ARM-Based Privacy Preserving for Medical Data Publishing

Zhang Fengli(✉) and Bai Yijing

School of Information and Software Engineering,
University of Electronic Science and Technology of China,
Chengdu, China
yijing_llll@163.com

Abstract. The increasing use of electronic medical records (EMR) makes the medical data mining becomes a hot topic. Consequently, medical privacy invasion attracts people's attention. Among these, we are particularly interested in the privacy preserving for association rule mining (ARM). In this paper, we improve the traditional reconstruction-based privacy preserving data mining (PPDM) and propose a new architecture for medical data publishing with privacy preserving, and we present a sanitization algorithm for the sensitive rules hiding. In this architecture, the sensitive rules are strictly controlled as well as the side effects are minimized. And finally we performed an experiment to evaluate the proposed architecture.

Keywords: Medical data · Privacy preserving data mining · Association rule mining

1 Introduction

As the size of healthcare data increase dramatically, more and more medical information system are using digitalized technology to realize the storage of the big data. The electronic form of healthcare data is called EMR (electronic medical records). The storage of these information are all precious wealth of human being. These medical information resources are very valuable for disease treatment, diagnosis and medical research. And data mining, that is to find the valuable knowledge hidden in these massive medical data resources, has become a very important research topic. One of the typical application conditions of data mining is association rule mining(ARM). ARM is to find a rule set such that the support and confidence value of each rule is bigger than the giving threshold.

Despite that a great number of valuable knowledge are discovered by ARM, these rules may include some privacy information for the patients and medical department, people have shown increasing concern about privacy violation brought by the technology. Because there are person-specific information contained in the medical system, publish of data will cause unconscious privacy leakage, which may bring some bother to victim. Typically privacy information can be classified into two categories [1]: one is that you can get directly from the original data, which can be protected by the methodologies like perturbation, sampling, generalization/suppression, transformation,

Z. Huang et al. (Eds.): ICCCS 2015, LNCS 9483, pp. 62–73, 2015.
DOI: 10.1007/978-3-319-27051-7_6

or anonymity technology etc. The other is sensitive knowledge patterns, which is hidden in the data. You can get them only in the results of data mining. To avoid the disclosure, you have to use privacy-preserving data mining (PPDM).

Take the medical data publishing as example: a typical EMR(Electronic Medical Record) template in China contains <Name, ID Number, Gender, Age, Birth Place, Nationality, Symptoms, Physical Examination, Diagnosis,…>. We suppose there is a medical institution negotiating with a medical factory. The medical factory offers the institution with a reduced-price products if they publish their database of EMR to the factory. The medical institution accepts the deal and preprocesses data to hide the sensitive information like name, ID number etc. However factory starts mining the data using ARM, they find a woman who bought Zyban is most likely pregnant. The female customers who has bought Zyban in the medical factory would receive some promotion of maternal medicine from the factory, and some of them are indeed pregnant, then they will feel a strong privacy invasion. And how to hide these sensitive association rules in order to avoid this kind of privacy invasion is this paper mainly focuses on.

Privacy preserving of ARM is to process the original medical dataset to get a new dataset, so that the sensitive association rules cannot be mined from the new dataset but all the non-sensitive rules can still be mined from it.

To complete the above process, we propose a new architecture of privacy preserving for medical data publishing. The rest of this paper will be organized as follows: we describe the related work in Sect. 2. Section 3 provides the system architecture we proposed and some preliminary we use in this paper. The specific procedure of the architecture and all of the algorithm details in each step are described in Sect. 4. Section 5 discusses performance evaluation of our architecture and algorithms.

2 Related Works

Since the concept of PPDM was first proposed in 1999 by Agrawal [2], there were a great number of achievements in this field by now [3–5]. Data mining technology is an inter-discipline includes lots of data analysis techniques like statistics and machine learning, and hence the diversity of privacy-preserving techniques on it.

Lei Chen identifies the incompatibilities between the traditional PPDM and the typical free text Chinese EMR. And he proposed a series of new algorithms to solve the problem in [6]. Then he also designed a new framework of privacy preserving for healthcare data publishing based on the method in [7]. There are also some achievements for medical data protection in [8, 9].

Typically, privacy-preserving technology in the field of data mining can be divided into two categories. For one kind is to protect the sensitive data itself, like name, ID number. For medical data, the Health Insurance Portability and Accountability Act (HIPAA) was enacted in US in 1996 [10]. The Act announced the personal health information privacy standards and guidelines for implementation. There are many kinds of technologies to protect it. The most common one is anonymization, including k-anonymity [11], l-diversity [12], t-closeness [13]. In [14] Aristides Gionis and Tamir Tassa extended the framework of k-anonymity to include any type of generalization operators and define three measures to count the loss of information more accurately.

They also proved that the problem of k-anonymity with minimal loss of data is NP-hard. And then proceed to describe an approximation algorithm with an approximation guarantee of O(ln k).

Another kind of category is to protect the sensitive data mining results that were produced in the process of data mining [15]. Generally, these techniques are focusing on the improvement of data mining algorithms. Among them, we are particular interested in approaches proposed to perform association rule hiding [16, 17]. In [18] a new algorithm for hiding sensitive rules using distortion techniques are proposed by Jain. The confidence of the sensitive rules are reduced by altering the position of them. In [19] Le proposed a heuristic algorithm to identify the items which has the least impact on the mining results, and remove these items from the original data set. There are also many achievements in [20–22]. In [23] Verykios proposed three groups of algorithms to hide the sensitive rules based on reducing the support and confidence values, which can be seen as the precursors to the algorithms we proposed in this paper.

3 Arm-Based System Architecture of Privacy-Preserving for Medical Data Publishing

As showed in Fig. 1, in our architecture, there are three types of data characters: data owner, data collector, and data user. Data owner provides the original medical dataset and privacy definition and policy; data collector processes ARM and sanitization algorithm and publishes new dataset to data user. And we assume that data collector and network communications are reliable.

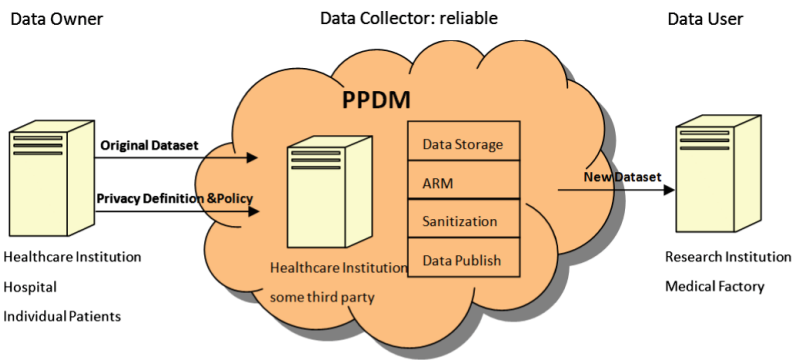

Fig. 1. ARM-based system architecture of privacy preserving for medical data publishing

Data owner could be a healthcare institution, hospital or individual patients, who has a collection of medical data provided to data collector. We use notation $D = \{T_1, T_2, T_3, \ldots, T_m\}$ to represent original medical dataset, and $I = \{i_1, i_2, i_3, \ldots, i_n\}$ to represents a set of items, $T_i \subseteq I$ is a transaction. We use TID to identify every transaction.

Data collector could be a healthcare institution or some third party(sometimes data owner and data collector could be the same one). Data collector's work includes: data storage, data processing(ARM and sanitization algorithm), and data publish. In this paper, we suppose that data storage and data publish are finished, and we only consider the data processing.

Data collector applied ARM algorithm on the dataset provided by data owner. We use r or $X => Y$ to denote a rule, where $X \subset I, Y \subset I$, and $X \cap Y = \emptyset$. Given an itemset X, support of an itemset is the number of transactions that contains the itemset, which is denoted as s(X). $X => Y$ means the number of occurrence of Y on condition of occurrence of X. Given an rule r(or $X => Y$), X is denoted as r_l, and Y is r_r, support of the rule is the number of transactions that contain both X and Y, $s(X => Y) = s(X \cup Y)$. Confidence of a rule is the frequent of itemsets that contain Y appear in transactions that contain X, $c(X => Y) = P(Y|X) = P(X \cup Y)/P(X)$. We define two threshold: α for minimum support value, ε for minimum confidence value. And ARM is to find all the rules that satisfy $s(X => Y) \geq \alpha$, $c(X => Y) \geq \varepsilon$. We use $R = \{r|\text{all rules such that } s(r) \geq \alpha \text{ and } c(r) \geq \varepsilon\}$ to denote the mining result.

Then data collector identifies the privacy rules according to the privacy definition and policy. Here let Rh be a set of sensitive rules to be hidden, $Rh \subset R$.

And finally data collector clean Rh directly from original dataset by sanitization algorithm. For a rule $r \in R$, we use Tr to denote a set of transactions that r appears, $Tr = \{t \in D | r \subseteq t\}$. If there exists a rule $r' \in R$ such that, the number of items appears both in r and r′ is larger than $|r|/2$ and $|r'|/2$, as well as r and r′ both appear in one or more transaction, $Tr \cap Tr' \neq \emptyset$, we call r and r' are brother rules. We use Rb to denote a set of brother rules, $Rb = \{r' | r' \in R, r' \text{ and } r \text{ are brother rules}\}$. After the sanitization algorithm, we got a new data set $D' = \{T'_1, T'_2, T'_3, \ldots, T'_n\}$. D′ is a "clean" dataset that satisfy the condition: no rules in Rh can be mined from D′, but all rules in (R-Rh) can still be mined from it.

Data user could be research institution, pharmaceutical factory etc. We consider data user is unreliable. The new dataset D′ is the final dataset to be published to data user.

4 Processing of the Archetecture

The process of the architecture is showed in Fig. 2. Data owner provides original datasets D and privacy definition and policy. Data collector is the main part of the procedure who transforms the original datasets D into a "clean" datasets D′ by applying a series a privacy-preserving algorithms, and finally publishes D' to data user.

The procedure of the framework is as follows:

step 1: Data owner provides the original datasets $D = \{T_1 T_2, T_3, \ldots, T_m\}$ and privacy definition and policy to data collector.
step 2: Data collector uses ARM algorithm to generate association rule set R.
step 3: Data collector identifies corresponding sensitive association rules Rh for different data users according to the privacy definition and policy.

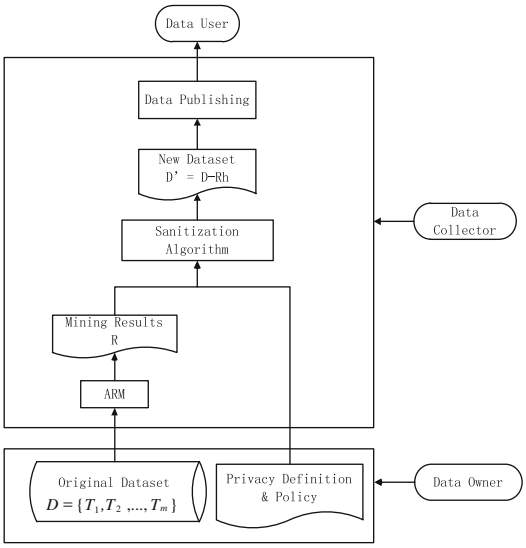

Fig. 2. Process of architecture

step 4: Data collector performs sanitization algorithm to hide the rules in Rh from the dataset.
step 5: Step 2 to step 5 is repeated until there are no sensitive rules could be found in the mining results. The rest of the dataset are "clean".
step 6: The new dataset D' with no sensitive patterns will be published to data user.

As depicted in Fig. 2, we do not use a reconstruction algorithm after sanitization as traditional framework does. Because it has been proven that the relationship between the transactions and their k-itemsets is one to one correspondence [24]. That is to say, direct modification in dataset D during sanitization algorithm, or cleaning sensitive rules from k-itemsets and then reconstructing a new dataset D′ using the clean k-itemsets, we can get the same result. In order to improve the efficient of our architecture, we hide rules directly from dataset D.

4.1 Applying Framework on Medical Data

A typical EMR(Electronic Medical Record) template in China contains <Name, ID Number, Gender, Age, Birth Place, Nationality, Symptoms, Physical Examination, Diagnosis,…>. For most of healthcare institution, the datasets stored in the system are in free text form instead of structured data. In order to apply the architecture on datasets, we shall label the EMRs at first place. Because our algorithms process involve only the patterns of datasets rather than specifics, it's OK to use just numbers to label the records.

Take the heart diseases diagnosis dataset as example, few samples of labeling EMRs is described in Tables 1 and 2.

Table 1. Sample of physical examination and Symptoms labeling

Code	Physical examination	Code	Symptoms
1	High blood cholesterol	31	Palpitation
2	High blood pressure	32	Dizzy
3	Subclinical atherosclerosis	33	Chest tightness
…	…	…	…

Table 2. Sample of age and diagnosis labeling

Code	Age	Code	Diagnosis
51	[0,5)	71	No heart disease
52	[5,10)	72	Heart disease
…	…	…	…

After the labeling, an item can be transformed into a transaction format as shown in the Table 3. The first column represents transaction ID, namely the patient number. The second column represents all the items recorded in his(her) EMR. The set of items I = {1, 2, 3, 4, 5, 6,…}. Our medical datasets can be transformed into a transaction format $D = \{T_1T_2, T_3, \ldots, T_m\}$. After the labeling, our framework can be easily used on the medical data.

4.2 Sanitization Algorithm

After ARM process, we have the association rule set R and the support, confidence values of each rule. In this part, we will use the sanitization algorithm to hide the sensitive patterns Rh by altering the support or confidence values of related items. The algorithm is depicted in the following codes. It starts by generating the transaction set Tr where the rule r comes from, and sort Tr in ascending order of transaction size |t|. Then it determines whether there exists brother rules of r in Rh. If there are brother rules, it would compute the loop_num of each rule in Rb and choose one of them to represent the Rb to be hidden in order to minimize the impact on the database. If there are no brother rules in Rh, it would compute the loop_num of rule r, and choose the smallest transaction t in Tr. Then it deletes the r_r of rule r from t. When the rule r is hidden, another rule is selected from set Rh. In the following codes, symbol [x) means the minimal integer that larger than but not equal to x.

Table 3. Sample of datasets D

TID	Records
1	1-2-31-55-72
2	1-2-3-34-58-72
3	5-11-40-64-71
…	…

```
Algorithm: Sanitization algorithm
Input: D, Rh, α, ε
Output: D'
Begin
  Foreach rule R ⊆ Rh do
  {
    Tr={t∈D| R⊆ t}
    //sort Tr in ascending order of |t|
    sort(Tr)
    if there exists brother rules of r in Rh
    {
      Tr = { t∈D| ∀r ∈ Rb , r⊆ t}
      Foreach rule r ∈ Rb do
      {
        Loop_supp = [S(r) - α)
        Loop_conf = [S(r) - ε*S(rl))
        Num = min(Loop_supp, Loop_conf)
      }
      Loop_num = maximal Num among Rb
      R' = {r| r ∈ Rb, r.Loop_num is the largest}
      if |R'| >= 2
        R is the one in R' who has the shortest rr
      else
        R = R'
        delete(Rb-{R}, Rh)
      }
    else
    {
      Loop_supp = [S(r) - α)
      Loop_conf = [S(r) - ε*S(rl))
      Loop_num = min(Loop_supp, Loop_conf)
    }
    For i=1 to Loop_num do
    {
      t = pop(Tr)
      delete(R. rr, t)
      delete(t, Tr)
    }
    delete(R, Rh)
  }
end
```

In our architecture, the sanitization algorithm will be repeated until there exists no sensitive patterns in the discovered rules. This will ensure that no unexpected information from Rh will be published to unreliable party.

5 Performance Evaluation

In this part, we perform our framework on a computer running windows server 2008 R2 operating system. The dataset we use is generated by ourselves as shown in the Table 4. In order for a more realistic simulation, we make the length of a transaction ranges from 2 to 7. All the performance are implemented on matlab R2012b.

We use three criteria to evaluate the performance of our framework: 1. time required of sanitation algorithm. 2. Number of lost rules. 3. Number of new rules. Lost rules are the non-sensitive rules that can be mined before the sanitization algorithm and lost after that. New rules are the non-sensitive rules that cannot be mined before the sanitization algorithm and being introduced to the mining result after that. Rules hiding would have some side effect on the mining result. Number of lost rules and new rules are used to evaluate the side effect of our sanitization algorithm.

5.1 Time Requirement

The main contribution of our framework is on the sanitization algorithm component, so we only consider the time requirement of that part. As shown in Fig. 3, the time required by sanitization algorithm is linear in |D|.

5.2 Lost Rules

Figure 4 shows the number of lost rules in different size of dataset. We can see it is almost a horizontal line except two peaks of 25 k and 50 k. By looking at the five hiding rules of the two dataset can explain the peaks.

As Table 5 shows that, in 25 k dataset, the first hiding rule [14,22,23=>5] and second hiding rule [5,22,23=>14] are brother rules. As depicted in our sanitization algorithm, only one of the two will represent the Rb to be hidden. Because they come from the same transactions, hiding one of them will cause the hiding of the other. That is to say, we only hide four rules in the 25 k dataset, which causes the lower values than other datasets. Then we look at the hiding rules in the 50 k dataset. The first hiding rule [9,27=>8] has a support value of 6, which is much larger than other support values. According to our sanitization algorithm, it has to make four loops to hide the rule. This is why the 50 k dataset has a much larger number of lost rules. Now we know that the number of lost rules does not depend on the volume of dataset, but the number of hiding rules and how many times the loop is performed. And proposing the concept of "brother rules" can effectively reduce the number of hiding rules, and in consequence of reducing number of lost rules.

Table 6 shows a comparison between hiding rules and lost rules of the 25 k dataset. It is obviously that almost all the lost rules are brother rules of the hiding rules except the last one [8,23=>5]. Actually [8,23=>5] and the hiding rules [14,22,23=>5], [5,22,23=>14] are very much alike though they are not brother rules. And in realistic EMR condition, for example, rule [12,22,23=>5] is sensitive rule, then rule

Table 4. Data set used in experiment

|D|	|I|	TL	|Rh|
15 k	50	2 ~ 7	5
20 k	50	2 ~ 7	5
25 k	50	2 ~ 7	5
30 k	50	2 ~ 7	5
35 k	50	2 ~ 7	5
40 k	50	2 ~ 7	5
45 k	50	2 ~ 7	5
50 K	50	2 ~ 7	5

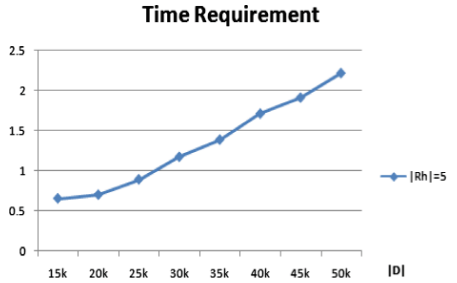

Fig. 3. Time requirement

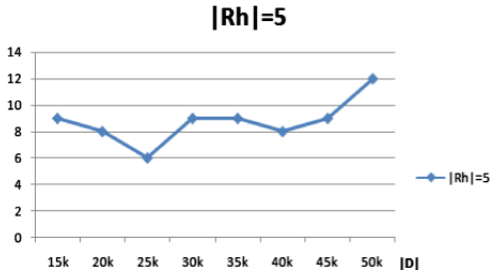

Fig. 4. Lost rules

[5,14,22=>23] and [5,14,23=>22] are very likely also sensitive rules. Consequently, the side-effect of lost rules is minimized in our architecture.

5.3 New Rules

As depicted in Fig. 5 that only one new rule appears in the 20 k dataset. The number of new rules generated during our sanitization algorithm is quite low and it tends to decrease with the increasing of dataset size. So we can say there is almost no effect of introducing new rules to the mining result during our sanitization algorithm.

Table 5. Five Rules to be Hidden

|D|	Rh	S	C	|D|	Rh	S	C
25 k	14,22,23=>5	3	50 %	40 k	25,45,48=>2	3	75 %
	5,22,23=>14	3	75 %		4,16,34=>28	3	60 %
	21,31,43=>47	3	100 %		12,39,44=>35	3	75 %
	27,43=>22	3	50 %		9,24=>36	3	50 %
	3,44=>23	4	57.143 %		26,35=>50	3	50 %
30 k	5,36=>34	3	50 %	50 k	9,27=>8	6	100 %
	12,39=>32	3	60 %		2,28,30=>46	3	50 %
	9,33,48=>43	3	75 %		24,25,45=>5	3	50 %
	13,30,31=>22	3	50 %		8,21,42=>15	3	75 %
	25,28,50=>26	3	100 %		23,24,42=>14	3	51.423 %

Table 6. Hiding Rules and Lost Rules

Rh	Lost Rules
14,22,23=>5 5,22,23=>14	5,14,22=>23 5,14,23=>22
21,31,43=>47	21,31,47=>43 21,43,47=>31
27,43=>22	22,43=>27
3,44=>23	
	8,23=>5

Fig. 5. New rules

6 Conclusion

We proposed an architecture of privacy preserving for medical data publishing. We don't use the traditional reconstruction algorithm to reconstruct a new dataset, but direct modification in original dataset, by which can effectively reducing the cost of time.

In addition, we propose a fundamental sanitization approaches in our architecture. The approach hides a sensitive rule by reducing its r_r until either support value or confidence value of the rule below the threshold. We also put forward a concept of brother rules to reduce the execution of the algorithm.

Finally, we experiment the algorithm on seven sets of dateset. And we use three criteria to evaluate the performance of algorithm: 1. time requirement of the process. 2. number of lost rules. 3. number of new rules. Lost rules are the rules mined before the algorithm but can't be mined after that. New rules are the rules can't be mined before the algorithm but being introduced to the mining result after the process. We believe that the proposed architecture could satisfy the demands of the health department on medical data publishing.

Our future plan is to work on the privacy measurement issues. We hope to develop different hiding strategies and different arguments according to different privacy metrics, in order to adapt to different data users. Moreover, we hope to use the actual medical dataset to perform the experiment in order to get a more real mining results.

References

1. Malik, M.B., Ghazi, M.A., Ali, R.: Privacy preserving data mining techniques: current scenario and future prospects. In: 3rd IEEE International Conference on Computer Communication Technology (ICCCT), pp. 26–32 (2012)
2. Agrawal, R., Srikant, R.: Privacy preserving data mining. In: Proceedings of ACM SIGMOD Conference, pp. 439–450 (2000)
3. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy preserving data publishing: a survey of recent developments, ACM Comput. Surv. **42**(4), art. id 14 (2010)
4. Xu, L., Jiang, C.: Information security in big data: privacy and data mining. IEEE **2**(10) (2014)
5. Matwin, S.: Privacy preserving data mining techniques: survey and challenges. In: Custers, B., Calders, T., Schermer, B., Zarsky, T. (eds.) Discrimination and Privacy in the Information Society, pp. 209–221. Springer, Berlin (2013)
6. Chen, L., Yang, J.: Privacy-preserving data publishing for free text chinese electronic medical records. In: IEEE 35th International Conference on Computer Software and Applications, pp. 567–572 (2012)
7. Chen, L., Yang, J.: A framework for privacy-preserving healthcare data sharing. In: IEEE 14th International Conference on e-Healthcare Networking, Applications and Services, pp. 341–346 (2012)
8. Hossain, A.A., Ferdous, S.M.S.: Rapid cloud data processing with healthcare information protection. In: IEEE 10th World Congress on Services, pp. 454–455 (2014)
9. Alabdulatif, A., Khalil, I.: Protection of electronic health records (EHRs) in cloud. In: 35th Annual International Conference of the IEEE EMBS Osaka, Japan, pp. 4191–4194 (2013)
10. HIPAA-General Infromation. http://www.cms.gov/HIPPAGenInfo/
11. Sweeney, L.: K-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl.-Based Syst. **10**(5), 557–570 (2002)
12. Machanavajjhala, A., Gehrke, J.. Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy Beyond k-anonymity. In: International Conference on Data Engineering (ICDE), pp. 24–35. IEEE Computer Society, Atlanta (2006)
13. Li, N.H., Li, T.C., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: 23rd IEEE International Conference on Data Engineering (ICDE), pp. 106–115. IEEE Computer Society, Istanbul (2007)
14. Gionis, A., Tassa, T.: k-anonymization with minimal loss of information. IEEE Trans. Knowl. Data Eng. **21**(2), 206–219 (2009)

15. Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y.: State of the art in privacy preserving data mining. ACM SIGMOD Rec. **33**(1), 50–57 (2004)
16. Sathiyapriya, K., Sadasivam, G.S.: A survey on privacy preserving association rule mining. Int. J. Data Mining Knowl. Manage. Process **3**(2), 119 (2013)
17. Zhu, J.M., Zhang, N., Li, Z.Y.: A new privacy preserving association rule mining algorithm based on hybrid partial hiding strategy. Cybern. Inf. Technol. **13**, 41–50 (2013)
18. Jain, D., Khatri, P., Soni, R., Chaurasia, B.K.: Hiding sensitive association rules without altering the support of sensitive item(s). In: Meghanathan, N., Chaki, N., Nagamalai, D. (eds.) CCSIT 2012, Part I. LNICST, vol. 84, pp. 500–509. Springer, Heidelberg (2012)
19. Le, H.Q., Arch-Int, S., Nguyen, H.X., Arch-Int, N.: Association rule hiding in risk management for retail supply chain collaboration. Comput. Ind. **64**(7), 776–784 (2013)
20. Dehkordi, M.N.: A novel association rule hiding approach in OLAP data cubes. Indian J. Sci. Technol. **6**(2), 4063–4075 (2013)
21. Bonam, J., Reddy, A.R., Kalyani, G.: Privacy preserving in association rule mining by data distortion using PSO. In: Satapathy, S.C., Avadhani, P.S., Udgata, S.K., Lakshminarayana, S. (eds.) Proceedings of the ICT Critical Infrastructure, Proceedings of 48th Annual Convention Computer Society India, vol. 2, pp. 551–558. Springer (2014)
22. Radadiya, N.R., Prajapati, N.B., Shah, K.H.: Privacy preserving in association rule mining. Int. J. Adv. Innovative Res. **2**(4), 203–213 (2013)
23. Verykios, V.S.: Association rule hiding methods. Wiley Interdiscipl. Rev. Data Mining Knowl. Discovery **3**(1), 28–36 (2013)
24. Chen, X., Orlowska, M., Li, X.: A new framework of privacy preserving data sharing. In: Proceedings of the 4th IEEE ICDM Workshop: Privacy and Security Aspects of Data Mining, pp. 47–56. IEEE Computer Society (2004)