association between items and it is relevant to the total number of users who have co-rated the items. The attack ones are generated by average attack, random attack and bandwagon attack models with different attack sizes and filler size. The values of track correlation degree for genuine users are larger than those of attack ones, indicating that there has obvious difference between genuine and attack users in track correlation degree.

Based on the latent topics of items, track similitude degree and interest uncertainty degree is used to capture the diversity of user interest by using the correlations between vectors of latent topics. By using LDA, hyperparameters α and η are set to 0.3 and 0.3 respectively and execute the model.

2. Track similitude degree

Track similitude degree may be calculated by cosine similarity. The cosine similarity measures the similarity by calculating the cosine angle between the two latent topic distributions.. If consider the cosine function, its value at 0 degrees is 1 and -1 at 180 degrees. This means for two overlapping vectors, the value of cosine will be maximum and minimum for two precisely opposite vectors. The attack ones are generated by the different attack models. The track similitude degree of attack users is lower than that of most genuine ones and concentrates on a small range, indicating that attack users' similitude differ from that of genuine users. The track similitude degree of attack users is lower than that of most genuine ones and concentrates on a small range, indicating that attack users' track similarity degree differ from that of genuine users.

3. Interest uncertainty degree

The interest uncertainty degree of user refers to the uncertainty of the user interest indicated by the rated items of user. It measures the average of stan-

dard deviations on all latent topics corresponding to the items on $Track_k$. For interest uncertainty degree of different users, higher value denotes more uncertainty or diversity of user interest. Here, the attack users are also generated by different attack model with different attack sizes and different filler sizes. The interest uncertainty degree of attack users is higher than that of most genuine users, indicating that the attack users' interests are more diverse.

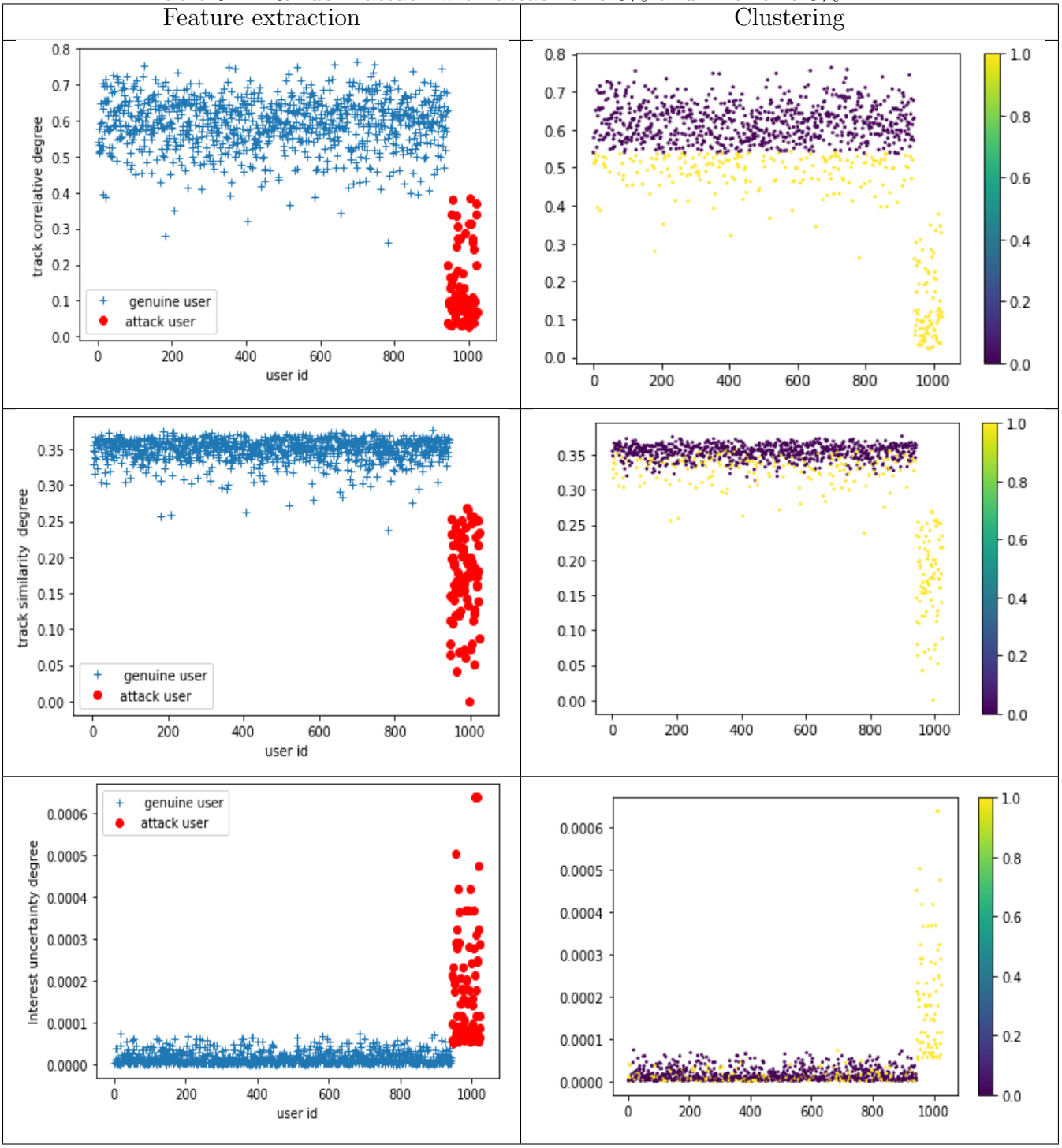
5.3.2 Unsupervised approach

The framework of unsupervised approach focused on behavior characteristics of users and present a k means clustering based unsupervised approach to detect shilling profiles. It start by the memory and inertial effect of human behaviors and extract three detection features reflecting behavior differences between genuine and shilling profiles, i.e., track correlation degree, track similitude degree and interest uncertainty degree. Then it construct the behavior feature space composed by all behavior feature vectors, and calculate the behavior similarly between vectors based on the euclidean distance. Finally detect the shilling profiles by using clustering algorithm in the behavior feature space, which is based on highly similar behaviors among shilling profiles. The features are extracted and detected by performing various attack models by random attack with attack size 5% and filler size 5% are illustrated in Table 5.1, average attack with attack size 15% and filler size 5% are illustrated in Table 5.2 and bandwagon attack with attack size 15% and filler size 5% are shown in Table 5.3. These methods are performed with an accuracy of 81.27%, 80.39% and 80% respectively.

5.3.3 Supervised Approach

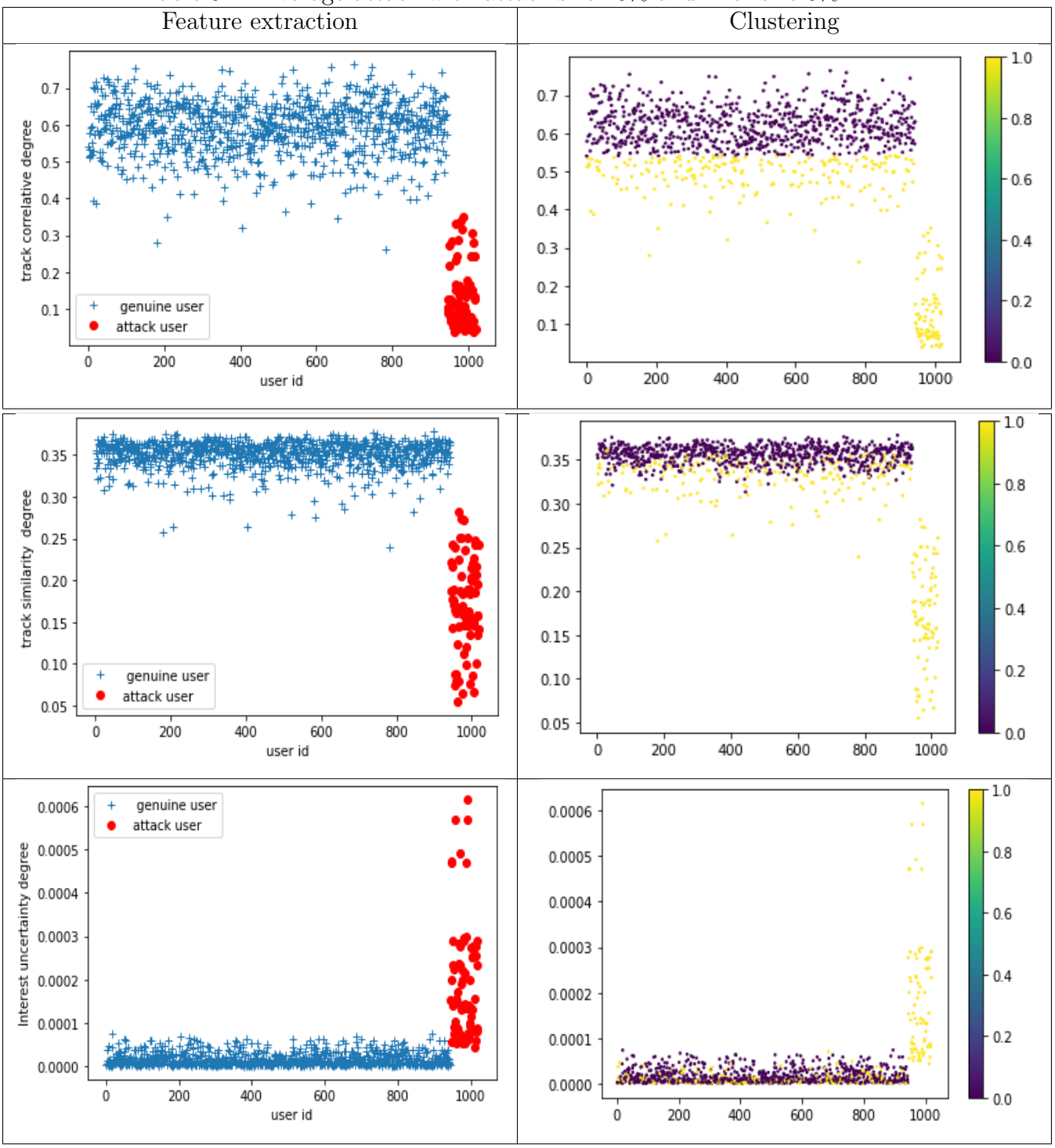
Classification starts from the selection of profile that needs to be created. Once the profile is selected,the detection features are extracted for the purpose of classification. The detection features such as track correlative degree, track similarity degree and interest uncertainity degree are detected by using bandwagon attack with an attack size of 15% and filler size of 10%. The extracted features are then fed to the trained classifier. In order to train the classifier, a training set first needs to be created. This is done by taking a set of profiles from the profile database, these profiles are assumed to be

Table 5.1: Random attack with attack size 5% and filler size 5%



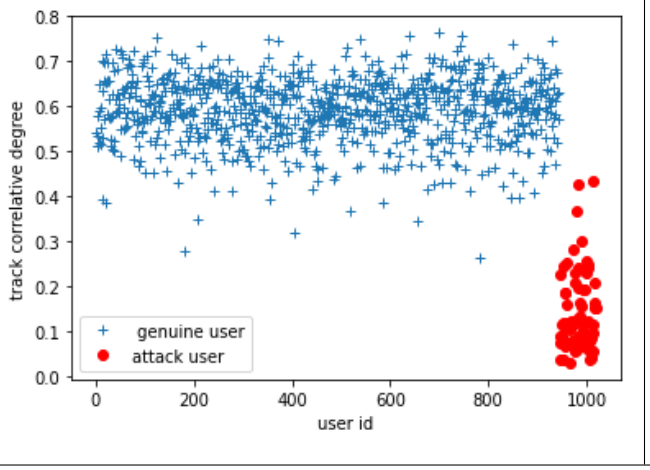
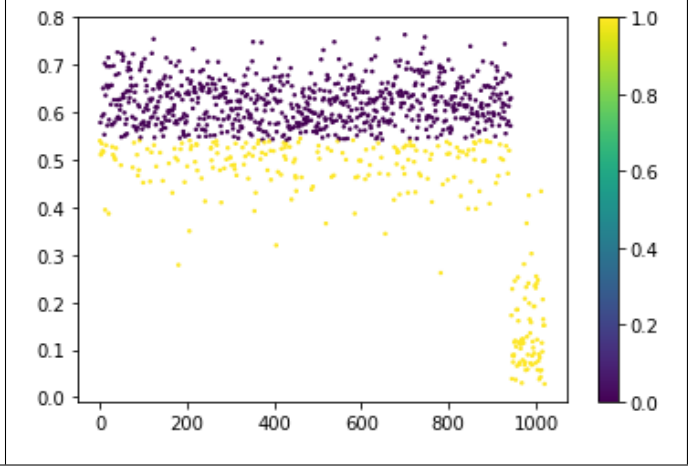
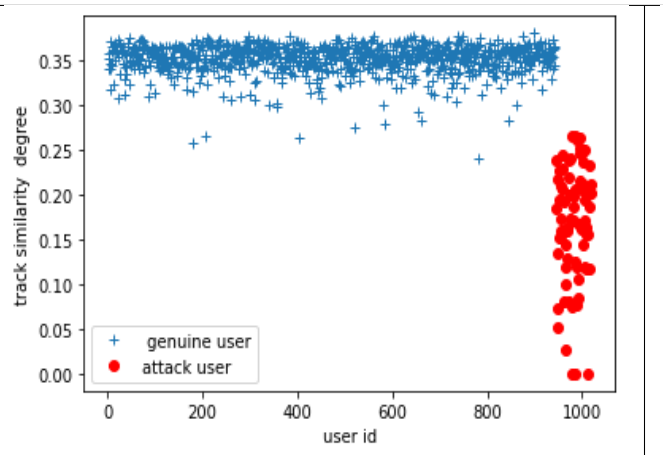
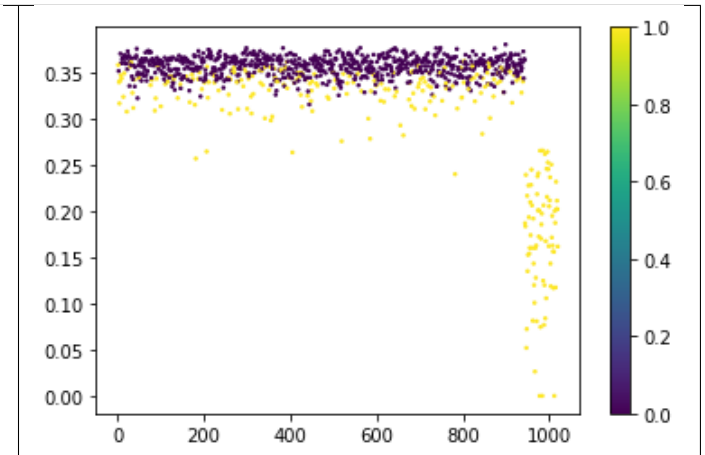
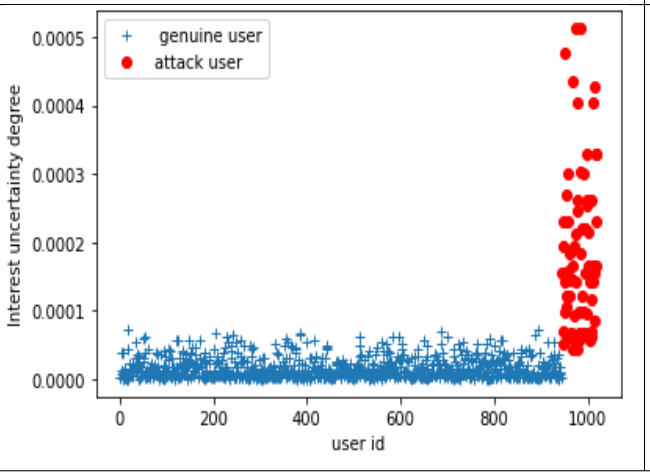
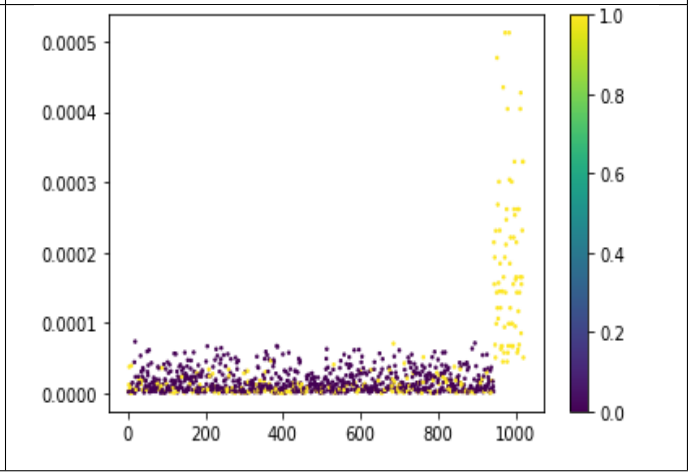
from non-malicious users and are labeled authentic. Into this training data, a mixture of attack types at various attack sizes and filler sizes is injected and labeled attack. The detection featues are then generated for each rating

Table 5.2: Average attack with attack size 15% and filler size 5%



profile, and only the detection features and label of profile is kept as part of the training set. Classifier is trained regularly as new data is fed into the classifier then determines whether the profile is genuine or fake. The result

Table 5.3: Bandwagon attack with attack size 15% and filler size 5%

Feature extraction	Clustering
	
	
	

of classification algorithm is then verified. As the number of training data increases, the classifier becomes more and more accurate in predicting the fake profile. The experimental results demonstrate that proposed supervised

Table 5.4: Supervised approach for detecting shilling attack

Attack model	Attack size	Filler size	SVM Classifier
Random attack	10%	10%	85.11%
		12%	83.19%
	30%	10%	85.98%
		12%	84.91%
	50%	10%	86.75%
		12%	85.76%
Average attack	10%	10%	86.10%
		12%	83.19%
	30%	10%	87.27%
		12%	86.14%
	50%	10%	87.42%
		12%	86.35%
Bandwagon attack	10%	10%	86.06%
		12%	84.66%
	30%	10%	87.27%
		12%	84.87%
	50%	10%	87.38%
		12%	86.06%

detection model achieve a better detection performance in shilling detection. The behavior features such as track correlative degree is depicted in Figure 5.2, track similitude degree is depicted in Figure 5.3 and interest uncertainity degree is depicted in Figure 5.4 are generated with an accuracy of 85.41% By analysis of Table 5.4, when attack size increases, all the models perform very well and it becomes easier to detect attack and authentic profiles. Some models give accuracy as low as low attack sizes. As compared to 943 authentic users attack profiles remains very less. Because of improper ratios of these two types of profiles some models does not give good performance. But when it increase the attack size, a sharp rise in the accuracy is clearly visible. At 50% of attack size, very good accuracy is given by many models in almost every case. Reason for this accuracy is at 50% attack size we insert 500 attack profiles in the system and the ratio between authentic and attack profiles become very good as a result proper training and testing is possible.

SVM classifier is likely to be more robust than other models since its classifier essentially incorporates all training examples simultaneously in evaluating a given profile. The SVM algorithm has been studied widely, in part due to

its theoretical basis and properties of its decision boundary. Specifically it mathematically finds the optimal decision hyperplane with the largest margin per attribute. What this means for the adversarial classification problem is that all features are considered such that they can meaningfully influence the classification. Conceptually this has the nice feature that it would likely be more difficult for an attacker to disguise their entire signature and still have an effective attack. Thus supervised classification method outperforms than unsupervised clustering method with an accuracy of 85.41%.

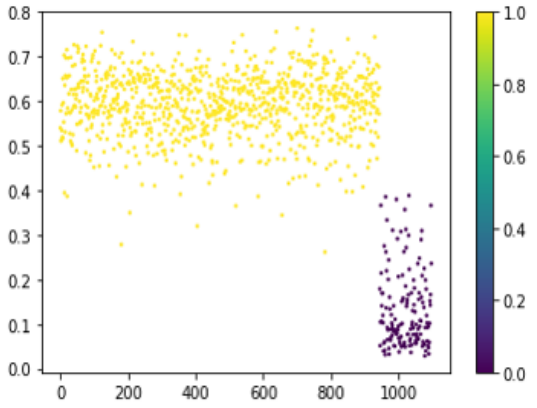


Figure 5.2: Track correlative degree

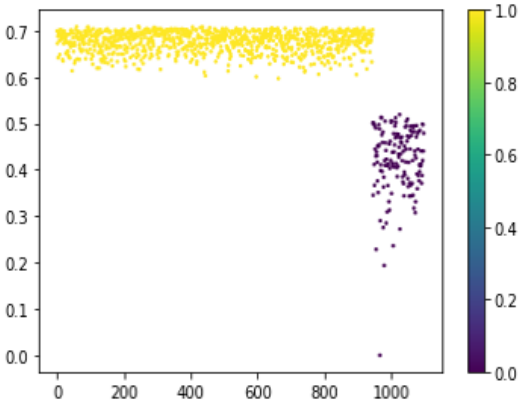


Figure 5.3: Track similarity degree

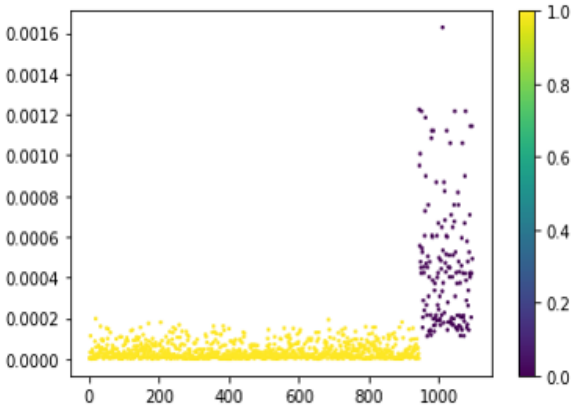


Figure 5.4: Interest uncertainity degree

Chapter 6

Conclusion

Shilling attacks have been shown to be effective threats to the robustness of collaborative recommender systems. Detection of shilling profiles can help to improve the credibility of recommender systems. It start by the memory and inertial effect of human behaviors and extract detection features reflecting behavior differences between genuine and shilling profiles, i.e., track correlation degree, track similarity degree and interest uncertainty degree. In this project, mainly focused on the behavior of attacks in the system. When attacks are inserted some change comes in the recommendations, measure these changes in the form of feature extraction. In the next phase, focused on the detection of these attacks. For this purpose, supervised profile classification approach is used. This project point out the vulnerabilities shared by the most commonly-implemented collaborative algorithms and demonstrate a supervised classification learning approach can add significant robustness to shilling attacks is accurate about 85.41%. Specifically incorporated the attack-specific feature extraction into the classifiers and SVM may provide additional protection against these types of attacks as well.

References

[1] F. Ricci, L. Rokach, and B. Shapira, Recommender Systems Handbook. Berlin, Germany: Springer, 2015.

[2] S. K. Lam and J. Riedl, "Shilling recommender systems for fun and profit," in Proc. 13th Int. Conf. World Wide Web, 2004, pp. 393–402.

[3] M. P. O'Mahony, N. J. Hurley, and G. C. M. Silvestre, "Recommender systems: Attack types and strategies," in Proc. 20th Nat. Conf. Artif. Intell., 2005, pp. 334–339.

[4] P.-A. Chirita, W. Nejdl, and C. Zamfir, "Preventing shilling attacks in online recommender systems," in Proc. 7th ACM Int. Workshop Web Inf. Data Manage., 2005, pp. 67–74.

[5] R. Burke, B. Mobasher, C. Williams, and R. Bhaumik, "Classification features for attack detection in collaborative recommender systems," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2006, pp. 542–547.

[6] R. Burke, B. Mobasher, C. Williams, and R. Bhaumik, "Detecting profile injection attacks in collaborative recommender systems," in Proc. 8th IEEE Int. Conf. E-Commerce Technol. 3rd IEEE Int. Conf. Enterprise Comput., E-Commerce, E-Services (CEC/EEE), Jun. 2006, p. 23.

[7] C. A. Williams, B. Mobasher, and R. Burke, "Defending recommender systems: Detection of profile injection attacks," Service Oriented Comput. Appl., vol. 1, no. 3, pp. 157–170, Oct. 2007.

[8] Z. Yang, L. Xu, Z. Cai, and Z. Xu, "Re-scale AdaBoost for attack detection in collaborative filtering recommender systems," Knowl.-Based Syst., vol. 100, pp. 74–88, May 2016.

[9]Y. Hao, F. Zhang, J. Wang, Q. Zhao, and J. Cao, "Detecting shilling attacks with automatic features from multiple views," Secur. Commun. Netw., vol. 2019, pp. 1–13, Aug. 2019.

[10] B. Mehta, T. Hofmann, and P. Fankhauser, "Lies and propaganda: Detecting spam users in collaborative filtering," in Proc. 12th Int. Conf. Intell. User Interface (IUI), 2007, pp. 14–21

[11] K. Bryan, M. O'Mahony, and P. Cunningham, "Unsupervised retrieval of attack profiles in collaborative recommender systems," in Proc. ACM Conf. Recommender Syst. (RecSys), 2008, pp. 155–162.

[12] Z. Yang, Z. Cai, and X. Guan, "Estimating user behavior toward detecting anomalous ratings in rating systems," Knowl.-Based Syst., vol. 111, pp. 144–158, 2016.

[13] Z. Yang, Z. Cai, and X. Guan, "Spotting anomalous ratings for rating systems by analyzing target users and items," Neurocomputing, vol. 240, pp. 25–46, 2017.

[14] H. Cai and F. Zhang,"BS-SC: An Unsupervised Approach for Detecting Shilling Profiles in Collaborative Recommender Systems",Vol. 33, NO. 4, APRIL 2021.