 CMC, vol.56, no.1, pp.137-149, 2018

Reliable Medical Recommendation Based on Privacy-Preserving Collaborative Filtering

Mengwei Hou[1], Rong Wei[1, *], Tiangang Wang[1], Yu Cheng[2] and Buyue Qian[3]

Abstract: Collaborative filtering (CF) methods are widely adopted by existing medical recommendation systems, which can help clinicians perform their work by seeking and recommending appropriate medical advice. However, privacy issue arises in this process as sensitive patient private data are collected by the recommendation server. Recently proposed privacy-preserving collaborative filtering methods, using computation-intensive cryptography techniques or data perturbation techniques are not appropriate in medical online service. The aim of this study is to address the privacy issues in the context of neighborhood-based CF methods by proposing a Privacy Preserving Medical Recommendation (PPMR) algorithm, which can protect patients' treatment information and demographic information during online recommendation process without compromising recommendation accuracy and efficiency. The proposed algorithm includes two privacy preserving operations: Private Neighbor Selection and Neighborhood-based Differential Privacy Recommendation. Private Neighbor Selection is conducted on the basis of the notion of k-anonymity method, meaning that neighbors are privately selected for the target user according to his/her similarities with others. Neighborhood-based Differential Privacy Recommendation and a differential privacy mechanism are introduced in this operation to enhance the performance of recommendation. Our algorithm is evaluated using the real-world hospital EMRs dataset. Experimental results demonstrate that the proposed method achieves stable recommendation accuracy while providing comprehensive privacy for individual patients.

Keywords: Medical recommendation, privacy preserving, neighborhood-based collaborative filtering, differential privacy.

1 Introduction

As Electronic Medical Records (EMRs) and wearable sensors become more widespread, medical datasets tend to be larger and specific methods of exploration are needed to extracting meaningful information. However, even experienced clinicians sometimes find it difficult to deal with the large amount of medical knowledge available to help them complete a particular goal. Thus, clinical organizations must exploit effective methods of

[1] The First Affiliated Hospital of Xi'an Jiaotong University, 277 West Yanta Road, Xi'an 710061, P.R. China.

[2] IBM Research AI, Yorktown Heights, 10593, USA.

[3] Xi'An Jiaotong University, No. 28, Xianning West Road, Xi'an 710061, P.R. China.

[*] Corresponding Author: Rong Wei. Email: weirong@xjtu.edu.cn.

discovering and recommending valuable knowledge to assist clinicians' work.

Recommender systems are becoming more and more important due to the increasing "information overload" problem [Davidson, Liebald, Liu et al. (2010); Das, Datar, Garg et al. (2007)]. Especially, it is difficult for users to determine the suitable information to optimize their process of decision making. Take into account this context, recommender systems provide useful selected information to the target user, which could optimize a large amount of decisions effectively. Some of the recently studies have started to make use of recommender systems for automatically recommending useful knowledge in clinical practices [Sun, Liu, Guo et al. (2016); Zhang, Chen, Tang et al. (2017)]. Collaborative filtering (CF) is one of the most popular recommendation techniques as it is insensitive to product details, and is adopted by many online service providers. In this paper, we aim to grapple with privacy preserving issue in the context of neighborhood-based CF methods in medical recommendation. In clinical environment, CF can be applied to provide clinicians with more correlative information, so as to improve the quality of medical service. During this process, patients' sensitive information is collected by the recommendation system, which arises privacy concerns. Enck et al. [Enck, Gilbert, Chun et al. (2012); Wondracek, Holz, Kirda et al. (2010); Li, Lv, Xia et al. (2011)] have shown that users' privacy could be exploited by service providers or malicious users to gain profits.

In this paper, we proposed a *private preserving medical recommendation (PPMR)* based on neighborhood-based collaborative filtering. The contribution of this work is summarized as follows:

- We provide *Private Neighbor Selection* to prevent the adversary from malicious hacking the patients' treatment information and demographic information. Specifically, a new de-identification *k*-anonymity method, *Optimal Lattice Anonymization (OLA)* is adopted to produce a globally optimal de-identification solution suitable for EMRs datasets. After the de-identification, the most similar neighbors are selected privately based on the de-identification datasets. Therefore, an adversary is unlikely to use the combination of quasi-identifier to identify an individual patient.
- We propose a *Neighborhood-based Differential Privacy Recommendation Algorithm*, with the aim of proving comprehensive privacy for the individual patient, as well as maximizing the accuracy of recommendations. Our algorithm consists of several steps, measuring (with noise) progressively more challenging aspects of the data before feeding the measurements to appropriately parameterized variants of the currently algorithms. We first describe the approach at a high level, before describing the sequence of precise calculations more concretely.
- At last, we conduct the security analysis and performance evaluation for the proposed scheme. The experiments carried out on the *real-world EMRs* dataset verify that the proposed medical recommendation scheme is effective and scalable.

2 Related work

In this section, we review the previous work in the literature related with our work. We will also explain the differences between our methods and the previous ones.

2.1 Medical recommendation system

In terms of applications, a lot of recent work has been done in mining the various kinds of EMRs data for actionable insights to improve the quality of healthcare delivery. For example, Zhou et al. [Zhou, Wang, Hu et al. (2014)] proposed a method to infer phenotypic pattern from EMRs. Lakkaraju et al. [Lakkaraju and Rudin (2016)] proposed to use a Markov Decision Process (MDP) to provide cost-effective recommendations based on a healthcare institution's financial restrictions. Hirano et al. [Hirano and Tsumoto (2014).] used occurrence and transition frequency to discover typical order sequences. Liu et al. [Liu, Wang, Hu et al. (2015)] developed a method to identify most significant and interpretable graphical feature from longitudinal EMRs. However, their work is mainly based on discovering effective recommendation algorithm in medical datasets without considering the privacy issue in the medical recommendation process.

2.2 Privacy preserving recommendation systems

A number of research has been working on privacy violations in the modern big data systems, including *cryptographic, perturbation and obfuscation*. Zhan et al. [Zhan, Hsieh, Wang et al. (2010)] solved a similar problem by applying homomorphic encryption and scalar product approaches. Han et al. [Han, Qian, Yang et al. (2016)] proposed a novel physical-layer identification system, utilizes unique features of wireless devices to provide authenticity and security guarantee. The cryptographic method preserves high performance but facing with serious scalability issues. Perturbation will change a user's rating by adding noise before submitting to the recommender system. Polatidis et al. [Polatidis, Georgiadis, Pimenidis et al. (2017)] proposed a multi-level privacy preserving method for collaborative filtering systems by perturbing each rating before it is submitted to the server. Obfuscation replaces a certain percentage of a user's rating by random values. Berkovsky et al. [Berkovsky, Eytani, Kuflik et al. (2007)] decentralized rating profile among multiple repositories and replaced some ratings with their mean.

In order to address these problems, differential privacy, a more rigid notion, has been proposed [Dwork (2006)]. Differential privacy provides a strong and provable privacy definition that can quantify the privacy risk to individuals. As a prominent privacy definition, Mcsherry et al. [Mcsherry and Mironov (2009)] were the first to introduce the differential privacy into recommender system using Laplace noise. Hardt et al. [Hardt and Roth (2011)] converted the recommendation problem into the Matrix Completion problem. Zhu et al. [Zhu, Ren, Zhou et al. (2014)] proposed a truncated similarity function in private neighbor selection so as to achieve differential privacy for neighbor-based collaborative filtering. However, the above methods only focused on the online commercial recommendation system.

In this paper, we provide a private preserving medical recommendation (PPMR) algorithm based on privacy-preserving collaborative filtering. The algorithm we proposed ensures that both the treatment information and demographic information are considered and protected.

3 Proposed method

In this section, we propose a private preserving medical recommendation (PPMR) algorithm to address the privacy preserving issue in medical recommendation process.

Firstly, we present an overview of the algorithm, followed by a detailed discussion. Then we provide a theoretical analysis on how PPMR achieve the differential privacy preserving purposes while retaining the utility for recommendation purposes.

3.1 The private preserving medical recommendation algorithm

For the privacy preserving issue in the context of neighborhood-based CF methods, the preserving targets differ between the user-based methods and item-based methods due to the different perspectives regarding definition of similarity. Traditional non-private user-based CF methods works as follows: The first stage aims to collects users' historical behaviors and users' basic information to identify the users of k nearest neighbors, and the second stage aims to predict the rating by aggregating the ratings on those items that identified neighbor users rated. We propose the PPMR algorithm to address this problem.

Detail for the first operations is presented in Section 3.2, and the second operation and theoretical analysis on privacy preserving is provided in Section 3.3.

3.2 Private neighbor selection

Private neighbor selection aims to privately select *k* neighbors from a list of candidates for the privacy preserving purpose. Prior to any anonymous process, direct identifiers (name, ID number, etc.) need of course to be suppressed from the dataset. However, some of the attributes that remain in the anonymized dataset may be *quasi-identifiers*, which may facilitate indirect re-identification of respondents through external data source (available as attackers' background knowledge) that combine those attributes with direct identifiers.

3.2.1 De-identifying patients' health data

In EMRs datasets, patients' attributes are recorded in patients' treatment dataset and demographic information. In this paper, we consider patients' gender, age, admission date, diagnosis name and treatment outcome as similarity measurement. Thus, we choose gender, age and admission date to be three quasi-identifiers, because these three preferences have been shown to lead to user re-identification.

Our method derives from a recently globally optimal *k*-anonymity method [Emam, Dankar, Issa et al. (2009)], which is called *Optimal Lattice Anonymization (OLA)*. The advantage of OLA is that it results in less information loss and has faster performance in medical dataset compared to the current de-identification algorithms.

A common way to satisfy the *k*-anonymity criterion is to generalize values in the quasi-identifies by reducing their precision. Examples of hierarchies can be represented in Fig. 1. The precision of variables is reduced as one move up the hierarchy.

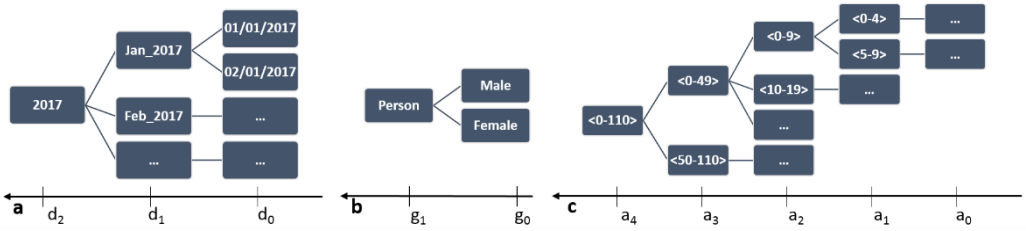

Figure 1: Examples of value generalization hierarchies for three common quasi-identifies: (a) Admission date (b) Gender (c) Age

The generalization hierarchies for the three quasi-identifies in Fig. 1 can be represented as a lattice. Each node in the lattice represents a possible instance of the dataset. One of these nodes is the globally optimal solution and the objective of a *k*-anonymous algorithm is to find it efficiently.

An information loss metric that takes into account is *suppressed rate*, which is defined as (1):

$$suppressed\ rate = \frac{\sum_i \delta(ec(i))}{EC} \tag{1}$$

Where $\delta(ec(i)) = 0$ if the equivalence class $ec(i) < k$, $\delta(ec(i)) = 1$, otherwise; EC is the total number of equivalence class.

All equivalence classes in the dataset that are smaller than *k* are suppressed. 85% of records were suppressed in the dataset represented by node $< d_0, g_0, a_0 >$ because these records were in small equivalence classes. As more generalization is applied, the extent of suppression goes down.

Suppression is preferable to generalization because the former affects single records whereas generalization affects all the records in the datasets. However, because of the negative impact of missing on the ability to perform meaningful data analysis, the limits on the amount of suppression need to be imposed. We present this limit as *MaxSup*. In this paper, we define the *MaxSup* as 5%, and all the nodes that satisfy "*suppressed rate*<*MaxSup*" criterion are *k*-anonymous nodes.

Once we have identified all the *k*-anonymous nodes, we need to select the one with the least information loss among them. Suppressed rate is not considered a good information loss metric because it does not account for the generalization hierarchy depths of the quasi-identifiers. For example, the generalization of "Male" to "Person" gives the equal weight to the generalization of age in one year to age in five years. In the former case, there is no information left in the gender variable, whereas the age interval still conveys a considerable amount of information and there are three more possible generalizations left in the age hierarchy.

Hence the gender information plays an important role for clinician to diagnose and prescribe, in this paper we choose g_0 as the gender variable. In this case the k-anonymous nodes are: $< d_2, g_0, a_1 >$, $< d_2, g_0, a_2 >$, $< d_2, g_0, a_3 >$, $< d_2, g_0, a_4 >$. We maintain the list of the four anonymous nodes and select the node $< d_2, g_0, a_1 >$ with the lowest height within their generalization strategies as the globally optimal

solution.

3.2.2 Similarity measure for patients

In this section, we give the definition of similarity measurement between patients. As is mentioned before, we consider patients' gender, age, admission date, diagnosis name and treatment outcome as similarity measurement. A patient can be formalized as (2):

$$P = \{P^G, P^A, P^D, P^{Diag}, P^O\} \quad (2)$$

In Section 3.2.1, we process the generalization of three quasi-identifies (Gender, Age and Admission Date) and select $< d_2, g_0, a_1 >$ as the globally optimal solution. Thus the gender of a patient PG can be "male" or "female", the age P^A can be "0-4", "5-9", "10-14", etc. the admission date P^D can be "2017", "2016", "2015", etc.

Diagnosis information P^{Diag} is given by doctors and consists the name of diseases. Outcome is evaluated and presented by doctors when a patient leaves hospital. An outcome of a patient can be "cured", "improved", "ineffective" or "dead". We use P^O to present the outcome of a patient in this paper.

In order to be easily understood, we present a toy example of quintuple *P* by Tab. 1.

Table 1: A toy example of *P*

P^G	P^A	P^D	P^{Diag}	P^O
female	55-59	2016	chronic gastritis	cured
male	0-4	2017	bronchopneumonia	Cured
male	60-64	2017	cerebral infarction	dead

In order to define similarity between different patients, we have to develop a method which can compute similarity between two such quintuples. P_i and P_j are represented two random selected patients, the similarity between P_i and P_j is defined as following:

Firstly, the similarity between P_i and P_j is determined by the diagnosis names P^{Diag}, if the diagnosis names of two patients are the same, then gender, age and admission Date are considered in a further step; otherwise, the similarity of P_i and P_j is set 0. Therefore, the similarity of P_i and P_j contains a multiplying term $\delta(P_i^{Diag}, P_j^{Diag})$, which equals 1 if P_i^{Diag} and P_j^{Diag} are the same, and equals 0 otherwise.

Secondly, the gender should be taken into account. The similarity between two P^G is described by term $\eta(P_i^G, P_j^G)$, which is 1 if two genders are the same, and equals 0.5 otherwise.

Thirdly, the age is considered. In this paper, P^A is indicated as interval, such as <5-9>. There are twenty intervals in all from <0-4> to <95-99>. We flag this intervals as sequence numbers P^{AI} from 0 to 20. The similarity between two PA is defined as (3),

$$\theta\left(P_i^A, P_j^A\right) = \begin{cases} 1 & if\left|P_i^{AI} - P_j^{AI}\right| < 2 \\ 0.5 & if 2 \leq \left|P_i^{AI} - P_j^{AI}\right| < 8 \\ 0.25 & if\left|P_i^{AI} - P_j^{AI}\right| \geq 8 \end{cases} \quad (3)$$

Lastly, the admission date also has large impact on the similarity determination. In this paper, admission date is generalized into year and is from 2008 to 2017, so the similarity between two P^D is defined by $\gamma(P_i^D, P_j^D)$, which is 1 if $\left|P_i^D - P_j^D\right| < 3$, and equals 0.5 if $3 < \left|P_i^D - P_j^D\right| \leq 6$, and equals 0.25 otherwise.

To sum up, similarity between Pi and Pj is finally defined as (3),

$$s\left(P_i, P_j\right) = \frac{\delta\left(p_i^{Diag}, p_j^{Diag}\right)[\eta\left(p_i^G, p_j^G\right) + \theta\left(p_i^A, p_j^A\right) + \gamma(p_i^D, p_j^D)]}{3} \quad (4)$$

The denominator 3 is to ensure the value of similarity drops in [0, 1].

3.2.3 Nearest neighbor selection

The goal of this section is to extract the most similar neighbors for the target patient. Given an active patient pa and his candidate neighbor list *P*. Firstly, we take into account the neighbor's outcome P^O in *P*. To guarantee that the treatment is effective, only the candidate neighbor patients who have the "cured" and "improved" outcome have been chosen. Secondly, we compute the similarity between p_a and other positive-outcome patient. In Section 3.2.2, we have given the definition of the similarity formula between two patients. So we get a corresponding similarity list $s(p_a) = \{s(a, 1), \cdots, s(a, m)\}$, which consists of similarities between p_a and other m positive-outcome patient. Finally, we choose the most similar *k* neighbors to form the KNN list $N(p_a)$ from the candidate list based on the similarity list $s(p_a)$.

3.3 Differential privacy recommendation

In Section 3.2, we have proposed an effective and secure-safe similarity measure between patients to find the most similar neighbors. In this section, we propose a neighborhood-based differential privacy recommendation algorithm, with the aim of proving comprehensive privacy for the individual patient, as well as maximizing the accuracy of recommendations.

3.3.1 Neighborhood-based recommend inference attack

For each patient $p_a \epsilon P$, a neighborhood-based recommendation algorithm publishes the related recommended medicine list $Q(M, p_a)$ based on his/her neighbor's medical records. For example, some hospital publishes the related list *L* of each medicine's usage. Supposing an attack knows some auxiliary information about a target patient p_a, usually some part of the medication administration record $\mathbb{R}_{p_a} = \{R_{p_{a1}}, R_{p_{a2}}, \ldots\}$. Suppose the target patient p_a interacts with the system with the time period $[t_1, t_2]$ and take the medicine *m*, which results in *L*. The covariance between *m* and all items in *L* must increase. Thus, the attacker can infer the purchasing activity of p_a by observing these related lists of medicines in *L*. If the same medicine *m* appears or move up in the related

lists of a sufficient large subset of the auxiliary medicines, the attacker can infer that p_a take the medicine *m*.

3.3.2 User-based differential privacy recommendation algorithm

Differential Privacy tends to maximize the accuracy of the output of the system while minimizing the chances of identifying the input to the system.

Definition 1. (ϵ*-differential privacy*) A randomized function *M* give ϵ-differential privacy if for any datasets D_1 and D_2 differing on at most one element (adjacent dataset), and any $S \subseteq \text{Range}(M)$, the following mathematical definition holds in (5):

$$\Pr[M(D_1) \in S] \leq e^{\epsilon} \times \Pr[M(D_2) \in S] \tag{5}$$

Differential privacy provides a bound on the ability to infer from any output *S*, whether the input to the computation was D_1 or D_2. Because:

$$\frac{pr(D_1|S)}{pr(D_2|S)} = \frac{pr(D_1)}{pr(D_1)} \times \frac{pr(S|D_1)}{pr(S|D_2)} \tag{6}$$

When $D_1 \approx D_2$, differential privacy limits the degree of inference possible about the input of system, and bounds it by a factor of $exp(\epsilon)$.

Definition 2. (*Exponential Mechanism*) Given a quality function *q*: $(R^{|U|\times|I|}, I) \to \mathbb{R}$, an input matrix M, Δ_q is the sensitivity of the quality function. The exponential mechanism $\text{expo}(M, I, q, \epsilon)$ outputs $i \in I$ with probability:

$$\Pr[\text{expo}(M, I, q, \epsilon) = i] = \frac{\epsilon q(M,i)/(2\Delta_q)}{\sum \epsilon q(M,i)/(2\Delta_q)} \tag{7}$$

satisfies ϵ-differential privacy.

In definition 1, the parameter ϵ refers to the privacy budget, which controls the level of privacy guarantee. A smaller ϵ represents a stronger privacy level.

Based on the above notions, we design a neighborhood-based Differential Privacy Recommendation algorithm. It consists of three steps:

- Given the target patient pa with his/her k nearest neighbors $N(p_a)$. We sample each patients' medication administration record in patient treatment dataset T, and form a clinical medicine score function $Q(m_i, p_a)$ to count the dosage of each drug mi used on the neighbor candidate list.

$$Q(m_i, p_a) = \sum_{p_k \in N_{p_a}} \zeta(m_i, p_k) \tag{8}$$

$$\zeta(m_i, p_k) = \begin{cases} 1 & if\ (p_k, m_i) \in T \\ 0 & if\ (p_k, m_i) \notin T \end{cases} \tag{9}$$

- Then we design a differential privacy algorithm by adding an exponential noise mechanism in $Q(m_i, p_a)$. Define $Q'(m_i, p_a)$ to be $Q(m_i, p_a) + exp^{(\frac{\varepsilon Q(m_i, p_a)}{2\Delta q})}$, which provides ϵ-differential privacy.
- We recommend top *n* score medicines to patient p_a according to the privacy-preserving medicine score function $Q'(m_i, p_a)$.
-

4 Experiment

The details of dataset and experimental setup used to evaluate the proposed privacy preserving medical recommendation algorithm is represented in the following subsections.

4.1 Real datasets

The dataset we experimented with are collected from Hospital Information Systems (HIS) of the First Affiliated Hospital of Xi'an Jiaotong University, which is the largest Grade Three Class A hospital in northwest China. The dataset contains 115,585,623 medical treatment records with 3,944 medicines and 946,429 patients from the year 2008 to 2017, where about 51% of the patients are male and rest are female, the age of patients is in between 0 year to 104 years.

The 946,429 patients are divided such that 662,500 (70%) patients are part of training set and 283,929 (30%) patients are part of the testing set. The predictions by the PPMR are compared with the actual treatment labelled by the medical expert to check the accuracy of the results.

4.2 Accuracy measures

To measure the quality of recommendation, in this paper, three performance metrics are used to evaluate the efficiency of our proposed algorithm: *Recall, Precision and Mean Absolute Error (MAE)* [Adomavicius and Tuzhilin (2005)]. These metrics are widely accepted for evaluating recommender systems. The Recall rate is defined as the ratio of the products used by the users in the recommending list to all the products that the user actually uses. The Accuracy rate is the ratio of products that users end up using in the recommended list to all recommended products. MAE is used for computing the deviation between the predicted and the real ratings. Note that lower values in MAE mean better recommending predictions. The three metrics is shown in (10), (11) and (12).

$$Recall = \frac{N_{rs}}{N_r} \tag{10}$$

$$Precision = \frac{N_{rs}}{N_s} \tag{11}$$

$$MAE = \frac{1}{|T|}\sum_{a,i\in T}|r_{ai} - \hat{r}_{ai}| \tag{12}$$

where N_{rs} is the number of the products used by the users in the recommending list, N_r is the number of all the products that the user actually uses, N_s is the number all recommended products. r_{ia} is the true rating of user u_a on the item t_i, and $\hat{r}_{ai}$ is the value of predicting the rating.

4.3 Performance of PPMR

In this section, we examine the performance of PPMR from the perspective of privacy preserving to the patients' information. Specifically, we apply the traditional neighborhood-based CF as the non-private baseline. The top score parameter *n* is set to be 20 because most patients have taken about twenty kinds of medicine during their treatment process. Moreover, the privacy budget ϵ is used to control the level of privacy,

so it need to be set to balance the privacy level and recommend accuracy. The algorithm proposed in this paper can be used to recommend medicine for various diseases. To illustrate and test our algorithm, we focus on the three kinds of patients with *coronary heart disease and pneumonia*, which are the most common diseases in China today.

4.3.1 Impact of parameter ϵ

Tab. 2 illustrates the effects of privacy budget ϵ on *coronary heart disease* and *pneumonia*. From the Tab. 2, we can see that when ϵ is in a larger value (for example ϵ =0.001), the probability of the best available medicine is amplified. On the other hand, when ϵ is small (for example ϵ=0.00001), the differences in usability for every medicines are suppressed and the probability of the medicine output tend to be equal. In this paper, the privacy budget parameter ϵ is fixed to 0.0001 to ensure the PPMR algorithm satisfies the *0.0001-differential privacy*, which could balance the privacy level and data accuracy.

Table 2: Effects of privacy budget ϵ on coronary heart disease and pneumonia

Coronary Heart Disease									
$Q(m_i, p_u)$	Probability (△q=1)				$Q(m_i, p_u)$	Probability (△q=1)			
	NonP	ϵ=0.001	ϵ=0.0001	ϵ=0.00001		NonP	ϵ=0.001	ϵ=0.0001	ϵ=0.00001
(1)36591	9.97%	94.27%	11.52%	5.47%	(11)16669	4.54%	0.00%	4.25%	4.95%
(2)29404	8.01%	2.59%	8.04%	5.28%	(12)15650	4.26%	0.00%	4.04%	4.93%
(3)28514	7.77%	1.66%	7.69%	5.26%	(13)14984	4.08%	0.00%	3.91%	4.91%
(4)27686	7.54%	1.10%	7.38%	5.24%	(14)13480	3.67%	0.00%	3.63%	4.88%
(5)24091	6.56%	0.18%	6.16%	5.14%	(15)11522	3.14%	0.00%	3.29%	4.83%
(6)22807	6.21%	0.10%	5.78%	5.11%	(16)11417	3.11%	0.00%	3.27%	4.83%
(7)21991	5.99%	0.06%	5.55%	5.09%	(17)10864	2.96%	0.00%	3.18%	4.81%
(8)17962	4.89%	0.01%	4.54%	4.99%	(18)10712	2.92%	0.00%	3.16%	4.81%
(9)17500	4.77%	0.01%	4.43%	4.98%	(19)9258	2.52%	0.00%	2.94%	4.77%
(10)17269	4.71%	0.01%	4.38%	4.97%	(20)8658	2.36%	0.00%	2.85%	4.76%
Pneumonia									
$Q(m_i, p_u)$	Probability (△q=1)				$Q(m_i, p_u)$	Probability (△q=1)			
	NonP	ϵ=0.001	ϵ=0.0001	ϵ=0.00001		NonP	ϵ=0.001	ϵ=0.0001	ϵ=0.00001
(1)12536	16.68%	69.98%	7.66%	5.22%	(11)2132	2.84%	0.39%	4.55%	4.96%
(2)8841	11.76%	11.03%	6.37%	5.13%	(12)2053	2.73%	0.37%	4.54%	4.96%
(3)7669	10.20%	6.14%	6.01%	5.10%	(13)1970	2.62%	0.36%	4.52%	4.95%
(4)6481	8.62%	3.39%	5.66%	5.07%	(14)1654	2.20%	0.30%	4.45%	4.95%
(5)5820	7.74%	2.43%	5.48%	5.05%	(15)1576	2.10%	0.29%	4.43%	4.95%
(6)4289	5.71%	1.13%	5.07%	5.01%	(16)1570	2.09%	0.29%	4.43%	4.95%
(7)4240	5.64%	1.11%	5.06%	5.01%	(17)1459	1.94%	0.28%	4.40%	4.94%
(8)3128	4.16%	0.63%	4.79%	4.98%	(18)1423	1.89%	0.27%	4.40%	4.94%
(9)3064	4.08%	0.61%	4.77%	4.98%	(19)1411	1.88%	0.27%	4.39%	4.94%
(10)2587	3.44%	0.48%	4.66%	4.97%	(20)1275	1.70%	0.25%	4.36%	4.94%

4.3.2 Experiment results and analysis

Tab. 3 shows the Recall, Precision and MAE of *coronary heart disease* and *pneumonia*

patients' recommendation in the two algorithm, i.e. he proposed PPMR algorithm and the traditional neighborhood-based CF, with *k* changing. Parameter *k* demotes the number of nearest neighbors. Here, *k* could be an integer from 100 to 1600. From Tab. 3, we can see that, with *k* increase, the Recall and Precision increase, and MAE decreases at first, but when *k* surpasses a certain threshold, the Recall and Precision decrease and the MAE increases with further increases in the value of *k*. We can observe that, the Precision, Recall and MAE achieves the best performance when *k* is around 800, while smaller values like *k*=400 or larger value *k*=1600 can potentially degrade the performance.

In addition, we compared the recommendation quality of PPMR algorithm and the traditional neighborhood-based CF to derive the predicted ratings on the medicines. It is discovered that on both Coronary Heart Disease and Pneumonia, the performance of PPMR is very close to that of the non-private baseline with *no more than 5% accuracy loss*. This indicates PPMR can retain the accuracy of recommendation while providing comprehensive privacy for individuals.

Table 3: Comparison on EMRs datasets

	Coronary Heart Disease					
k	MAE		Precision		Recall	
	Non-private	PPMR	Non-private	PPMR	Non-private	PPMR
100	0.7504	0.7635	0.2571	0.2501	0.2083	0.2079
200	0.7114	0.7407	0.2610	0.2613	0.2338	0.2311
400	0.7087	0.7182	0.2831	0.2819	0.2410	0.2385
800	0.7071	0.7208	0.3211	0.3113	0.2627	0.2626
1600	0.7078	0.7211	0.3200	0.3124	0.2629	0.2620
	Pneumonia					
k	MAE		Precision		Recall	
	Non-private	PPMR	Non-private	PPMR	Non-private	PPMR
100	0.7912	0.8018	0.2310	0.2242	0.2110	0.2098
200	0.7614	0.7691	0.2564	0.2533	0.2313	0.2286
400	0.7527	0.7527	0.2678	0.2600	0.2465	0.2409
800	0.7478	0.7481	0.2680	0.2675	0.2511	0.2505
1600	0.7481	0.7531	0.2597	0.2501	0.2509	0.2500

5 Conclusion

Privacy preserving is one of the most essential aspects of collaborate filtering as it protests the sensitive information of users in recommendation systems. In clinical environment, privacy preserving problem is more important since the healthcare data of patients involves high personal and sensitive nature.

This paper proposes an effective privacy preserving method for neighborhood-based collaborative filtering and makes the following contributions:

- Private Neighbor Selection algorithm is provided to prevent the patients' healthcare information from being attacked. In addition, a new de-identification *k*-anonymity method is adopted to produce a globally optimal de-identification solution suitable for EMRs datasets.
- A novel Neighborhood-based Differential Privacy Recommendation Algorithm is

proposed to provide privacy protection for patients and maximize the accuracy of recommendation at the same time.

- The security analysis and performance evaluation is carried out. Experimental results show the effectiveness and robustness of the proposed PPMR algorithm in various metrics.

Most notably, to the best of our knowledge, this is the first study to investigate the privacy preserving collaborative filtering in medical recommendation. It has been proven that our algorithm can guarantee a better quality of recommendation accuracy. However, the current study only concentrates on the privacy of neighborhood-based CF. Other recommendation techniques, such as Matrix Factorization, still suffer from privacy problem. Therefore, future work should consider the privacy issue for other recommendation techniques.

Acknowledgement: We thank the valuable comments from our reviewers and editors. This work is supported by the Fundamental Research Funds of the First Affiliated Hospital of Xi'an Jiao Tong University (No. 2017RKX-06).

References

Adomavicius, G.; Tuzhilin, A. (2005): Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering*, vol. 17, no. 6, pp. 734-749.

Berkovsky, S.; Eytani, Y.; Kuflik, T.; Ricci, F. (2007): Enhancing privacy and preserving accuracy of a distributed collaborative filtering. *Conference on Recommender Systems*, pp. 9-16.

Das, A. S.; Datar, M.; Garg, A.; Rajaram, S. (2007): Google news personalization: Scalable online collaborative filtering. *International Conference on World Wide Web*, pp. 271-280.

Davidson, J.; Liebald, B.; Liu, J.; Nandy, P.; Vleet, T. V. et al. (2010): The youtube video recommendation system. *ACM Conference on Recommender Systems*, pp. 293-296.

Dwork, C. (2006): Differential privacy. *Lecture Notes in Computer Science*, vol. 26, no. 2, pp. 1-12.

El Emam, K.; Dankar, F. K.; Issa, R.; Jonker, E.; Amyot, D. et al. (2009): A globally optimal *k*-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association Jamia*, vol. 16, no. 5, pp. 670.

Enck, W.; Gilbert, P.; Chun, B. G.; Cox, L. P.; Jung, J. et al. (2012): Taintdroid: An information flow tracking system for real-time privacy monitoring on smartphones. *ACM Transactions on Computer Systems*, vol. 32, no. 2, pp. 1-29.

Han, J.; Qian, C.; Yang, P.; Ma, D.; Jiang, Z. et al. (2016): Geneprint: Generic and accurate physical-layer identification for UHF RFID tags. *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 846-858.

Hardt, M.; Roth, A. (2011): Beating randomized response on incoherent matrices. *Proceedings of the Annual ACM Symposium on Theory of Computing*, pp. 1255-1268.

Hirano, S.; Tsumoto, S. (2014): Mining typical order sequences from EHR for building clinical pathways. *Trends and Applications in Knowledge Discovery and Data Mining*, vol. 8643, pp. 39-49.

Lakkaraju, H.; Rudin, C. (2016): Learning cost-effective and interpretable regimes for treatment recommendation. *Machine Learning*.

Li, D.; Lv, Q.; Xia, H.; Shang, L.; Lu, T. et al. (2011): Pistis: A privacy-preserving content recommender system for online social communities. *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 79-86.

Liu, C.; Wang, F.; Hu, J.; Xiong, H. (2015): Temporal phenotyping from longitudinal electronic health records: A graph based framework. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 705-714.

Mcsherry, F.; Mironov, I. (2009): Differentially private recommender systems: Building privacy into the net. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 627-636.

Polatidis, N.; Georgiadis, C. K.; Pimenidis, E.; Mouratidis, H. (2017): Privacy-preserving collaborative recommendations based on random perturbations. *Expert Systems with Applications*, vol. 71, pp. 18-25.

Sun, L.; Liu, C.; Guo, C.; Xie, Y.; Xie, Y. (2016): Data-driven automatic treatment regimen development and recommendation. *International Conference on Knowledge Discovery and Data Mining*, pp. 1865-1874.

Wondracek, G.; Holz, T.; Kirda, E.; Kruegel, C. (2010): A practical attack to de-anonymize social network users. *Security and Privacy*, vol. 41, pp. 223-238.

Zhan, J.; Hsieh, C. L.; Wang, I. C.; Hsu, T. S.; Liau, C. J. et al. (2010): Privacy-preserving collaborative recommender systems. *IEEE Transactions on Systems Man & Cybernetics Part C*, vol. 40, no. 4, pp. 472-476.

Zhang, Y.; Chen, R.; Tang, J.; Stewart, W. F.; Sun, J. (2017): LEAP: Learning to prescribe effective and safe treatment combinations for multimorbidity. *International Conference on Knowledge Discovery and Data Mining*, pp. 1315-1324.

Zhou, J.; Wang, F.; Hu, J.; Ye, J. (2014): From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records. *International Conference on Knowledge Discovery and Data Mining*, pp. 135-144.

Zhu, T.; Ren, Y.; Zhou, W.; Rong, J.; Xiong, P. (2014): An effective privacy preserving algorithm for neighborhood-based collaborative filtering. *Future Generation Computer Systems*, vol. 36, no. 36, pp. 142-155.