

Privacy vs. Utility: An Enhanced K-coRated

Ze Xiang[1], Ghada El-Haddad[2(✉)], and Esma Aïmeur[2]

[1] Beihang University, Beijing Shi, China
seanxiangze@buaa.edu.cn
[2] Department of Computer Science and Operations Research, University of Montreal, Montreal, Canada
{elhaddag,aimeur}@iro.umontreal.ca

Abstract. In recommender systems, collaborative filtering (CF) techniques are becoming increasingly popular with the evolution of the Internet. Such techniques are based on filtering or evaluating items through the opinions of online consumers. They use patterns learned from their behavior or preferences to make recommendation. In this context, it is of great importance to protect users' privacy when there is a need to publish data for a specific purpose which conduct to the usefulness of collaborative recommender systems. However, too much protection to individual privacy will lead to the loss of data utility. How to balance between privacy and utility is challenging. In this paper, we propose a privacy-preserving method based on k-means and k-coRating privacy-preserving model. First, we evaluate the k-coRated model by privacy and utility. Then, according to the drawbacks of it, we introduce our solutions to address the problem. Finally, we make a comparison between our model and k-coRated model in different aspects. As a result, our model outperforms k-coRated model with respect to utility as well as privacy.

Keywords: Privacy · Utility · Balance · K-coRated · Recommendation system · K-means

1 Introduction

In 2006, Netflix held a competition to award $1 million to the best winner who could improve the performance of its recommendation system[1]. In 2007, two researchers from The University of Texas claimed that they were able to re-identify users in the published data with film ratings on the Internet Movie Database. They proposed an attack model [1] based on the assumption of the knowledge gained by attackers. By comparing the auxiliary information gained by attackers and published data, the model could either re-identify a user or tell that he is not in the data released.

On December 28, 2009, a news report[2] announced that Netflix was sued for the privacy breach. On March 2010, Netflix decided [2] to cancel the pursue a second Prize

[1] Netflix. 2006. Netflix Prize. Retrieved December 17, 2018 from https://www.netflixprize.com/.

[2] Netflix Sued for Largest Voluntary Privacy Breach To Date. Retrieved December 17, 2018 from https://privacylaw.proskauer.com/2009/12/articles/invasion-of-privacy/.

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11619, pp. 566–578, 2019.
https://doi.org/10.1007/978-3-030-24289-3_42

competition being previously announced in the same year and to respond to a lawsuit and Federal Trade Commission privacy concerns.

Although the event of Netflix privacy disclosure has passed for many years, privacy-preserving receives much attention recently as data mining technology is booming. Indeed, privacy is getting much attention in the literature [3–5]. Many models are proposed to define privacy awareness and how to protect people from the aspect of cyber-security.

However, nowadays as a significant change in business domains, it is unavoidable for organizations and companies to cooperate. On the one hand, they have to share their data to get their targeted result; on the other hand, they are afraid of invading users' privacy when sharing data. Therefore, filling methods to evaluate the performance of balancing privacy and utility become a challenge [6].

Data publication usually involves two main issues. One is to protect users' privacy. The other is to ensure the utility of data. Because when the publisher tries to anonymize user record, it must lead to data modification, which will further lead to a decrease in data utility. It is a trade off between individual privacy and data utility.

Nowadays, there is a large stream of literature on privacy, scientific studies have widely addressed privacy models [7–9]. They tried to classify attributes of user record into sensitive data (such as disease and salary), non-sensitive data and quasi-identifiers (such as zip code, age, and gender). Based on that, they proposed methods to protect sensitive data and quasi-identifiers in order to make people indistinguishable. However, such classification is heuristic and varies from person to person because people have different definitions of privacy. Additionally, prior studies proposed privacy protection models based on encryption [10–12], but generally, it is impractical considering the computational time.

As for data utility, as per our knowledge, there is no universal standard to evaluate the utility of data since it depends on the scenario adopted and the methods applied. In the same context, previous works have proposed a data utility model in the scenario of fog computing environments [13]. However, it just gives an abstract definition of utility and is not effective in practice. In another study, the authors proposed an anonymization method to improve data utility for classification [14], but for utility assessment, they evaluate it based on the performance of classification, which does not apply to other scenarios.

When considering both privacy and utility, the compromise between them is problematic. In previous studies [15], the authors specified that there is no trade off actually if assuming that attackers and scientists have the same background in knowledge and technology, while other works [16–18] tried to reach a balance in specific fields.

In the field of the recommendation system, the data of users involves their history of purchases, personal preferences and browsing records, which can be sensitive sometimes. Therefore, publishers like Netflix should give the same emphasis to user privacy as data utility when they try to release the data. On the one hand, they must anonymize users' information, on the other hand, they must ensure data utility. Since there is no universal standard for utility, we propose two aspects to increase utility empirically and based on those we propose a creative method to enhance a privacy-preserving model named k-coRated.

In Sect. 2, we briefly introduce some related work about privacy, utility and the balance between them. In Sect. 3, we introduce the k-coRated model and highlight its advantages and disadvantages, and based on that; we propose our method to solve the problem. In the same section, we simulate a cyber-attack model to evaluate our model and make a comparison between it and the k-coRated model. We present the results with the experiments in Sect. 4. Finally, we conclude this paper in the last section and discuss future work.

2 Background and Related Work

As the technology of data mining is booming and collecting people's data is increasing, there are rising concerns over people's privacy. Preserving confidential data of users while producing accurate predictions is one of the extremely important directions of the researches about recommendation systems [19]. However, how to define privacy and where is the boundary of privacy remain questionable. Several kinds of research have been done attempting to define what privacy is and to provide methods to protect it. For example, K-Anonymity [20] proposed an innovative model explaining privacy by the difficulty to distinguish someone from others. Based on K-Anonymity, L-Diversity [21] and T-Closeness [22] offered stronger protection considering the diversity and distribution of attributes of people. They try to define and extend privacy in different aspects, but there is no universal formalized definition of privacy until the appearance of differential privacy [23], which offers a theoretical framework to protect privacy in statistical databases.

While literature has widely explored privacy protection from malicious cyber-attack, privacy concerns within the publication of data are receiving more and more attention. It is necessary to protect users' privacy while ensuring the utility of data. Brickell and Shmatikov [15] stated that there could be no tradeoff between privacy and utility if we assume that attackers and scientists have the same level of knowledge.

However, Li and Li [17] pointed out that they have got the wrong way and offered a new perspective on data publishing. Most work relates utility to data modification. Chen et al. [24] proposed a novel way to evaluate utility based on the framework of differential privacy. Johnson et al. [16] proposed to trade off utility and privacy on full data. Rastogi et al. [18] discussed the boundary between privacy and utility. Based on prior knowledge of attackers, they were able to analyze when it is possible to design an algorithm to ensure privacy and utility.

Encryption/decryption is another technique for privacy preserving. This kind of technique is applied in distributed computing scenario with a very time-consuming. It consists by encrypting original data before being used. Although such method can ensure recommendation accuracy however there are some limitations regarding to its application. It is unpractical to encrypt/decrypt large datasets and many data mining computations cannot be achieved by encryption/decryption. As a result, the scopes of its usage are restricted.

Zhang et al. [6, 8] proposed a privacy preserving model inspired from k-anonymity technique named k-coRated in order to provide privacy to the data of the movie

recommendation system. The authors proposed this model with k-coRating to retain data privacy of rating-style datasets by replacing some null ratings with "well-predicted" scores and enhance the data utility measured by prediction accuracy. However, they omitted to balance privacy with utility. In the following section, we discuss their work in detail and propose our enhanced method.

3 Methodology

In this section, we show our methodology step by step. First, in Subsect. 3.1, we introduce the k-coRated method. Second, in Subsect. 3.2, we discuss its drawback with respect to utility and propose our method based on the hybrid of k-means. In the remaining subsections, we give a more detailed discussion about our method.

3.1 Introduction to K-coRated

The k-coRated model defined user's privacy from the items they rated. It only cares whether a user has rated an item or not. The model is suitable for ratings with a small scale, like the 5-star rating system. However, with large scale rating system, it is not that effective regarding privacy protection. Below we show the definition of k-coRated privacy and related concepts.

Definition 3.1. For any user $u \in U$ and any item $i \in I$, $R_{u,i}$ is the rating of user u to item i, we define that

$$u[i] = \begin{cases} 0, & R_{u,i} = NULL \\ 1, & otherwise \end{cases} \tag{1}$$

Definition 3.2 (coRated Equivalence). Any two users $u,v \in U$ are of coRated Equivalence if for all $i \in I$, $u[i] = v[i]$.

Definition 3.3 (k-coRated Privacy). A rating matrix satisfies k-coRated privacy if each user $u \in U$ has at least k-1 coRated Equivalence users.

In the work of k-coRated model [20], the authors proved that to make a rating matrix satisfy k-coRated privacy is NP-hard problem and proposed a heuristic algorithm to do it; we summarize the process to two steps. First, they start to group the users by a specific criterion. Second, they fill the original rating matrix with a predicted rating to make it satisfy k-coRated privacy. Although they claimed that it could provide a considerable degree of privacy protection to the users in the data published, they omitted to weigh the balance between user privacy and data utility.

In this paper, we implemented their proposed algorithm and found that after processing by the k-coRated method, the original rating matrix of MovieLens 100 k dataset, which has 80,000 ratings at first, had about 370,000 ratings after anonymization. Thus, the number of ratings that it added to the matrix is so large.

However, as per our knowledge, there is no universal standard to evaluate utility, because it depends on the scenario to which data is going to be applied. Even in the field of the recommendation system, it is still hard to define utility for the fact that it is uncertain that what property of rating leads to a better recommendation.

3.2 Utility Concern and Solutions

Since it is impossible to evaluate utility absolutely, we propose a method with comparably better utility. In the best case, we wish to realize our objective without any modification to data. For that reason, we choose the method of anonymizing data by adding noise and getting less modification. Consequently, if it is unavoidable to make some changes to data, we hope it as little as possible.

Furthermore, the reason why we bypass modification to data is that original data is the best because people in their daily life generate it and the numbers that we add to the matrix are fake. However, if what we add to the original matrix equals what will happen or happened in the real world, then the utility of data should remain the same. So, a better prediction will also result in better utility.

Less Modification: We hope to reach k-coRated privacy with the least modification to data. First, we adopt k-means to improve clustering of users hoping that users in the same cluster have already co-rated most items. Then we propose a new way to do the modification, called Delete Little to Add Less (L2L).

Better Prediction: User-based Collaborative Filtering (CF) methods are typically used in the field of recommendation system and proved to have better results than item-based CF methods. In order to predict the preference of a user u to item i, denoting as $P_{u,i}$, user-based CF methods would like to refer to the preferences of K-nearest neighbors (k-NN) [25] of u, denoting as NN_u. In our work, we adopt the formula below to make the prediction.

$$P_{u,i} = \bar{u} + \frac{\sum v \in NNu \,\mathrm{sim}(u,v) * Rv, \mathrm{i}}{\sum v \in NNu \,\mathrm{sim}(u,v)} \tag{2}$$

$\bar{u}$ is the average rating of user u, sim(u,v) is the degree of similarity between user u and user v.

3.3 Less Modifications: K-Means

Since we only care about whether a user has rated an item or not, before we apply k-means to our scenario, we should do some normalization.

Definition 3.4 (Normalized Rating Matrix). A rating matrix R is normalized if $R_{u,i}$ satisfy the following equation.

$$R[u,i] = \begin{cases} 1, & \text{User } u \text{ has rated item } i \\ 0, & \textit{otherwise} \end{cases} \tag{3}$$

Since the performance of k-means largely depends on the selection of initial centers, a selection, closer to the real final center, can provide better clusters. In order to get a better basis for measuring the initial set of centers, we conclude two properties of an initial and right set of center heuristically that help to determine a way to find initial centers.

Distance: the initial centers should be distant enough to each other and not located in the same cluster.

Density: the initial center must have many other points around it.

Below, we give the mathematical definitions of density and distance based on the Euclidean distance between users. To balance between distance and density, we propose a weight function between the normalized results of distance and density. Based on those definitions, we design Algorithm 1, as shown in Table 1, to find better initial centers.

Definition 3.5 (dis). for any user u and $v \in U$,

$$dis(u,v) = \sqrt{\sum_i (u[i] - v[i])^2} \tag{4}$$

Definition 3.6 (density). for any user $u \in U$,

$$Density(u) = \frac{1}{\sum_{v \in U} dis(u,v)} \tag{5}$$

Definition 3.7 (distance). for any user $u \in U$,

$$Distance(u) = \sum_{v \in Centers} dis(u,v) \tag{6}$$

Definition 3.8 (weight). For any user $u \in U$,

$$weight(u) = \alpha * \widehat{distance}(u) + (1-\alpha) * \widehat{density}(u) \tag{7}$$

where $* \widehat{distance}(u)$ and $\widehat{density}(u)$ are the normalized results between [0,1].

3.4 Less Modification: L2L

L2L is inspired by the scenario that in a cluster of users, some items are rated by most of the users while a few people rate some items. To those rated by few, it is better to delete them. Figure 1 shows the process of L2L and makes a comparison between it and k-coRated method.

In Fig. 1, the illustrated Table 1 is the original matrix. Four users have rated items from 1 to 5, in order to make them k-coRated following the old method, we add predicted ratings to fill the null blank and get the illustrated Table 2. However, some items are rated by only a few users, like item 1 and item 5 in the illustrated Table 1. Based on our approach, we delete those rated by few users; thus in Fig. 1, we delete the ratings of ($User_1$, $Item_1$), ($User_2$, $Item_1$) and ($User_3$, $Item_5$), then we add predicted ratings to ($User_2$, $Item_4$) and ($User_3$, $Item_2$).

However, how to decide which part to be deleted? Which items are the part that rated by few?

Table 1. Algorithm 1: DD-based initial centers

Input:
k, the number of initial centers
R, normalized rating matrix
Output:
k initial centers
1: Initialize centers=[]
2: Compute density for every user in *R*
3: Set the user with max density as the first center
4: **while** *len*(centers)! = *k* **do**
5: Compute weight for every user in *R* except those in centers
6: Add a user with max weight to centers
7: **end while**
8: **return** centers

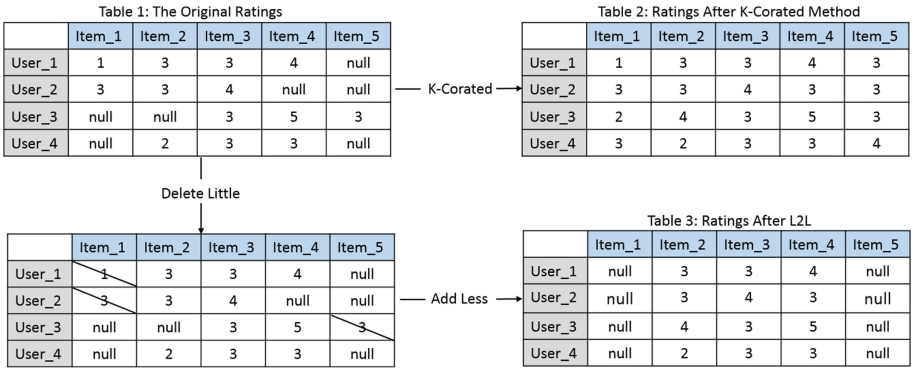


Fig. 1. Process of L2L

Definition 3.9. Below, we define what is the centroid of a cluster *C*, denoting as *ctr*(*C*). For each item $i \in I$,

$$ctr(C)[i] = \frac{\sum_{u \in U} u[i]}{|C|} \tag{8}$$

Based on *ctr*(*C*), we propose that for any item i, if $ctr(C)[i] \leq \beta$, then *i* belongs to the part to be deleted. β is a hyperparameter, which is given by the user of the algorithm based on different scenarios.

3.5 Better Prediction

We base our prediction on Eq. 2. A Better prediction comes from a better definition of similarity between two users. Meanwhile, a better set of nearest neighbors will also result in better prediction.

We make a comparison between the currently popular definitions of similarity, including cosine-based, correlation-based, distance-based, Jaccard-based, and trust-based and pick the best among them. Moreover, we choose the top N similar users of user u as $N\ N_u$.

4 Experiments and Results

In this section, we evaluate our model regarding privacy and utility and compare it with [8]. In order to evaluate privacy, we introduce an attack model first. Considering that there is no universal standard evaluation of utility, we build our assessment by comparing the number of ratings in the matrix and evaluating the performance of data regarding collaborative filtering.

4.1 Experiment Design

Our experiments are performed using MovieLens 100k dataset, which consists of 100,000 ratings from 943 users on 1682 items. Our algorithms are implemented in python on a machine with an Intel Core i7-8700 CPU 3.20 GHz, 32 GB RAM running Windows10 64-bit operation system.

4.2 Privacy

Narayanan and Shmatikov proposed the attack model [1]. In this model (NS attack model), they assume that attackers have gained auxiliary information of users, denoted as *Aux*. They suppose that *Aux* contains m pieces of information, and n of them are correct considering that attackers may not have the right source of information. Every user has a piece of record in the database, denoted as *Rec*. The model defines a function called *Score*, which takes *Rec* of a user and *Aux* that attackers have gained as parameters and output a number as an indicator of the extent to which the user may be the target user of *Aux*.

$$Score(Rec, Aux) = \sum_{i \in supp(Aux)} wt(i) * Sim(Reci, Auxi) \tag{9}$$

For all the users in the published data, it gets a list of scores by applying Score to every Rec of users. Max_1, Max_2, and σ denote the highest score, second highest score, and the standard deviation of scores, respectively. If $\frac{Max_1 - Max_2}{\sigma} < \phi$, where ϕ is a fixed parameter called eccentricity, then there is no match for *Aux*, otherwise the record with the highest score is of the target user. The value of $\frac{Max_1 - Max_2}{\sigma}$ indicates how much the best-match record outperforms others and ϕ is a threshold. Therefore, the model not only requires that the target user record is the one with the highest score, but also that it must outstand other records. The Algorithm 2, shown in Table 1, explains the NS attack model step by step.

We simulated the NS attack by assuming that attackers gained 8 and 16 pieces of information in total, respectively. Figure 2 compares the performance of privacy between our method and k-coRated method. The y-axis indicates the probability to be

re-identified for users in the data, and the x-axis indicates the information gained by attackers. From both two sub-figures, we can see that although there is little difference between our method and k-coRated method when attackers gain few pieces of information, our method outperforms k-coRated method almost two times regarding the possibility to be re-identified as attackers know more.

Table 2. Algorithm 2: scoreboard-RH [1]

Input:
Aux, the auxiliary information gained by attackers
D, the rating dataset ϕ, eccentricity
Output:
the match record or empty set
1: For each record $Rec \in D$, compute $Score(Rec, Aux)$
2: Compute $Max_1 = Max_1(S)$, $Max_2 = Max_2(S)$, $\sigma = std(S)$, where $S = (Score(Rec,Aux), Rec) \in D$
3: **If** $\frac{Max_1 - Max_2}{\sigma} < \phi$, **then**
4: **return** an empty set
5: **else**
6: **return** the match record with the highest
7: **endif**

4.3 Utility: Number of Ratings

We compare the number of ratings in the matrix, which reflects the modification to the data. In order to make the two methods comparable, we make them have the same number of clusters. Figure 3 shows the results of two methods when clustering users into 10, 20 and 30 groups. From Fig. 3, we can determine that after being processed by our method, the number of ratings in the matrix stays around 260,000 while after being processed by the k-coRated method, it increased from about 400,000 to 600,000 as the number of user clusters increases. **Therefore, our method significantly reduces data modification and provides better data utility**.

4.4 Utility: Performance of CF

Since our data is about the recommendation system and collaborative filtering is commonly used in the field, so we decided to test the utility of the data based on the performance of collaborative filtering. We implement different prediction strategy proposed by [8] and use RMSE to evaluate the performance of CF. Below, we explain what RMSE is.

Given the N actual/predicted pairs ($R_{u,\ I}$, $P_{u,\ i}$), the *RMSE* of the N pairs is computed as follows:

$$RMSE = \sqrt{\frac{\sum (Ru, i - Pu, i)^2}{N}} \tag{10}$$

Figure 4 shows the performance of our method compared to the k-coRated method. In each case, we use the same definitions of similarity according to Eq. 2 to make the prediction, and we choose the nearest neighbors from the top 5 to 100 most similar users.

From Fig. 4a, the best *RMSE* is about 0.98 when adopting correlation-based similarity definition and choosing nearest neighbors from top 80, while from Fig. 4b, the best *RMSE* is about 0.94 when adopting correlation-based similarity definition and choosing nearest neighbors from the top 40. **Therefore, our method increases the utility of data considering the performance of CF.**

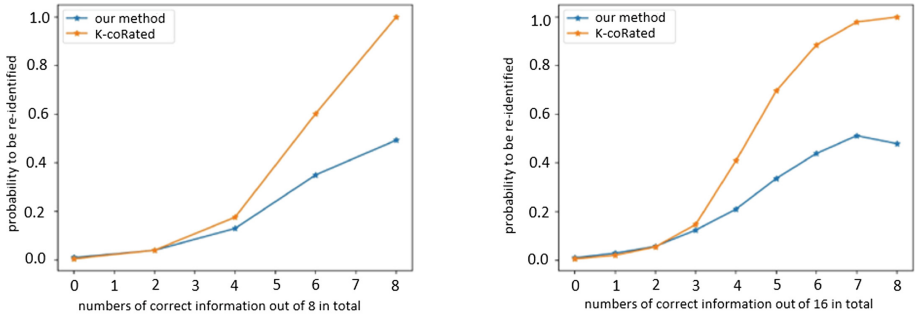


Fig. 2. Privacy comparison

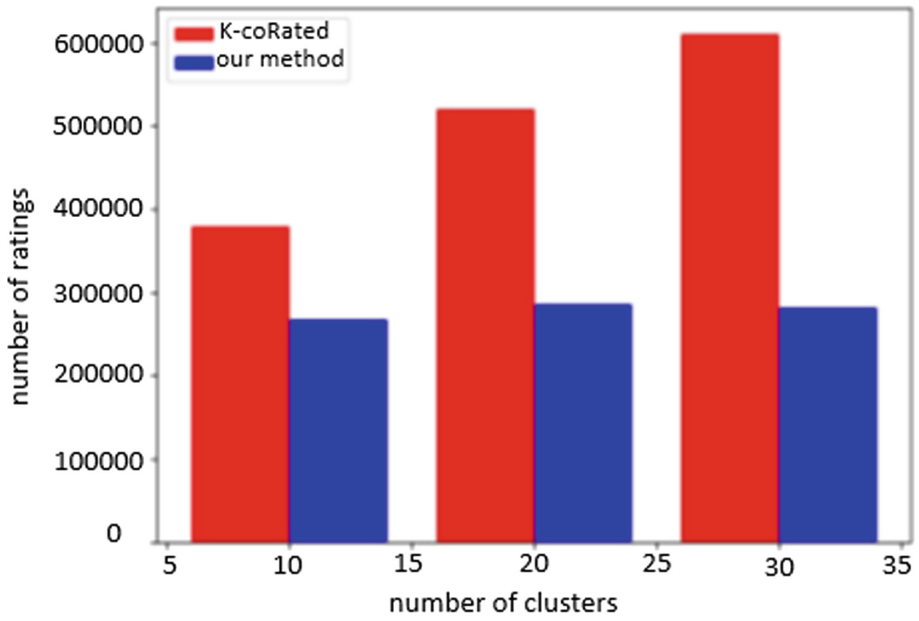


Fig. 3. Comparison of number of ratings

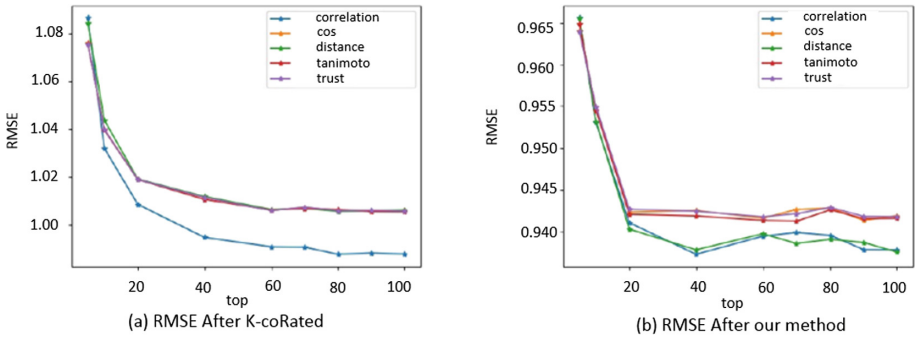


Fig. 4. Utility comparison

5 Conclusions and Future Work

Zhang et al. [8] proposed an innovative model to protect user privacy when data is about to be released. However, they omitted to weight the balance between user privacy and data utility. Based on the drawbacks of its method, we propose a new method involving less modification to the data and providing a better prediction. By adopting k-means and L2L, we achieve a significant reduction in data modification. Meanwhile, under the evaluation of NS attack, we even provide a better degree of privacy protection.

However, since there is no universal standard to evaluate utility, it is hard to judge whether the utility of our method is acceptable or not. Even though, it is intuitively right that less modification leads to better utility, how to define modification remains doubtful. The relationship among modification, addition, and deletion is complicated.

Besides, we only implement and tested our work on MovieLens 100k dataset. In future work, we believe in considering other datasets such as MovieLens 10M, Epinions or Netflix. The experiment on those datasets requires further considerations. Indeed, since MovieLens 100k is not a big dataset, we do not lay our emphasis on computational time. However, we will consider it in more extensive data. Another future work is to design a model to evaluate the utility of data in different fields.

References

1. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: IEEE Symposium on Security and Privacy, SP 2008, pp. 111–125 (2008)
2. Lohr, S.: Netflix cancels contest after concerns are raised about privacy, p. B3. New York Times (2010)
3. El Haddad, G., Aïmeur, E., Hage, H.: Understanding trust, privacy and financial fears in online payment. In: 2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), pp. 28–36 (2018)
4. Gai, K., Qiu, M., Xiong, Z., Liu, M.: Privacy-preserving multi-channel communication in edge-of-things. Future Gener. Comput. Syst. **85**, 190–200 (2018)

5. Barth, S., de Jong, M.: The privacy paradox–investigating discrepancies between expressed privacy concerns and actual online behavior–a systematic literature review. Telemat. Inform. **34**, 1038–1058 (2017)
6. Zhang, F., Lee, V.E., Jin, R.: k-CoRating: filling up data to obtain privacy and utility. In: AAAI, pp. 320–328 (2014)
7. Ye, Y., Wang, L., Han, J., Qiu, S., Luo, F.: An anonymization method combining anatomy and permutation for protecting privacy in microdata with multiple sensitive attributes. In: 2017 International Conference on Machine Learning and Cybernetics (ICMLC), pp. 404–411 (2017)
8. Zhang, F., Lee, V.E., Jin, R., Garg, S., Choo, K.-K.R., Maasberg, M., et al.: Privacy-aware smart city: a case study in collaborative filtering recommender systems. J. Parallel Distrib. Comput. **127**, 145–159 (2018)
9. Zhang, Q., Liu, H., Wu, Y., Lin, C., Lin, G.: Grey maximum distance to average vector based on quasi identifier attribute. In 2017 International Conference on Grey Systems and Intelligent Services (GSIS), p. 119 (2017)
10. Bos, J.W., Castryck, W., Iliashenko, I., Vercauteren, F.: Privacy-friendly forecasting for the smart grid using homomorphic encryption and the group method of data handling. In: International Conference on Cryptology in Africa, pp. 184–201 (2017)
11. Wu, L., Chen, B., Zeadally, S., He, D.: An efficient and secure searchable public key encryption scheme with privacy protection for cloud storage. Soft Comput. **22**, 1–12 (2018)
12. Raisaro, J.L., Choi, G., Pradervand, S., Colsenet, R., Jacquemont, N., Rosat, N., et al.: Protecting privacy and security of genomic data in I2B2 with homomorphic encryption and differential privacy. IEEE/ACM Trans. Comput. Biol. Bioinf. **15**, 1413–1426 (2018)
13. Cappiello, C., Plebani, P., Vitali, M.: A data utility model for data-intensive applications in fog computing environments. In: Mahmood, Z. (ed.) Fog Computing, pp. 183–202. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-94890-4_9
14. Han, J., Yu, J., Lu, J., Peng, H., Wu, J.: An anonymization method to improve data utility for classification. In: International Symposium on Cyberspace Safety and Security, pp. 57–71 (2017)
15. Brickell, J., Shmatikov, V.: The cost of privacy: destruction of data-mining utility in anonymized data publishing. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 70–78 (2008)
16. Johnson, M.P., Zhao, L., Chakraborty, S.: Achieving Pareto-optimal MI-based privacy-utility tradeoffs under full data. IEEE J. Sel. Top. Signal Process. **12**, 1093–1105 (2018)
17. Li, T., Li, N.: On the tradeoff between privacy and utility in data publishing. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 517–526 (2009)
18. Rastogi, V., Suciu, D., Hong, S.: The boundary between privacy and utility in data publishing. In: Proceedings of the 33rd International Conference on Very Large Data Bases, pp. 531–542 (2007)
19. Batmaz, Z., Kaleli, C.: Methods of privacy preserving in collaborative filtering. In: 2017 International Conference on Computer Science and Engineering (UBMK), pp. 261–266 (2017)
20. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl.-Based Syst. **10**, 557–570 (2002)
21. Machanavajjhala, A., Gehrke, J., Kifer, D.: ℓ-density: privacy beyond k-anonymity. In: Proceedings of the International Conference on Data Engineering (ICDE 2006), Atlanta, Georgia (2006)

22. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: IEEE 23rd International Conference on Data Engineering, ICDE 2007, pp. 106–115 (2007)
23. Dwork, C.: Differential privacy: a survey of results. In: International Conference on Theory and Applications of Models of Computation, pp. 1–19 (2008)
24. Chen, K.-C., Yu, C.-M., Tai, B.-C., Li, S.-C., Tsou, Y.-T., Huang, Y., et al.: Data-driven approach for evaluating risk of disclosure and utility in differentially private data release. In: 2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA), pp. 1130–1137 (2017)
25. Keller, J.M., Gray, M.R., Givens, J.A.: A fuzzy k-nearest neighbor algorithm. IEEE Trans. Syst. Man Cybern. 580–585 (1985)