International Journal of Mathematical Archive-1(3), Dec. - 2010, Page: 86- 91
Available online through www.ijma.info ISSN 2229 – 5046

AN IMPROVED APPROACH FOR MINING PRIVACY - PRESERVING FREQUENT ITEMSETS

Bharat Solanki*, Rashmi Awasthy and Rajesh Shrivastava

Shri Ram Institute of Technology-MCA
E-mail: Jabalpur_bharat@yahoo.co.in
Shri Ram Institute Of Technology, Computer Science Department Jabalpur, Madhya Pradesh ,India
E-mail: Rashmi.8sept@gmail.com
Shri Ram Institute of Technology-MCA
(Received on: 03-12-10; Accepted on: 11-12-10)

Abstract

Due to the increasing use of very large databases and data warehouses, mining useful information and helpful knowledge from transactions is evolving into an important research area. Frequent Itemsets (FI) Mining is one of the most researched areas of data mining. In order to mining privacy preserving frequent itemsets on large transaction database efficiently, a new approach was proposed in this paper.

Keywords- *Data mining, frequent itemsets, privacy preserving*

---

1. INTRODUCTION:

Data mining services require accurate input data for their results to be meaningful, but privacy concerns may influence users to provide spurious information. In order to preserve the privacy of the client in data mining process, a variety of techniques based on random perturbation of data records have been proposed recently.

Randomization and Distortion are the two dominant methods provided as a means to preserve the privacy. Randomization process modifies each transaction by replacing some of the existing items with non-existing items, and adding some fake items, thereby preserving the privacy. Distortion process operates on a transaction database by probabilistically changing some of the items in each transaction.

We focus on an improved distortion process that tries to enhance the accuracy by selectively modifying the list of items. The normal distortion procedure does not provide the flexibility of tuning the probability parameters for balancing privacy and accuracy parameters, and each item's presence/absence is modified with an equal probability. In improved distortion technique, frequent one item-sets, and non-frequent one item-sets are modified with a different probabilities controlled by two probability parameters *fp*, *nfp* respectively. The owner of the data has a flexibility to tune these two probability parameters (*fp* and *nfp*) based on his/her requirement for privacy and accuracy. The experiments conducted on real time datasets confirmed that there is a significant increase in the accuracy at a very marginal cost in privacy.

---

*Corresponding author: Bharat Solanki**
E-mail: Jabalpur_bharat@yahoo.co.in
Shri Ram Institute Of Technology, Computer Science Department Jabalpur, Madhya Pradesh, India

A. Model of Data Miners:

Two classes of data miners are considered in this system. One is legal data miners. These miners always act legally in that they perform regular data mining tasks and would never intentionally breach the privacy of the data. On the other hand, *illegal data miners* would purposely discover the privacy in the data being mined. Illegal data miners come in many forms. In this paper, we focus on a particular sub-class of illegal miners. That is, in our system, illegal data miners are *honest but curious*: they follow proper protocol (i.e., they are honest), but they may keep track of all intermediate communications and received transactions to perform some analysis (i.e., they are *curious*) to discover private information.

Even though it is a relaxation from Byzantine behavior, this kind of honest but curious (nevertheless illegal) behavior is most common and has been widely adopted as an adversary model in the literatures. This is because, in reality, a workable system must benefit both the data miner and the data providers. For example, an online bookstore (the data miner) may use the association rules of purchase records to make recommendations to its customers (data providers). The data miner, as a long-term agent, requires large numbers of data providers to collaborate with. In other words, even an illegal data miner desires to build a reputation for trustworthiness. Thus, honest but curious behavior is an appropriate choice for many illegal data miners.

B. Randomization Model:

Let us consider the entire mining process as an iterative one. In each stage the data miner obtains a perturbed transaction from a different data provider. With the randomization approach, each data provider employs a randomization operator R(.) and applies it to one transaction t which the data provider holds.

Upon receiving transactions from the data providers, the legal data miner must first perform an operation called *support recovery* which intends to filter out the noise injected in the data due to randomization, and then carry out the data mining tasks. At the same time, an illegal data miner may perform a particular privacy recovery algorithm in order to discover private data from that supplied by the data providers.

Clearly, the system should be measured by its capability in terms of supporting the legal miner to discover accurate association rules, while preventing illegal miner from discovering private data.

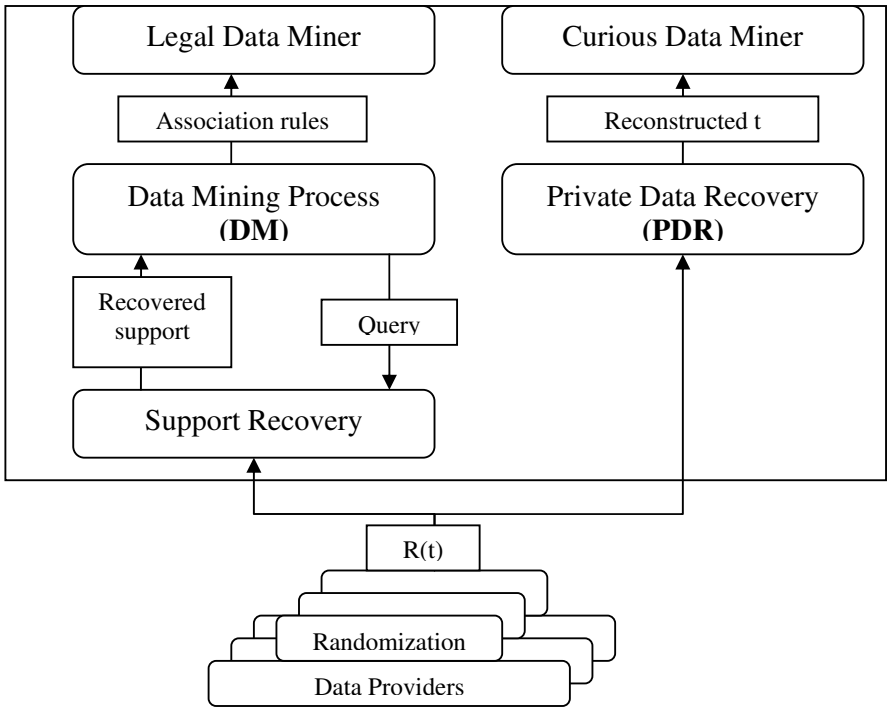

Figure 1: Infrastructure of a Typical Randomization System

C. New Model:

Figure 2 shows the infrastructure of the newly proposed system. The legal data miner contains two components, Data Mining process (**DM**) and Perturbation Guidance (**PG**). When a data provider C_i initializes a communication session, **PG** first dispatches a reference V_k to C_i. Based on the received V_k the data perturbation component of C_i transforms the transaction t to a perturbed one $R(t)$ and transmits $R(t)$ to **PG**. **PG** then updates V_k based on the recently received $R(t)$ and forwards $R(t)$ to the Data Mining process **DM**.

The key here is to properly design V_k so that correct guidance to data provider on how to distort the data transactions. In this system, V_k is an algebraic quantity derived from *T (Transaction Database)* which enables us to effectively maintain the accuracy of data mining while significantly reduces the leakage of private information.

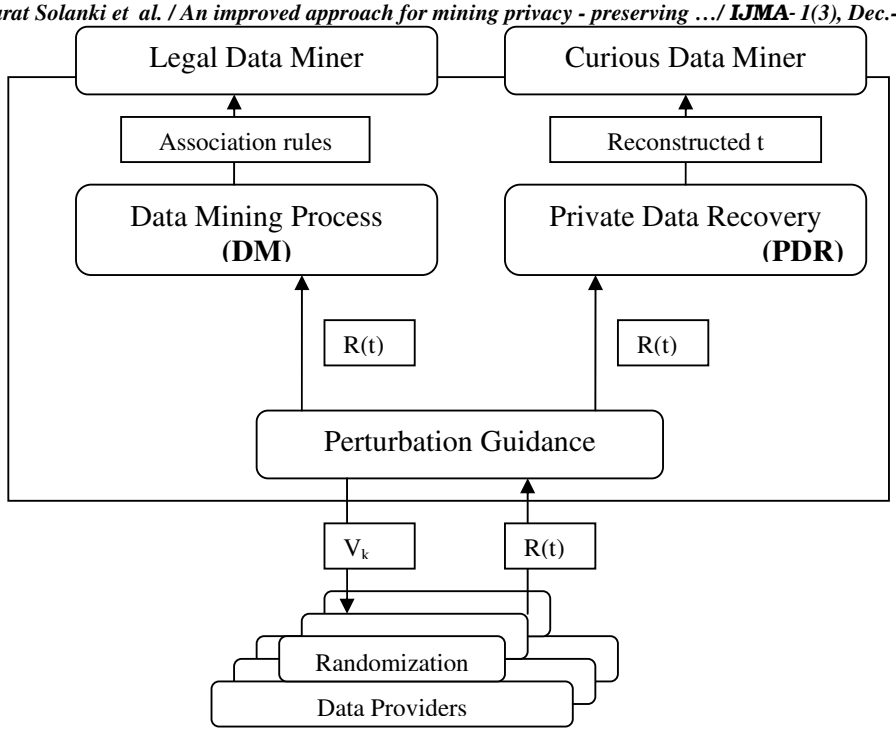

Figure 2: Infrastructure of a newly Proposed Randomization System

D. Communication Protocol:

The details of the communication protocol used between data providers and data miners are as follows. On the side of the data miner there are two current threads that perform the following operations iteratively after initializing V_k.

Thread of registering Data provider
R1: Negotiate on the truncation level *k* with a data provider R2: Wait for a ready message from a data provider R3: Upon receiving the ready message from a data provider Register the data provider Send the data provider current V_k R4: Goto R1
Thread of Receiving data transaction
T1: Wait for a (perturbed) data transaction *R(t)* from a data provider T2: Upon receiving the data transaction from a registered data provider, Update V_k based on the newly received perturbed data transaction Deregister the data provider T3: Goto T1

For a data provider, it performs the following operations to transfer its transaction to the data miner:

P1: Send the data miner a ready message indicating that this provider is ready to contribute to the mining process. P2: Wait for message that contains V_k from the data miner. P3: Upon receiving the message from the data miner, compute *R(t)* based on *t* and V_k P4: Transfer *R(t)* to the data miner.

II. RELATED WORK:

In (Agrawal et al) [6], efficiency issues in privacy preserving mining are addressed. They demonstrated that it is possible to bring the efficiency to *well within an order of magnitude* with respect to direct mining, while retaining satisfactory privacy and accuracy levels. This improvement is achieved through changes in both the distortion process and the mining process of MASK (Mining Associations with Secrecy Konstraints), resulting in a new algorithm called *EMASK* (Efficient MASK).

In (Kun Liu et al) [7], the following problem is discussed. Suppose there are N organizations O_1, $O_2 \ldots O_N$; each organization O_i has a private transaction database DB_i. A third party data miner wants to learn certain statistical properties of the union of these databases $\bigcup_{i=1}^{N} DB_i$ .These

organizations are Comfortable with this, but they are reluctant to disclose their raw data. How could the data miner perform data analysis without compromising the privacy of the data? In this scenario, the data is usually distorted and its new representation is released; anybody has arbitrary access to the published data. The authors provide randomized multiplicative data perturbation technique to solve some of the problems of additive random perturbation.

The authors in [8] used an efficient updating technique in privacy preserving frequent itemset mining, and an incremental algorithm called *IPPFIM (Incremental Privacy Preserving Frequent Itemset Mining)* is proposed.

III. IMPROVED DISTORTION ALGORITHM:

We propose an extension to the so called distortion technique MASK (Mining Associations with Secrecy Konstraints) proposed in [2]. Accuracy and Privacy are typically contradictory in nature in the sense that, improving one normally incurs a cost in the other. The distortion approach proposed in [2] aimed at providing as much privacy as possible at the same time maintaining good accuracy in mining results. This method does not provide the flexibility of tuning the probability parameters for balancing privacy and accuracy parameters, and each item's presence/absence is modified with an equal probability. Our approach further improves the distortion technique to provide better accuracy while keeping the privacy also as an important factor. In our improved distortion technique, frequent one itemsets are modified with a lesser probability (*fp*), and non-frequent one itemsets are modified with a greater probability (*nfp*). The owner of the data has a flexibility to tune these two probability parameters (*fp* and *nfp*) based on his/her requirement for privacy and accuracy.

Another advantage of our distortion method is that any off-the-shelf algorithms can be used to find the frequent itemsets from the distorted database without any modification. So the time taken to mine the distorted database is same as that of the original database. Some of the previous distortion algorithms assume that the transactions are stored as bitmap files consisting of 0s and 1s. We apply distortion procedure not on the bitmap file but on the item-list representation of the database which is common representation for transactional databases. Each transaction is dynamically converted into a bitmap before distortion, and converted back to item-list representation before being stored on to the disk. So our solution is also space efficient, since much less space is consumed by a database (especially if the database is sparse) when it is represented as item lists rather than large bitmap.

A. Solution Framework:

Let 'I' be a set of 'n' items $\{a_1, a_2 \ldots a_n\}$ and 'T' be a set of transactions $\{t_1, t_2 \ldots t_n\}$ where each transaction t_i is a subset of 'I'.

Each transaction can be considered to be a random Boolean vector $X = \{X_i\}$, such that X_i is either 0 or 1. $X_i = 1$ (or 0) indicates that the transaction represented by X (does not) include(s) the item a_i. We generate the distorted vector from this transaction by computing $Y = distort(X)$ where $Y_i = X_i \oplus R_i'$ and R_j' is the complement of R_i, a random variable with density function

$$f(R) = Bernouli(p), (0 \le p \le 1) \quad (1)$$

i.e., R_i takes a value 1 with probability p and 0 with probability $(1 - p)$. Each bit in the vector X is flipped with a probability of p.

In normal distortion scheme [2], each bit is distorted with equal probability. But in optimal distortion technique frequent items are distorted with one probability, and non-frequent items are distorted with a different probability. This is to ensure that good accuracy is achieved even after distortion. These two probability parameters can be tuned as per the user's requirements for privacy and accuracy.

B. Distortion Algorithms:

Let *Bitmap[1...n]* contains a bitmap representation of a transaction t_i and p be the distortion probability. *Convert_to_bitmap( )* converts an item-list of a transaction to a bitmap. *get_random_double( )* generates a random real number between 0 and 1 with uniform probability.

C. Normal Distortion Algorithm:

The normal distortion algorithm changes every item with an equal probability say *p*. This algorithm scans the database only once.

Algorithm 1 Normal Distortion

```
For i ← 1 to m
  Bitmap ← Convert_to_bitmap(t_i)
For j ← 1 to n
  Rand_num ← get_random_double()
  If Rand_num > p
    Bitmap[i] ← (Bitmap[i]+1)%2
```

Figure 3

D. Improved Distortion Algorithm:

The improved distortion algorithm changes frequent items with a less probability (*fp*) and non-frequent items with a greater probability (*nfp*). The values of *fp* and *nfp* can be changed by the user.

This algorithm makes two scans over the entire database. In first scan the supports are calculated for each item, and stored in an array. Let *freqs[1…n]* stores the frequencies of all the items and supp be the minimum support. In the second scan, the actual distortion process takes place as per the following algorithm.

Algorithm 2 Improved Distortion

```
For i ← 1 to m
Bitmap ← Convert_to_bitmap(t_i)
```



```
For j ← 1 to n
Rand_num ← get_random_double()
If freqs[j]< supp & bitmap[j]=0
If Rand_num > nfp
Bitmap[j] ← (Bitmap[j]+1)%2
Else
Rand_num ← get_random_double()
If Rand_num > fp
Bitmap[j] ← 0
```

Figure 4

IV. EXPERIMENTAL RESULTS :

To assess the effectiveness of our algorithms, the experiments are conducted on three popular real time datasets *Retail, BMS-Webview-1, BMS-Webview-2* [7].

For each of the three databases privacy (P), accuracy (A) metrics are calculated for various distortion probabilities (*nfp*) with an interval of 0.5 (Tables 1 to 3) and *fp*=0.95. The bench mark supports used for retail, BMS-Webview-1, BMS-Webview-2 are 0.003, 0.002, and 0.003 respectively. These two metrics are calculated and compared for normal distortion, and improved distortion. As the distortion probability decreases privacy increases, and accuracy decreases. The experimental results show that with a minor reduction in privacy, accuracy can be improved significantly with the improved distortion technique.

Table 4 reports the execution times of a distortion algorithm implemented using item-list file representation (T_I) and bitmap file (T_B) representation of the transactional database. Experiments are conducted in the three datasets used in this paper. For all three databases, distortion algorithm implemented using item-list representation performed better than its bitmap counterpart. In general the former consumes less space and performs significantly better for the databases in which the available items are more, and the average transaction length is less.

Table 1. BMSWebView1

Distortion Probability	Normal Distortion		Improved Distortion (fp=0.95)	
	P	A	P	A
0.95	79.14	85.51	79.14	86.80
0.90	88.52	83.51	79.21	85.18
0.85	92.53	67.57	79.23	82.67
0.80	94.74	62.64	79.74	80.94
0.75	96.10	54.70	80.12	81.31
0.70	96.99	50.9	81.76	81.80

Table 2. Retail

Distortion Probability	Normal Distortion		Improved Distortion (fp=0.95)	
	P	A	P	A
0.95	81.09	85.04	81.09	86.43
0.90	85.63	71.61	81.68	85.66
0.85	88.07	60.11	81.88	84.88
0.80	89.72	50.60	81.97	84.74
0.75	90.93	43.07	82.03	85.23
0.70	91.83	36.33	82.07	84.74

Table 3: BMSWebView2

Distortion Probability	Normal Distortion		Improved Distortion (fp=0.95)	
	P	A	P	A
0.95	90.44	73.37	90.44	75.47
0.90	95.12	56.58	91.07	68.71
0.85	96.93	45.48	91.28	69.82
0.80	97.88	38.53	91.30	70.85
0.75	98.45	32.10	91.45	70.56
0.70	98.82	27.58	91.48	69.52

Table 4: Execution times of distortion algorithms

Dataset	**T_B (sec)**	**T_I (sec)**
Retail	300	43
BMS-Webview-1	134	20
BMS-Webview-2	184	37

V. CONCLUSION:

An improved distortion technique for privacy preserving frequent itemset mining is proposed. Two probability parameters (*fp* and *nfp*) are introduced. Better accuracy values can be obtained by tuning these two parameters with a minor reduction in privacy. This algorithm produces the best results when the fraction of frequent items among all the available items is less. The distortion technique proposed in this paper assumes that transactions are stored in the file as item-lists rather than Boolean arrays which saves the disk space and hence enhances the performance of the algorithm by reducing the disk access time. The database distorted through our

distortion technique will provide reasonably accurate results without reconstruction.

The general frequent itemset mining algorithm can be extended to mine the distorted databases so as to improve the accuracy of the discovered patterns.

REFERENCES

[1] Evfimievski, A., Srikant, R., Agrawal, R. & Gehrke, J. (2002), Privacy Preserving Mining of Association Rules, in 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, pp. 217(228).

[2] Rizvi, S.J., and Haritsa, J.R.: Maintaining data privacy in association rule mining. In Proceedings of the 28th Conference on Very Large Data Bases. (2002).

[3] Zhang, N., Wang, S., and Zhao, W. 2004. A new scheme on privacy preserving association rule mining. In Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (Pisa, Italy, September 20 - 24, 2004). Pages: 484-495.

[4] Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., and Verykios, V. 1999. Disclosure Limitation of Sensitive Rules. In Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange (November 07 - 07, 1999). KDEX. IEEE Computer Society, Washington, DC, 45.

[5] Jun-Lin Lin, Julie Yu-Chih Liu, Privacy preserving Itemset mining Through Fake Transactions, In proceedings of the 2007 ACM Symposium on Applied Computing, Seoul, Korea, Pages: 375-379.

[6] Agrawal, Shipra and Krishnan, Vijay and Haritsa, Jayant R (2004) On Addressing Efficiency Concerns in Privacy-Preserving Mining. Proceedings 4th International Conference on Database Systems for Advanced Applications: DASFAA '04 (LNCS), pages Vol.2973, 113-124, Jeju Island, Korea.

[7] Liu, K. and Ryan, J. 2006. Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining. IEEE Trans. on Knowledge and Data Eng. 18, 1 (Jan. 2006), 92-106.

[8] Jin-Lang Wang, Cong-fu Xu, and Yun-He Pan 2006. An Incremental Algorithm for mining privacy preserving frequent Itemsets. Proceedings of Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August, 2006.

[9] Chen, T. 2006. A Novel Method for Protecting Sensitive Knowledge in Association Rules Mining. In Proceedings of the Sixth international Conference on intelligent Systems Design and Applications (ISDA'06) - Volume 01 (October 16 - 18, 2006). ISDA. IEEE Computer Society, Washington, DC, 694-699.

[10] Aris Gkoulalas-Divanis, Vassilios S. Verykios. An integer programming approach for frequent itemset hiding, ACM, CIKM'06, November 5-11 2006, Arlington, Virginia, U.S.A.

[11] S. Oliveira and O. Zaïane. Privacy preserving frequent itemset mining. CRPITS'14: Proceedings of the IEEE International Conference on Privacy, Security, and Data Mining, pages 43–54, 2002.

[12] Oliveira, S. R. and Zaïane, O. R. 2003. Protecting Sensitive Knowledge by Data Sanitization. In Proceedings of the Third IEEE international Conference on Data Mining (November 19 - 22, 2003). ICDM. IEEE Computer Society, Washington, DC, 613.

[13] Oliveira, S. R. and Zaïane, O. R. 2003. Algorithms for balancing privacy and knowledge discovery in association rule mining. In Proceedings of the IEEE seventh International Database Engineering and Applications Symposium, 16-18 July 2003, Hong Kong, China. Pages 54-63.

[14] Y. Saygin, V. S. Verykios, and C. Clifton. Using Unknowns to Prevent Discovery of Association Rules. SIGMOD Record, 30(4):45–54, December 2001.

[15] http://fimi.cs.helsinki.fi/data/.

