Neurocomputing 207 (2016) 374–386

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Reconstructing rated items from perturbed data

Burcu Demirelli Okkalioglu[a], Mehmet Koc[b], Huseyin Polat[c,*]

[a] Computer Engineering Department, Yalova University, 77100 Yalova, Turkey
[b] Electrical & Electronics Engineering Department, Bilecik Seyh Edebali University, 11210 Bilecik, Turkey
[c] Computer Engineering Department, Anadolu University, 26470 Eskisehir, Turkey

ARTICLE INFO

Article history:
Received 13 February 2015
Received in revised form
4 May 2016
Accepted 5 May 2016
Communicated by Peng Cui
Available online 12 May 2016

Keywords:
Data reconstruction
Noise reduction
Auxiliary information
Privacy
Randomized perturbation
Collaborative filtering

ABSTRACT

The basic idea behind privacy-preserving collaborative filtering schemes is to prevent data collectors from deriving the actual rating values and the rated items. Different data perturbation methods have been proposed to protect individual privacy. Due to different privacy concerns, users might disguise their data variably to meet their own privacy concerns. In addition to reconstructing the true rating values, data collectors might try to reconstruct the rated items.

In this paper, our goal is to reconstruct the rated items with the help of auxiliary information when users mask their confidential data inconsistently in privacy-preserving prediction systems. We first need to estimate the number of the rated items. Then we have to predict the rated items. To do so, we first use existing methods to eliminate noise from the disguised data. We improve our predictions by utilizing the auxiliary information. Our real data-based empirical outcomes show that our proposed approaches are able to reconstruct the rated items with decent accuracy in spite of variable data masking.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

With the growth of the Internet, the number of e-commerce Web sites has gradually increased every passing year. E-commerce companies provide a fast way for people to purchase what they search or desire without spending too much time. People do not have to visit any physical store anymore to buy things. They prefer purchasing products online. For example, an item located at different cities or countries can be easily bought online. However, as the number of the products gets higher, e-commerce companies have begun to handle too much information to provide quality services for customers. Thus, a new problem called *information overload* has emerged. To deal with this problem, collaborative filtering (CF) has been proposed [1,2]. In the past, people shared their opinions with like-minded friends and relatives about books, movies, places, and restaurants; and they gave advices to each other about things they like. Now, CF systems take the place of the previous idea of exchanging recommendation between people. A CF system makes recommendations by matching users based on their past preferences or tastes [3,4].

Cranor et al. [5] conduct a survey to figure out online users' privacy concerns. According to the survey's results, people have different levels of privacy concerns. Most of the people do not want to disclose personally identifiable information. The other important result is that people think that some information such as social security number, credit card number, or phone number is more sensitive than other information such as e-mail address, age, or favorite TV. People are much less willing to share data with the third parties because they want to be sure that their information will not be shared with other companies for different purposes. Unsolicited communication is another problem for people. On the other hand, some people do not worry about privacy. They are willing to share information to provide advantages such as getting free coupons or pamphlets. All results show that each person has different privacy concerns and some people have more privacy concerns than others. Likewise, CF systems have similar privacy risks for users. Since CF systems fail to protect users' privacy, researchers propose privacy-preserving collaborative filtering (PPCF) schemes. PPCF methods allow users to share their data while protecting privacy as well as providing accurate predictions. To preserve privacy, sensitive/private data has to be disguised before sending it to the CF system. Several data disguising methods have been proposed to preserve privacy including but not limited to homomorphic encryption [6,7], anonymization [8,9], randomized response techniques [10,11], secure multi-party computation [12–14], and randomization [15,16].

Randomization is a very common data disguising method in PPCF. Polat and Du [16] propose randomization for PPCF to protect individual privacy. According to their data disguising method, each user first transforms the original ratings into z-score values and then adds random numbers to them using randomization. The disguised data is sent to the server instead of the original data.

[*] Corresponding author. Tel.: +90 222 321 3550; fax: +90 222 323 9501.
E-mail address: polath@anadolu.edu.tr (H. Polat).

http://dx.doi.org/10.1016/j.neucom.2016.05.014
0925-2312/© 2016 Elsevier B.V. All rights reserved.

However, some researchers show that randomization may not protect individual privacy [17–19]. Private data can be reconstructed from the disguised data. Zhang et al. [19] claim that a great number of data is derived from the disguised one when the approach proposed by Polat and Du [16] is used. Thus, Polat and Du [20] propose to use inconsistent data disguising method to increase privacy instead of the naïve solution [16]. The naïve solution may not sufficiently protect user's privacy [19]. The server or the CF system knows the rated items. Users might think that which items they purchase or rate are also confidential. Hence, this kind of information may decrease privacy. To deal with varying levels of privacy concerns, Polat and Du [20] suggest different data disguising methods and allow users to disguise their sensitive data inconsistently. According to one of their scenarios, each user disguises some of the uniformly randomly selected empty items with random numbers besides the rated items. There are different scenarios how empty items are handled based on users' concerns. Some users want to disguise all empty items, although others disguise a predefined percentage of empty items only. There are some recent studies based on inconsistent data disguising methods to hide the ratings and the rated items with decent accuracy [21–24].

The aim of this paper is to reconstruct the rated items from disguised data, which masked inconsistently. The disguised data contain the rated items and the filled items with fake ratings. There are two cases to define the number of the filled cells with fake ratings. In the first case, each user randomly selects a predefined percentage of empty items in the same way. The number of empty cells to be filled can differ for each user depending on her concerns in the second case. Inconsistent data disguising method increases privacy compared with the naïve method. Our purpose is to predict the real rated items when inconsistent data disguising method is preferred by users to mask their sensitive data. Different from the previous works, we present several approaches to estimate the rated items in several scenarios. There are two important tasks in the paper to be solved. The first task is to estimate the number of the real ratings from the disguised data and the second one is to identify the rated items. We also claim that auxiliary information can help us reconstruct the rated items in addition to existing data reconstruction methods.

The main contributions of the paper can be listed as follows:

(1) We study how to reconstruct the rated items from inconsistently perturbed numeric data. To the best of our knowledge, this study is the first one focusing on deriving the rated items in numeric ratings-based PPCF systems.
(2) Two approaches are proposed to estimate the number of the real ratings from the disguised data.
(3) In PPCF, matrix factorization methods are widely used for prediction. However, we use the matrix factorization methods to decrease the effects of the noise data.
(4) We finally scrutinize the joint effect of auxiliary public information and characteristic properties of CF systems in addition to existing matrix factorization methods.

The rest of the paper is structured as follows. In Section 2, the related work is explained. Section 3 gives the detailed information about inconsistent data disguising scenarios. Privacy and problem definition are explained in Section 4. Then we show that how the number of real ratings from the disguised ones are estimated. We continue with describing different noise elimination methods and different approaches utilizing auxiliary information in Section 5. In Section 6 and Section 7, several experiments are performed to evaluate our approaches and Section 8 demonstrates our conclusions and future research directions.

2. Related work

Protecting individual privacy is the main concern in PPCF. Polat and Du [16] propose randomization as a data disguising method in PPCF to alleviate users' concerns. According to their data disguising scheme, each user transforms the original ratings into z-score values and then disguises z-score values using randomization. Users send the disguised data rather than the original data to the CF system or the server. Randomization is first proposed in the privacy-preserving data mining (PPDM) by Agrawal and Srikant [15]. The idea behind randomization is to add a random number (r_i) to the original data (x_i) so that unauthorized people cannot learn the original data from masked data ($x_i + r_i$). Polat and Du [16,25] show that the CF systems can perform CF computations with decent accuracy on perturbed data although they do not know the original data.

Randomization is a common and simple method to protect individual privacy in PPCF. However, some studies show that randomization may not keep private information as much as believed. Agrawal and Srikant [15] show that reconstructing the distribution of the original data from the disguised data is possible. Their work is extended by Agrawal and Aggarwal [26] using expectation maximization (EM) algorithm. They show that EM algorithms work well to estimate the original distribution when the data is large enough. Kargupta et al. [17,27] claim that randomization may not hide the private information. They propose a random matrix-based spectral filtering method to prove their hypothesis. They show how the original data can be obtained from the disguised data using random matrices properties. Huang et al. [18] conduct a similar study with Kargupta et al. [17]. They propose two data reconstruction methods, which are principal component analysis (PCA) and Bayes estimation. They show that when the correlations between attributes are increased, the reconstruction results improve. Guo and Wu [28] and Guo et al. [29,30] also use the spectral filtering method to show how an attacker can compare their results with the original data. Guo et al. [29,30] propose singular value decomposition (SVD)-based data reconstruction method. By using this method, the data owners can decide the least amount of random data that should be added to the original data to protect individual privacy. Zhang et al. [19] show the derivation of the private data from the disguised one by using k-means clustering and SVD-based data reconstruction method. Their study differs from the previous studies because previous studies conduct experiments with complete matrices. Used data sets are usually artificial and do not contain empty cells. As the nature of recommender systems, CF systems contain many empty cells; the study of Zhang et al. [19] is prominent.

There are different existing data reconstruction methods in the literature. Apart from data reconstruction methods, characteristic properties of CF system and auxiliary information can help to derive the original data from the disguised one. With the emergence of social networks, there are different sources of obtaining auxiliary information about users. Calandrino et al. [31] use auxiliary information to infer users' transaction, which is performed on a CF system. They claim that public information can be obtained from some popular services and can be used to derive user related data from CF systems. In another study, auxiliary information is used to alleviate data sparsity and cold-start problems in CF [32]. The authors claim that social network data is a huge potential to overcome the data sparsity and the cold-start problems. They employ a star-structure graph and utilize useful auxiliary information from cross-domains. Their experiments provide better results with the integration of auxiliary cross-domain information. Learning insights about users from different social domains could be an important factor for business intelligence [33]. Liu et al. [33,34] aim to link users among different

Download English Version:

https://daneshyari.com/en/article/494467

Download Persian Version:

https://daneshyari.com/article/494467

Daneshyari.com