http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 9 (3): 500-505, 2010
ISSN 1812-5638

A Novel Hybrid Protection Technique of Privacy-Preserving Data Mining and Anti-Data Mining

Tung-Shou Chen, Jeanne Chen and Yuan-Hung Kao
National Taichung Institute of Technology,
Graduate School of Computer Science and Information Technology, Taichung, Taiwan

Abstract: In this study, we proposed a novel hybrid protection scheme to protect privacy information and clustering knowledge in mined data. The scheme involves integrating the privacy-preserving data mining technique with that of the knowledge-preserving anti-data mining technique. Hierarchical clustering is used and the clustering structure is manipulated by perturbation to create original data where the mined data has the appearance of similar information and knowledge from the original dataset but with misleading and non-useful contents. The scheme is novel in that it allows users to tailor the amount of protection on personal basis. Experimental tests were conducted on ten public datasets results showed that the privacy information in datasets is preserved and the clustering knowledge cannot be revealed in the mining process. Furthermore, the original dataset can be restored using the key values in the reverse order of the two phases perturbation procedures.

Key words: Knowledge discovery, anti-data mining, privacy-preserving data mining, clustering

INTRODUCTION

Data Mining is a technology which mines the potential supporting decision-making knowledge from a large number of databases by using information technologies and statistics methods (Dunham, 2003). The main techniques in data mining include association rules, clustering, classification, sequential pattern analysis and more (Dunham, 2003). These techniques could easily leaked sensitive knowledge in the mined data which might result in great loss to a company. Others include loss of personal privacy data such as age, wage and more.

Enterprises are expecting not only the potential knowledge analyzed by data mining but the mined knowledge can be protected in case the lost knowledge brings the loss of company. In the meantime, due to the datasets always have personal privacy such as age, wage and more, companies respects more in the issue of safety about these information and reinforces the control authority of using data and information protection.

In this modern generation with fast-developing network, enterprise often put their datasets on the network to provide personnel and clienteles an easy form of information delivery and service. The easy access, however, resulted in more information security issues. Currently, attentions on information security issues are focused on security in the system frame. Examples of which include protecting data access, sending by symmetric and asymmetric encryption (Rapuano and Zimeo, 2008; Rowan, 2007) or using identity identification and authority control techniques (Bertino *et al*., 2002; Bertino, 1998; Hwang and Lee, 2001). However, these techniques are unable to completely protect privacy information and knowledge security because enterprise personnel are usually the source of data loss and disclosure of datasets in companies.

Clifton first proposed the threatening and clash of data mining to protect database (Clifton and Marks, 1996). He concluded that when databases were mined some sensitive data or rules could be revealed in the mined data. Therefore, he proposed a scheme to protect the privacy or sensitive information from disclosure which include using methods like data perturbation (Fung *et al*., 2009), decomposition (Jiang *et al*., 2008) and more. However, not all information was protected with the exception of some sensitive data in databases. They concluded that some important knowledge from the database could have been disclosed in the mined process.

According to the problem of data mining, Chen *et al*. (2008) proposed the Anti-Data Mining (ADM) technique

Corresponding Author: Jeanne Chen, Graduate School of Computer Science and Information Technology, National Taichung Institute of Technology, Taichung 404, Taiwan
Tel: 886-4-2219-6617 Fax: 886-4-2219-6341

to protect knowledge. The technique involved adding noise data to perturb the original knowledge in the dataset so that mined data would contain only non-sensitive information.

In this research, we integrate the protection methods in two data mining techniques to design a novel hybrid data and knowledge protection technique. The aim is to hide the sensitive information of dataset and also protect knowledge from being disclosed in mining. Therefore, the knowledge and privacy information in dataset protected by the proposed method is guaranteed against disclosure.

PPDM and ADM schemes: This study proposed an integrated technique to protect both data and knowledge. However, the data is protected by perturbation; while knowledge is protected by adding noise to the data. The combined methods can tailor mined data to outwardly behave like normal data but actually contained insensitive non-useful information.

Data perturbation for Privacy-Preserving Data Mining (PPDM): The purpose of privacy-preserving data mining (PPDM) is to remove or hide the privacy information before the dataset is accessible in public (Divanis and Verykios, 2009). However, mining the removed privacy information could still gather some if the original knowledge in the dataset. According to the research by Liu *et al.* (2006), the six methods in PPDM are data perturbation, data swapping, k-anonymity, secure multiparty computation, distributed data mining and rule hiding (Liu *et al*., 2006). The methods can be applied to protect privacy data from different data mining techniques. For example, data perturbation can be used to generate noise data to replace the original data fields so that the privacy and sensitive information in original data can be protected (Liu *et al*., 2006).

Thorough consideration must be made when generating the noise data by data perturbation. The original data characters and normal mining procedures in the original dataset are important factors to the noise data to meet the purpose of mining data without privacy disclosure. The common way of data perturbation is that the protected field values are multiplied a noise data to change the original field values. For example, the data waiting for protecting A uses additive or multiplicative parameter c to perturb and change the original value. The additive equation is $B = A + c$; the multiplicative equation is $B = A \times c$, then B is the data through the protecting procedure (Liu *et al*., 2006). However, PPDM can only protect privacy data.

Clustering against mining knowledge for Anti-Data Mining (ADM): Chen *et al.* (2008) proposed the Anti-Data Mining (ADM) to protect the privacy data and knowledge from disclosure. They proposed adding noise data to change the knowledge in the original dataset in order to prevent knowledge from being mined. The main focus was to make use of the random seed in the clustering technique to add a certain number of noise data or noise fields to change the clustering structure of the original data. For example, the aim is to add a noise field to change the data character and then generating a different clustering result to meet the purpose of protecting the clustering result in the original data. The ADM does not modify the original data. On the other hand, the privacy issue was also not considered.

In this study, we integrate the advantages from PPDM and ADM protection techniques to protect both privacy and knowledge from being mined.

A hybrid privacy-knowledge protection technique: The proposed protection technique is in two phases. The first phase applies data perturbation to do PPDM privacy information protection in original dataset; its purpose is to hide the values of the original data and also to analyze the correct mining result. The second phase adds the noise data to the data set protected by first phase and to use ADM to protect knowledge. The purpose is to protect the correct clustering knowledge in the protected dataset from being mined. Therefore, the dataset protected by the proposed technique can not only protect its own privacy information but also prevent the clustering knowledge from being mined by the illegal users.

In this study, the scheme involves using Hierarchical Clustering (HC) (Bolshakova *et al*., 2005) to protect dataset in two phases; PPDM and ADM. The proposed scheme is a retrievable protection technique; the random seed function is used to generate the key values in dataset for retrieval. The random $Seed_1$ and $Seed_2$ are used in the two phases to flexibly undo the protection by PPDM and ADM. Furthermore, users are allowed to set a deviation parameter r as a noise data to change the original data values in PPDM phase and to be the proportional value of generating noise data in ADM phase.

Protecting dataset algorithm: The proposed scheme uses HC to be a data mining tool to analyze dataset D. A tree H is constructed with the clustering result of D. H includes n-1 non-leaf nodes, n is the total data number. The different nodes are divided up to get the different clustering H_i, $1 \le i \le k$, $H_i = \{d_{ij}, 1 \le j \le n_i\}$, d_{ij} representing the data in H_i; k is the clustering number and n_i is the data number of H_i. The proposed scheme make use of data perturbation where noise data is added to the two phases to perturb the original dataset. Therefore, dataset D' is

generated which protected by PPDM and the dataset D'' is protected by ADM. The datasets, D' and D'', are analyzed by HC where the clustering results H' and H'' are generated. The un-protected, protected and unchanged clustering data number s' and s'' are then calculated for the different phases. For example, in the first phase PPDM, Eq. 1 calculates s' and s'' where, $\|H_i \cap H'_i\|$ represents the data number of the set. Also, the data number of $d_{ij} \in H_i$ and $dij \in H'_i$, which represents the unchanged part of H_i in H'_i:

$$s' = \| H_i \cap H'_i \| = \max(\| H_i \cap H'_i \|), \text{ for } \forall i \in [1,k]$$

$$s'' = \| H'_i \cap H''_i \| = \max(\| H'_i \cap H''_i \|), \text{ for } \forall i \in [1,k] \quad (1)$$

For easier examination of the effect, Eq. 2 is used to generate the rate value RC' and RC'' of unchanged data number s' and s'' where total number n represents the degree of perturbation in the clustering result. When RC value is higher more changes are realized in the clusters of the datasets. Therefore, a lower RC' value in the PPDM phase meant that the dataset perturbed by data perturbation can be analyzed by clustering resulting in knowledge that appeared similar with the original dataset (however, the knowledge is purposely released non useful information). On the contrary, a higher RC'' value in the ADM phase represents remarked differenced between the clustering result and the original dataset:

$$RC' = \left(1 - \frac{s'}{n}\right) \times 100$$

$$RC'' = \left(1 - \frac{s''}{n}\right) \times 100 \quad (2)$$

The hybrid data and knowledge protecting technique proposed in this paper are illustrated as following:

- **Input:** Original dataset D where D = {di, i = 1, 2, 3, ..., n}, i is the data number of D with n data numbers with each data record dl = (d_{i1}, d_{i2}, d_{i3}, .., d_{im}) and m is the total data columns of d_i; virtual random $Seed_1$ and $Seed_2$ values; perturbation parameter rϵ(0, 1]; limited executive times T_t with threshold values T_{r1} and T_{r2}
- **Output:** Protected dataset D'' where D'' = {d_i'', i = 1, 2, 3, ..., n}; perturbation parameter r'

Phase I: Data perturbation of PPDM:

- **Step 1:** Analyze the dataset D by HC clustering to generate the clustering result H
- **Step 2:** Set r' = r, calculation times t = 1, threshold value T_{r1} as the clustering evaluation by data perturbation, the random generator $Seed_1$ value in Phase I is as the Rand() function and its generating range [0,1)
- **Step 3:** Proceed data perturbation in dataset D. Use Eq. 3 to change the data value of dataset D to generate the protected dataset D', D' = {d'_i, I = 1, 2, 3, .., n}

$$d'_{ij} = d_{ij} \times (1 + (\text{Rand}() - 0.5) \times r'), \text{ for } \forall j \in [1,m] \quad (3)$$

- **Step 4:** Analyze dataset D' by HC to generate the clustering result H' and to calculate the RC'' value in phase I
- **Step 5:** Check RV''>T and t<T_t, if it is true that means the clustering structure changes too much, then delete D' and to set r' = r'×r, t = t+1 and reset.Rand(), return to Step 1.3
- **Step 6:** Generate dataset D' in Phase I data perturbation

Phase II: Adding noise data in ADM:

- **Step 1:** Analyze dataset D' by HC to generate the clustering result H'
- **Step 2:** Set the calculation times t = 1, threshold value T_{r2} as a predictable perturbation rate, the random generator $Seed_2$ value in Phase II is as the Rand() function and its generating range [0,1)
- **Step 3:** Use Eq. 4 to generate the noise data y_1, i is $1 \le i \le n$, q is the range of perturbation, which can be set by users, or to get the max(d_{ij}) and min(d_{ij}) in dataset, j is $1 \le j \le m$ as the border value of the noise data range for meeting the original data character. To add column y_i into D' to generate D'':

$$y_i = \text{Rand}() \times q + \min(d_{ij}) \quad (4)$$

- **Step 4:** Analyze dataset D'' by HC to generate the clustering result H'' and to calculate the RC'' value in phase II
- **Step 5:** Check RC''<T_{r2} and t<T, if it is true that means the changed clustering structure does not meet the expectation, then delete D'' and to set q = q×(1+r'), t = t+1 and reset.Rand(), return to Step 3
- **Step 6:** Execute RAnd()×m to generate the column order of noise column in D' and insert the noise data by the generating value order to output the protected dataset D'' in phase II

After completing the two phases of data and knowledge protection procedures, the dataset has double security protection against data mining which meets the purpose of effective privacy and knowledge protection.

The two phases of protection method proposed in this study has individual key values which can be flexibly applied by users to phase data mining analysis. Also, the clustering characteristic of the original dataset should be considered when setting the noise data range and adding noise column setting to prevent easy guessing and removal by hackers. At the same time, more than one noise data should be used to increase dataset robustness.

Restoring the protected dataset: The retrieval process is the reverse of the perturbation procedures. Since, the perturbation procedure is in two phases; the retrieval process starts from the second phase with dataset D'' on the reverse ending in the first phase. The algorithm describing the retrieval process follows:

- **Input:** The protected dataset D'' where D'' = {d_i'', i = 1, 2, 3,..., n}, n is the total data number of D''; each data record d_i'' = (d''_{i1}, d''_{i2}, d''_{i3}, ..., $d''_{im'}$) where, m' is the total data column of d_i''; virtual random $Seed_1$ and $Seed_2$; perturbation parameter r'
- **Output:** The original dataset D where D = {d_i, i = 1, 2, 3, .., n}

Phase II: Remove the added noise data:

- **Step 1:** Set random generator $Seed_2$ value, set the random generator Rand() range is [0,1) to execute n times of Rand() which is the execute times of generating noise data
- **Step 2:** Execute Rand()×(m'-1) to find out the order of the noise data in dataset D''
- **Step 3:** Remove the noise data and output the retrieved dataset D' in phase II

Phase I: Retrieve the values of the original data:

- **Step 1:** Set the random generator $Seed_1$ value, set the random generator Rand() range is [0,1)
- **Step 2:** Use Eq. 5 to retrieve the data values of the original dataset D:

$$d_{ij} = d'_{ij} \div (1 + (\text{Rand}() - 0.5) \times r'), \text{ for } \forall j \in [1,m] \quad (5)$$

In the random seed function, the same seed value is used to generate the same random values in each repeated execution. Therefore, in the proposed scheme the random $Seed_1$ and $Seed_2$ values are used as protecting key values in the two phases with noise parameter r' to generate a complete protection procedure. Only users have the correct key value to retrieve the protected dataset. Also, the correct data and clustering knowledge in dataset may be retrieved in different phase by applying the keys to the different protection layers.

Experimental results and effect analysis: For experimental testing, ten example datasets from UCI Machine Learning Repository (Asuncion and Newman, 2007) website were used (Table 1 for details). The experiment also makes use of Cluster 3.0 as the clustering tool for HC (De Hoon *et al*., 2004) and sets the centroid linkage as the clustering method of HC.

Table 2 illustrates the clustering results for the famous Iris dataset in Phase I. The Iris dataset comes with 150 four-column records, which is divided into three clusters (Asuncion and Newman, 2007) based on the data character, that means the clustering number k = 3. In the experiment, we set the Iris dataset as D clustering by HC and generate the correct clustering result H_i = {H_1,H_2,H_3}. Each clustering data number is listed as row n_i on Table 2. We set the perturbation parameter r in the median 0.5 of [0,1] (r = 0.5) and the limit executive times T_t = 10 with threshold values T_{r1} = 5 and T_{r2} = 50; the threshold values are empirical. We allow modification to the clustering structure to be smaller than 5% in phase I and bigger than 50% in phase II.

Next, $Seed_1$ is set to change the original dataset D in the protection procedure of phase I PPDM; the protected dataset D' is generated in this phase. The HC is used to analyze D' and to generate the clustering result H'_i = {H'_1, H'_2, H'_3}. The number of intersecting data from the each clustering of D and D' are also analyzed. Table 2 shows the results of H'_1, H'_2 and H'_3 clustering data of D' overlay in D. As seen, the clustering data of H'_1 is similar with H_1; H'_2 includes not only all records in H_2 but also a record from H_3 and, H'_3 has only 35 records from H_3. The

Table 1: Details of the different datasets from the UCI website

Datasets	No. of records	No. of fields per record	No. of clusters
BUPA liver disorders	345	6	2
Dermatology database	366	34	6
Glass identification database	214	9	7
Haberman's survival data	306	3	2
Hayes-roth and hayes-roth database	132	5	3
Iris plants database	150	4	3
Pima Indians diabetes database	768	8	2
Protein localization sites	336	7	8
Teaching assistant evaluation	151	5	3
Wine recognition data	178	13	3

Table 2: The clustering result of Iris protected by Phase I

	Original HC label		
HC label after protection	H_1	H_2	H_3
n_i	50	64	36
H'_1	50	0	0
H'_2	0	64	1
H'_3	0	0	35

Table 3: Experimental result of Iris dataset in two-phase protections

Result	H_1	H_2	H_3
Phase I			
n_1, n_2, n_3	50	64	36
$\{H'_1, H'_2, H'_3\} \Rightarrow (n'_1, n'_2, n'_3)$	50	65	35
$\{H'_1, H'_2, H'_3\} \Rightarrow (s'_1, s'_2, s'_3)$	50	64	35
RC"			0.667
Phase II			
$\{H'_1, H'_2, H'_3\} \Rightarrow (n'_1, n'_2, n'_3)$	50	65	35
$\{H''_1, H''_2, H''_3\} \Rightarrow (n''_1, n''_2, n''_3)$	25	59	66
$\{H''_1, H''_2, H''_3\} \Rightarrow (n''_1, n''_2, n''_3)$	25	35	11
RC"			52.667

Table 4: Experimental result of UCI ten datasets in two-phase protections

Datasets	RC'	RC"
BUPA liver disorders	2.609	51.884
Dermatology database	2.732	61.202
Glass identification database	0.467	81.308
Haberman's survival data	0.634	51.961
Hayes-roth and hayes-roth database	2.272	58.333
Iris plants database	0.667	52.667
Pima Indians diabetes database	1.562	50.391
Protein localization sites	0.326	51.961
Teaching assistant evaluation	3.973	54.305
Wine recognition data	4.494	83.708

calculated results for s' and RC' in phase I are listed in Table 3. The Hierarchical Anti-Clustering (HAC) is used in the experiment with phase II ADM to observe the clustering structure to confirm the effect of the data and knowledge protection of the proposed scheme in the two phases.

According to the values on row s' in Table 3, with the perturbing phase I by PPDM, only one record of D' has changed its cluster. The whole clustering result has changed very little. This meets the purpose of protecting data and to allow mining of the clustering knowledge similar to the original ones (the clustering knowledge, however contains useless information). Since, the range of change in parameter r' is radical, a wide gap difference is realized between RC' and $T_{r1} = 5$. However, r' may be adjusted by users to get different dataset characteristics. In this study, we use the same parameter values to test the all dataset for consistency. The row s" on Table 3 has not changed the data number of clustering clusters after perturbing in phase II. Eq. 2 is used to calculate RC" = 52.667 in phase II to meet the perturbation expectation $T_{r2} = 50$, which means that the knowledge protection of ADM in phase II truly can change the original structure over 50% in order to protect the knowledge.

Finally, experimental results from all ten datasets are shown in Table 4. From the table, it can be seen the effects of the proposed perturbation of RC' and RC" on all the en datasets in the two phases of data and knowledge protection. As seen in Table 4, after each dataset in phase I was perturbed, all the RC' values are small than 5; that is the clustering knowledge of the protected dataset by phase I meets the expectation of under 5% difference and all RC" values are bigger than 50 which means the perturbation used in ADM of phase II can truly change the clustering structure of the original data over 50% to meet the purpose of clustering knowledge protection.

CONCLUSIONS

The proposed scheme was aimed at the security issue of data mining to propose a novel hybrid data and knowledge protection technique for perturbing datasets to meet the purpose of protecting privacy and clustering knowledge in datasets. In this study, we use hierarchical clustering in data mining technology as an example which generates a novel information-security-protection scheme against mining of the privacy and knowledge in datasets. The proposed scheme is novel in that users can tailor the degree of protection by adjusting the perturbation parameter r to the required degree of security protection.

The proposed scheme is different from the traditional data encryption and decryption schemes to protect against data mining. The scheme integrated the two protection techniques by PPDM and ADM to manipulate and perturb the data structure and information in the datasets. Protection is doubled and the protected dataset gives misleading mined knowledge and information to illegal users. The original dataset may be retrieved with the key values used in the perturbation procedures.

REFERENCES

Asuncion, A. and D. Newman, 2007. UCI machine learning repository. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Bertino, E., 1998. Data security. Data Knowledge Eng., 25: 199-216.

Bertino, E., S. Castano, E. Ferrari and M. Mesiti, 2002. Protection and administration of XML data sources. Data Knowledge Eng., 43: 237-260.

Bolshakova, N., F. Azuaje and P. Cunningham, 2005. An integrated tool for microarray data clustering and cluster validity assessment. Bioinformatics, 21: 451-455.

Chen, T.S., J. Chen, Y.H. Kao and T.C. Hsieh, 2008. A novel anti-data mining technique based on hierarchical anti-clustering (HAC). Proc. 8th Int. Conf. Intel. Syst. Design Appl., 3: 426-430.

Clifton, C. and D. Marks, 1996. Security and privacy implications of data mining. Proceedings of the ACM SIGMOD Workshop on Data Mining and Knowledge Discovery, June 1996, Montreal, Canada, pp: 15-19.

De Hoon, M.J.L., S. Imoto, J. Nolan and S. Miyano, 2004. Open source clustering software. Bioinformatics, 20: 1453-1454.

Divanis, A.G. and V.S. Verykios, 2009. Exact knowledge hiding through database extension. IEEE Trans. Knowledge Data Eng., 21: 699-713.

Dunham, M.H., 2003. Data Miming: Introductory and Advanced Topics. Prentice Hall, New Jersey.

Fung, B.C.M., K. Wangb, L. Wanga and P.C.K. Hung, 2009. Privacy-preserving data publishing for cluster analysis. Data Knowledge Eng., 68: 552-575.

Hwang, M.S. and C.H. Lee, 2001. Secure access schemes in mobile database systems. Eur. Trans. Telecommun., 12: 303-310.

Jiang, W., C. Clifton and M. Kantarcýoðlu, 2008. Transforming semi-honest protocols to ensure accountability. Data Knowledge Eng., 65: 57-74.

Liu, K., H. Kargupta and J. Ryan, 2006. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Trans. Knowledge Data Eng., 18: 92-106.

Rapuano, S. and E. Zimeo, 2008. Measurement of performance impact of SSL on IP data transmissions. Measurement, 41: 481-490.

Rowan, T., 2007. VPN technology: IPSEC vs SSL. Network Security, 2007: 13-17.