Privacy Preserving Data Mining:

Case of association rules

Sarra Gacem[1], Djamila Mokeddem[2] and Hafida Belbachir[3]

[1] University of Science and Technology Mohamed Boudiaf, Department of Computer Science, Laboratory Systems Signals Data Oran, Algeria

[2] University of Science and Technology Mohamed Boudiaf, Department of Computer Science, Laboratory Systems Signals Data Oran, Algeria

[3] University of Science and Technology Mohamed Boudiaf, Department of Computer Science, Laboratory Systems Signals Data Oran, Algeria

Abstract

Data mining has become an important technology to discover a hidden and nontrivial knowledge from large amounts of data. A major problem is to achieve this discovery process with preserving privacy of extracted data and / or knowledge. Privacy preserving data mining (PPDM) is a new area of research that studies the side effects of knowledge mining methods on individuals and organizations privacy. We present in this paper a state of the art of the PPDM in the case of association rules. We propose taxonomy of existing techniques and a classification of work realized in this context. This synthesis is followed by a discussion of certain issues and perspectives.

Keywords: *Privacy, data hiding, rule hiding, association rules.*

1. Introduction

The advances in digital data storage devices allow to companies and organizations to store a large volume of data. Data mining field which focuses on the extraction of useful knowledge from large databases has seen considerable progress over the past two decades. Knowing that these electronic data cover all aspects of our lives (e.g. purchases with credit card, information (news), medical records, etc...), the goal remains how to ensure that data or information as sensitive will be kept away from any abuse. This question is crucial in data mining.

Privacy preserving data mining is a new area of research which treats the negative side of knowledge extraction techniques. We can classify privacy problems related to data mining application into two major categories; the first bonded to data called data hiding while the second concerns information or knowledge that data mining process may discover after analyzing the data, called knowledge hiding .data hiding seeks to delete a private or confidential data before their revelation. On the other side, knowledge hiding is interested to eliminate confidential knowledge that can be extracted from the data. We present in this paper synthesis of some work on the PPDM in the case of association rules.

The formal framework of association rules mining algorithms is as follow: Let I = {I1, I2... Im} a set of items. All $X \subseteq I$ is called an itemset. In addition, an itemset with k elements is called a k-itemset. Let D= {T1, T2…Tn} be a set of transactions where each transaction Ti ($i \subseteq [1 .. n]$) is an itemset. The support of an itemset $X \subseteq I$ in D, denoted supp (X), is defined as the percentage of transactions containing X in D. supp (X) = | x | / | D | (where | D | is the number of transactions in D). X is a frequent itemset if the support of X is greater than a minimum threshold predefined min_supp. An association rule is an implication the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \Phi$. the confidence of a rule is denoted by conf ($X \Rightarrow Y$), is defined as the percentage of transactions containing X that also contain Y in D. Conf ($X \Rightarrow Y$) = | XUY | / | D |. We say that the rule $X \Rightarrow Y$ holds in D if the confidence of the rule is greater than a predefined minimum threshold min_conf.

After this introduction, we discuss the main dimensions on which our synthesis is based. A classification of work in this context is detailed in Section 3. Subsequently, we

name some evaluation criteria of proposed algorithms. We conclude with a discussion about the techniques used and the problems still open.

2. The principal dimensions in the methods of the PPDM

There are various taxonomies in the domain of privacy preserving data mining .We propose synthesis of those works [22-25-17-8] based on the following dimensions:

- Data or rule hiding.
- Centralized or distributed data.
- Privacy preserving techniques.

A synthesis linking these three dimensions is presented in Table 1

Table 1: Classification of PPDM algorithms

Data / Rule Hiding	***Privacy preserving techniques***	***Centralized/ distributed data***
Data Hiding	SMC	Distributed data
	Data perturbation	Distributed and centralized data
	Data anonymization	Distributed and centralized data
Rule Hiding	Data distorsion	Distributed and centralized data
	Data blocking	
	Data reconstruction	

2.1 Data or rule hiding

Data hiding consists to develop new treatment methods of the original data in order that sensitive data stays protected during and after the mining process .Other shares, rule hiding(knowledge hiding) try to hide some sensitive knowledge in order that they cannot be discovered by data mining techniques.

2.2 Data distribution

The second dimension corresponds to the distribution data. Some approaches have been developed in a centralized environment where data are all stored in one database, while others were developed in a distributed environment where data resides in different databases. The data partitioning strategy can be horizontal (different record of the same data attributes is found in different sites) or vertical (different attributes of the same data record are situated in different places).

2.3 Privacy preserving techniques

This dimension refers to the methods used to preserve privacy. It is tightly related to two dimensions mentioned above. We present in the following basic principle of the techniques most used.

• **Data perturbation** consists to falsify the sensitive data by random noise based on statistical techniques.

• **Data anonymization** is to modify the data in such a way to remove any information that could directly link data to individuals.

• **Secure Multi-party Computation SMC**, the problem of secure multiparty computation is to allow calculating any function on a set of data distributed across multiple sites. Each site has part of the data and calculation should be realized in such a way that any party can deduct in some way the data of other sites from the calculation results and from its own data.

•**Data distortion,** This technique aims at minimizing or increasing the support /confidence of certain transactions below the threshold. The process is to change the value of the itemset from '1 'to '0' (Delete an itemset) or '0' to '1' (add an itemset) to hiding certain rules considered sensitive.

• **Data blocking,** in certain cases, the addition or deletion of items (distortion) can generate false data that can create side effects, as in the case of a Medical DB. Blocking approach is implemented for reducing support and confidence of sensitive rules by replacing certain items with a question mark "?".This special value unknown brings uncertainty to the, making the thresholds min_sup and Min_Conf two uncertain intervals respectively.

• **Data reconstruction,** this technique treats the inference problem in databases. It is to recreate a new database from the knowledge (rules) extracted, after selecting the sensitive rules that should not be generated after having done the data mining process of reconstructed database.

3. Existing work

Several works were developed in the area of PPDM. They can be classified according to the dimensions shown above:

3.1 Data hiding

3.1.1 Work based on the data perturbation in the centralized case

In the works [2-11], Additive perturbation method (AP) was proposed by adding noise to the original data in order to ensure privacy .The authors challenged the use of the additive noise by showing how in some cases the noise can be effectively filtered by revealing a good approximation of the original data. Another approach Multiplicative Perturbation (MP) is to multiply each data element by a random number that has a Gaussian distribution truncated with a small average and variance [12]. The random projection was proposed in [15] to preserve both the correlations between attributes and the Euclidean distance between data vectors by multiplying the original data with a random matrix with lower dimension [7-10].

3.1.2 Work based on the data perturbation in the distributed case

The authors in [14] use the random projection matrix as a tool to preserve the privacy of data distributed by using the lemma johnson-lindenstauss. They showed that the properties related to the statistical distance of the original data are still well preserved without disclosing the original data values after the process of the perturbation. They also show in their work that this technique can be successfully applied to the various tasks of data mining.

3.1.3 Work based on the data anonymization in the centralized case

The principle of K-anonymity approach is to hide K identities at a time so that sensitive data stays private [23] The authors in [14] showed how a lack of diversity among the sensitive attribute values can be used to link individuals and sensitive values and they asserted that the K-anonymity does not always guarantee data privacy. To remedy to this problem, they proposed a new definition of privacy called L-diversity [1]. Another method called (α, k)-anonymization [20] is to consider the relative frequency of sensitive value in each equivalence class that must be less or equal than α.

3.1.4 Work based on the data anonymization in the distributed case

In cases where data is distributed, the methods k-anonymity and L-diversity was adapted to this context in [13-14].

3.1.5 Work based on the SMC

SMC can be used to secure the transfers of partial results in the case of distributed data mining. The authors in [6] proposed a secure method based on the FDM algorithm proposed in [4] to maintain the confidentiality of mining association rules on horizontally data distributed. The goal is to keep the efficiency of this algorithm without any site does not disclose its local itemsets, its support value or its transactions size. In [21], the authors present a new method to compute the globally frequent itemsets from data sources horizontally distributed, while preserving the privacy of participating sites. Another method based on homomorphism encryption is to make a confidential correspondence (matching) and an intersection of set confidential, for two parts with data vertically partitioned [9].

Table 1: Classification of existing works in data hiding

		Distributed Data		*Centralized Data*
		Horizontal	***Vertical***	
Data Hiding	***SMC***	[21],[6]	[9]	
	Data perturbation	[16]		[2],[11], [12],[15], [7],[10]
	Data anonymization	[13], [14]		[1], [20], [23]

3.2 Rule hiding in centralized case

3.2.1 Work based on data distortion

The authors in [18] propose two algorithms RRA (The Round Robin Algorithm) and RA (Random Algorithm). The main idea is to eliminate the items from sensitive transactions to reduce the impact on the modified database and extract only the non-sensitive association rules. The authors in [26] propose an approach to distorted data by a matrix "sanitization matrix" to hide sensitive items.The work [27] presents two algorithms, ISL (Increase Support of LHS) and DSR (Decrease Support of RHS) to hide sensitive association rules. The ISL algorithm consists to increment the support of the sensitive rule by changing the left part, while the DSR algorithm decrements the Support by changing the right side of the rule. In order to reduce the modification rate of the original database caused by ISL and DSR, authors propose a new algorithm in [28]. The idea is to hide first rules which sensitive item (we want to hide) is in the right side then hide rules which the sensitive item is in the left side.

3.2.2 Work based on data blocking

The authors in [24] proposed three algorithms GIH, CR and CR2, the algorithm GIH hide sensitive itemset by reducing support below minimum threshold. However the algorithm CR and CR2 hides sensitive rules by reducing the confidence of the rule. The difference between the two algorithms is that the CR algorithm hides sensitive rules by replacing only the values of items equal to '0 ' with '? ' and CR2 hides sensitive rules by replacing only the values of items equal to '1 ' with '? '.

3.2.3 Work-based on data reconstruction

In [19], we present an extraction algorithm based on the approach CIILM (Constrained base Inverse Lattice Mining itemset) to hide sensitive frequent itemset using the lattice for the reconstruction of the new generated database. It is on this principle that the authors used in [30] to hide sensitive rules by using the FP-tree structure for the reconstruction of the new database.

Table 3: Classification of existing works in rule hiding

	Data Distortion	***Data Blocking***	***Data Reconstruction***
Itemset Hiding	RRA;RA[18] [26]	GIH[24]	CIILM[19]
Rule Hiding	[27] [28]	CR[24] CR2[24]	[30]

4. The evaluation criteria

In work related to PPDM there is no standard framework for evaluating the performance of the proposed algorithms. In this paper, we present a non exhaustive list the measures most used in the literature.

4.1 Calculation and communication Cost

Temporal complexity is an important factor in assessing the proposed algorithms. It measures the time required for transformation of specific set sensitive information. If distributed, communication load between sites should be minimal to ensure scalability of the algorithm.

4.2 Information loss

At the end of the process of preserving confidentiality, loss of information (knowledge) and data is an important question because that for sensitive information is hidden; the database must be modified in some cases.

4.2.1 Data loss

In the case of the insertion of false information (distortion) or through blocking of certain data values, the data loss can be measured by the following formula:
DISS (D,D')= $\frac{1}{\sum_{i=1}^{n} f_{D(i)}} \times \sum_{i=1}^{n} [f_{D(i)} - f_{D'(i)}]$ with D: the original database, D': the modified database, f D(i): the frequency of the item i in the original database D, f D'(i): the frequency of the item i in the modified database D '.

4.2.2 Knowledge (rules) Loss

Measuring the amount lost depending on the algorithm of data mining used. In the case of association rules loss of consciousness can be defined as variations of support and the confidence of all rules. It can be evaluated according to the following metrics:

4.2.2.1 Hiding failure

Represents the percentage of rules which are not sensitive hidden successfully, it can be calculated by the following formula: HF=|RS (D')|/|RS (D)|, such as Rs: represents the set of sensitive rules.

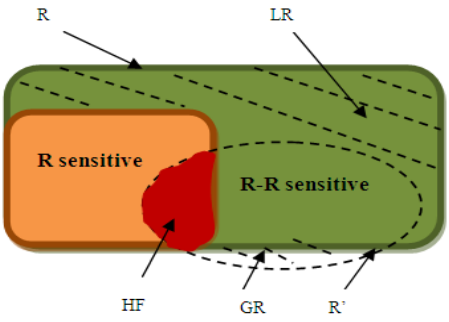
4.2.2.2 Lost rules ratio

Represents the percentage of non-sensitive rules wrongly hidden, it is calculated by the following formula:
LR = | RNS (D) | - | RNS (D) '| / | RNS (D) |, such as RN S: represents all non-sensitive rules.

4.2.2.3 Ghost rules ratio

Represents the percentage of new rules wrongly generated (ghosts), it is calculated by the following formula:
GR = (| R '| - | R ∩ R' |) / | R ') |, such that R and R' represent respectively the set of association rules that can be generated from D and D '.

FIG1: Tree metric of knowledge (rules) Loss



4.3 The level of confidentiality (resistance)

The resistance of the algorithms PPDM is linked to opportunities for scaling the algorithm to different techniques of data mining. The objective is to protect sensitive information against disclosure. In this case, it is important do not forget that intruders try to compromise the information using various data mining algorithms. Thus an algorithm developed PPDM against a particular technique of data mining and ensuring the preservation of confidentiality may not achieve similar protection against all possible algorithms.

5. Discussion and perspectives

Any existing algorithms are more efficient than all others on all possible criteria. Rather an algorithm may be more optimal than another on specific criteria such as complexity.

Perturbed data preserve individual privacy, but they are problematic in data mining. Two crucial questions to ask are: How to extract information from randomized data and how the results based on randomized data are comparable to results probable of the original data base.

In terms of computation efficiency, rule hiding (distortion, locking and reconstruction) are less effective than those used in data hiding. Before we hide a sensitive rule, we must first identify the corresponding itemsets by accessing the database. Some researchers found that hiding the rules is more critical than hide data because it is possible to infer confidential information from sensitive rules that are not hidden successfully.

In term of information loss, the reconstruction method has a high rate in data loss, but it allows for a minimum rate of loss rules. The problem to be able to infer sensitive knowledge is placed in the majority of approaches. Based on the work presented in this paper, the two issues (hide data or rules) are treated separately. We think it is important to develop algorithms that can treat both problems at a time.

In the era of cloud computing, privacy preserving in distributed domain should have more attention given the complexity of the problem. In effect, the majority of work is based on cryptographic technique SMC that is very costly in terms of communication cost. In addition, all these works only deal with the problem of data hiding distributed either vertically or horizontally. Practical, we may need to share data with preserving privacy of certain knowledge. This problem may be further complicated with a hybrid partitioning data (horizontal and vertical).

6. Conclusion

After maturity of many data mining algorithms such as association rules, we try in recent years to answer the question: how can search without disclosing? In this work we proposed taxonomy and a detailed description of various PPDM algorithms in order to explore issues research. The work presented shows the growing interest of researchers in the area of data security and knowledge privacy.

References

[1] A.Machanavajjhala,J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity Privacy beyond k-anonymity", ACMTransactions on Knowledge Discovery from Data, 1(1), 2006.

[2] Agrawal, R., & Srikant, R, " Privacy preserving data mining.", Paper presented at the 2000 ACM SIGMOD Conference on Management of Data, pp. 439-450,2000.

[3] Aggarwal, , Yu, P.S.," Privacy-Preserving Data Mining: Models and Algorithms", pp. 267–286. Springer, Heidelberg (2008)

[4] Cheung, D., Ng, V., Fu, A. & Fu, "Efficient Mining of Association Rules in Distributed Databases", IEEE Transactions on Knowledge and Data Engineering. 8(6), 911-922,1996.

[5] Chirag Modi, U.P. Rao, and Dhiren R. Patel," A Survey on Preserving Privacy for Sensitive Association Rules in Databases",In: BAIP, Vol. 70Springer (2010) , p. 538-544.

[6] Clifton, C., Kantarcioglu, M. & Vaidya, J , "Defining privacy for data mining". Book Chapter Data Mining, Next generation challenges and future directions.

[7] Domingo-Ferrer, J., & Mateo-Sanz, J. M., "Practical data-oriented microaggregation for statistical disclosure control",

IEEE Transactions on Knowledge and Data Engineering, 14(1), 189-201,2002.

[8] E. Bertino, I. Nai Fovino , L.P. Provenza, "A Framework for Evaluating Privacy Preserving Data Mining Algorithms", 18 August 2005.

[9] Freedman, M.J., Nissim, K. and Pinkas, "Efficient private matching and set intersection', International Association for Cryptologic Research (IACR), Interlaken,Switzerland, 2–6 May 2004.

[10] Hansen, S. L., & Mukherjee," A polynomial algorithm for optimal univariate microaggregation". Knowledge and Data Engineering, IEEE Transactions on, 15(4), 1043-1044,2003.

[11] Huang, Z., Du, W., & Chen, " Deriving private information from randomized data",In Proceedings of the 2005 ACM SIGMOD international conference on management of data, pp. 37-48).

[12] Kim, J. J., & Winkler, W. E," Multiplicative noise for masking continuous data",Washington D.C.: Statistical Research Division, U.S.Bureau of the Census,2003.

[13] L SWEENEY," k-anonymity: a model for protecting privacy". International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002,10(5):557-570.

[14] LI Ninghui, LI Tiancheng. "t-Closeness: Privacy Beyond k-Anonymity and i-Diversity", IIProc of 23rd Int'l Conf. on Data Engineering. 2007:106- ll5.

[15] LIU. Kun,H .Kargupta, J. Ryan, "Random Projection based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining", Knowledge and Data Engineering, 2006,18(1):92-106.

[16] Liu, K., Kargupta, H., & Ryan, " Random projection-based multiplicative data perturbation for privacy preserving distributed data mining", IEEE Transactions on Knowledge and Data Engineering, 18(1), 92-106.2006.

[17] Madhu V. Ahluwalia, Aryya Gangopadhyay,"computer security,Privacy,and Politics", Book Chapter Privacy.Preserving. Data.Mining: Taxonomy.of.Existing.Techniques.

[18] Oliveira, S.R.M. and Zaïane, O.R. Algorithms for balancing privacy and knowledge discovery in association rule mining. In: Proc. of the 7th P Int'l Database Engineering and Applications Symposium (IDEAS'03). IEEE Computer Society, 2003. 54-63.

[19] Orlowska, Chen, X., M., and Li, X, "A new framework for privacy preserving data sharing", In: Proc. of the 4th P IEEE ICDM Workshop: Privacy and Security Aspects of Data Mining. IEEE Computer Society, 2004. 47-56.

[20] R. Chi-Wing Wong, J. Li, A. Wai-Chee Fu, and K. Wang, " (α, k)- anonymity: an enhanced k-anonymity model for privacy preserving data publishing", In Proceedings of the 12th ACM SIGKDD Conference (KDD'06), pages 754–759, Philadelphia, PA, August 2006.

[21] Rajalakshmi, M., Purusothaman ,T. and Pratheeba, S. "Collusion-Free Privacy Preserving Datamining", International Journal of Intelligent Information Technologies (IJIIT), Vol.6, No. 4,pp.30-45, October 2010.

[22] S. Verykios, E. Bertino, I. Fovino ,"State-of-the-art in Privacy Preserving Data Mining",2001.

[23] Samarati, P. & Sweeney," Protecting privacy when disclosing information: k-Anonymity and its enforcement through generalization and suppression", Proceedings of the IEEE Symposium on Research in Security and Privacy,1998.

[24] Saygin, Y., Verykios, V.S., and Clifton, C, "Using unknowns to prevent discovery of association rules". SIGMOD Record, 2001, 30(4):45-54.

[25] W.Xiaodan , C.Chao-Hsien , W.Yunfeng , L.Fengli , Y.Dianmin ," Privacy Preserving Data Mining Research:Current Stat,"us and Key Issues",2007.

[26] Wang, E.T., Lee, G., and Lin, Y.T, "A novel method for protecting sensitive knowledge in association rules mining",In: Proc. of the 29th Annual Int'l Computer Software and Applications Conf.. IEEE Computer Society, 2005.

[27] Wang, S.L., B. Parikh and A. Jafari." Hiding informative association rule sets". Exp. Syst. Appli., 33: pp. 316-323, 2007.

[28] Yogendra kumar jain , Vinod kumer yadac et Geetika S.panday ; "An efficient association rule hiding- algorithm for privacy preserving data mining" ; 7 july 2011.

[29] Yong Yin, Ikou Kaku, Jiafu Tang et JianMing Zhu," Application for Privacy-preserving Data Mining, Decision Engineering, springer , 2011

[30] Yuhong Guo ,"reconstruction-based association rule hiding" ,june 10,2007,china.