Robustness Analysis of Naïve Bayesian Classifier-Based Collaborative Filtering

Cihan Kaleli and Huseyin Polat

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

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

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

1 Introduction

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

Malicious users can insert bogus profiles, referred to as shilling attacks, in a very straightforward way into recommender systems’ databases to manipulate the estimated predictions on behalf of their advantages. The advantage helping people be part of recommender systems easily then becomes a vulnerability for the systems. Consequently, CF algorithms can be faced with various profile injection attacks [2-3]. In a traditional example of attacking scenario, any product producer may want to increase its product’s popularity. To do so, it tries to insert fake user profiles into the system in which the target product is extremely liked. In another scenario, the same producer might intend to decrease the popularity of one of its competitor’s product by creating and inserting bogus profiles.
CF algorithms suffer from shilling attacks. Thus, researchers introduce several studies examining the robustness of CF algorithms against them [4,5]. However, previous works examine numerical ratings-based CF algorithms and there is no work covering the case when the ratings are in binary form. Hence, we primarily focus on how the common basic attack models can be applied to NBC-based CF. All users having rating for the target item participate in recommendation process in NBC-based scheme. Thus, vulnerability of NBC-based CF algorithm might increase against profile injection attacks. We particularly introduce binary forms of six mostly implemented attack types, i.e., segmented attack intends to push a product, reverse bandwagon and love/hate attacks are employed as nuke attacks, while random, average, and bandwagon attacks can be considered for achieving both goals. We investigate how robust NBC-based CF algorithm under such attacks. For the purpose of measuring success of binary attacks, we propose a new metric. We perform real data-based experiments and their results clearly show that the proposed binary forms of shilling attacks are capable of biasing prediction results of NBC-based CF algorithm in the direction of their aims.

2 Related Work

Dellacoras [6] discusses negative effects of fraudulent behaviors of users on online reputation systems inspiring shilling attacks concept. O’Mahony et al. [23] introduce the first works about shilling attacks, where the authors analyze vulnerabilities of CF systems against biasing prediction results. Initially, shilling attack strategies are discussed by O’Mahony [7]. The proposed attacks are performed by inserting fake user data to the CF systems. Later, Lam and Riedl [8,9] introduce four open questions related to effectiveness of shilling attacks. Mobasher et al. [10,11] determine attack strategies and present the basic attack types such as random, average, bandwagon, and love/hate attacks. Burke et al. [4] examine bandwagon and popular item attacks. Burke et al. [5] propose a different attack type called segmented attack targeting a set of particular users. Cheng and Hurley [12] propose diverse and obfuscated attack models to be effective on model-based CF schemes. To bias users’ reputation, copied-item injection attack is presented by Oostendorp and Sami [13]. Gunes et al. [14] present a comprehensive survey about shilling attack studies explaining attack descriptions, detection methods, designing robust recommender algorithms, and evaluation metrics and data sets.

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

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