An Overview of Privacy Preserving Data Mining

by Aris Gkoulalas-Divanis and Vassilios S. Verykios

Introduction

Significant advances in data collection and data storage technologies have provided the means for inexpensive storage of enormous amounts of data in data warehouses that reside in companies and public organizations. Apart from the benefit of using this data per se (e.g., for keeping up-to-date profiles of customers and their purchases, maintaining a list of available products, their quantities and price, etc.), the mining of these datasets with existing data mining tools [5] can reveal invaluable knowledge that was unknown to the data holder beforehand. The extracted knowledge patterns can provide insight to the data holders as well as be invaluable in tasks such as decision-making and strategic business planning. Moreover, companies are often willing to collaborate with other entities who conduct similar business toward the mutual benefit of their industries. Significant knowledge patterns can be derived and shared among the collaborative partners through the aggregate mining of their datasets. Furthermore, public sector organizations and civilian federal agencies usually have to share a portion of their collected data or knowledge with other organizations having a similar purpose, or even make this data and knowledge public. For example, the National Institutes of Health (NIH) endorses research that leads to significant findings, which improves human health and provides a set of guidelines that sanction the sharing of NIH-supported research findings with research institutions.

As it becomes evident, there exists an extended set of application scenarios in which information or knowledge derived from the data must be shared with other (possibly untrusted) entities. The sharing of data and/or knowledge may come at a cost to privacy, primarily due to two reasons:

- If the data refers to individuals, e.g., as in customers' market basket data, then the disclosure of this data or any knowledge extracted from the data may potentially violate the privacy of the individuals if their identity is revealed to untrusted third parties.
- If the data contains business (or organizational) information, then the disclosure of this data or any knowledge extracted from the data may potentially reveal sensitive trade secrets, whose knowledge can provide a significant advantage to competitors and could cause the data holder to lose business.

The aforementioned privacy issues encountered in the course of data mining are amplified due to the fact that untrusted entities (adversaries) may utilize other external and publicly available sources of information, e.g., the yellow pages or public reports, in conjunction with the released data or knowledge, in order to reveal sensitive information.

Since its inception in 2000 with the pioneering work of Agrawal & Srikant [3] and Lindell & Pinkas [6], privacy preserving data mining has gained increasing popularity in the data mining research community. As a result, a new set of approaches was introduced to allow for data mining, while, at the same time, prohibiting leakage of private and sensitive information. Most existing approaches can be classified into two categories: methodologies that protect the *sensitive data itself* in the mining process, and methodologies that protect the *sensitive data mining results*, i.e., the extracted knowledge, that were produced by the application of data mining.

The first category refers to methodologies that apply such techniques as perturbation, sampling, generalization/suppression, transformation, etc., to the original datasets in order to generate their sanitized counterparts that can be safely disclosed to untrusted third parties. The goal of this category of approaches is to enable data miners to obtain accurate results when they are not provided with the real data. As a part of this category, we highlight Secure Multiparty Computation methodologies that have been proposed to enable a number of data holders to collectively mine their data without having to reveal their datasets to each other. On the other hand, the second category deals with techniques that prohibit the disclosure of sensitive knowledge patterns derived through the application of data mining algorithms, as well as techniques for downgrading the effectiveness of the classifiers in classification tasks, such that they do not reveal sensitive knowledge. In what follows, we further investigate each of these two categories of approaches.

Protecting Sensitive Data

A wide range of methodologies have been proposed in research literature to effectively shield the sensitive information contained in a dataset by producing its privacy-aware counterpart that can be safely released. The goal of these privacy preserving methodologies is to ensure that the sanitized dataset (a) properly shields all the sensitive information that was contained in the original dataset and (b) has properties similar to the original dataset, e.g., first/second order statistics, etc., possibly resembling it to a high extent, and (c) maintains reasonably accurate data mining results when compared to those attained when mining the original dataset.

The protection of sensitive data from disclosure has been extensively studied in the context of **microdata** release, where methodologies have been proposed for the protection of sensitive information regarding individuals recorded in some dataset. In microdata, we consider each record in the dataset to represent an individual for whom the values of a number of attributes are being recorded, e.g., name, date of birth, residence, occupation, salary, etc. Among the complete set of attributes, there exists some attributes that explicitly identify the individual, e.g., name, social security number, etc., as well as attributes that, once combined together or with publicly available external resources, may lead to the identification of the individual, e.g., address, gender, age, etc. The first type of attributes, also known as **identifiers**, must be removed from the data prior to its publishing. On the other hand, the second type of attributes, also known as **quasi-identifiers**, have to be handled by the privacy preservation algorithm in such a way that in the sanitized dataset, the knowledge of their values regarding an individual no longer poses a threat to the identification of his/her identity.

Existing methodologies for the protection of sensitive microdata can be partitioned in two directions: (a) data modification approaches and (b) synthetic data generation approaches. Willenborg & DeWaal [9] further partition the data modification approaches into **perturbative** and **non-perturbative** approaches, depending on whether they introduce false information in the attribute-values of the data, e.g., by the addition of noise based on some data distribution, or they operate by altering the precision of the existing attribute-values, e.g., by changing a value to an interval that contains it.

Secure Multiparty Computation

The approaches discussed so far aim at generating a sanitized dataset from the original one, which can be safely shared with untrusted third parties, as it contains only non-sensitive data. Secure Multiparty Computation (SMC) provides an alternative family of approaches that can effectively protect the sensitive data. SMC considers a set of collaborators who wish to collectively mine their data but are unwilling to disclose their own datasets to each other. As it turns out, this distributed privacy preserving data mining problem can be reduced to the secure computation of a function based on distributed inputs and is thus solved by using cryptographic approaches. Pinkas [8] elaborates on this close relation that exists between privacy-aware data mining and cryptography. In SMC, each party contributes to the computation of the secure function by providing its private input. A secure cryptographic protocol that is executed among the collaborating parties ensures that the private input that is contributed by each party is not disclosed to the others. Most of the applied cryptographic protocols for multi-party computation result to some primitive operations that have to be securely performed: secure sum, secure set union, and secure scalar product. Clifton et al. [4] discuss these operations.

As a final remark, we should point out that the operation of the secure protocols in the course of distributed privacy preserving data mining depends highly on the existing distribution of the data in the sites of the collaborators. Two types of data distribution have been investigated: In a **horizontal data distribution**, each collaborator holds a number of records, and for each record, he/she has knowledge of the same set of attributes as his/her peers. On the other hand, in a **vertical partitioning** of the data, each collaborator is aware of different attributes referring to the same set of records.

Protecting the Sensitive Knowledge

In this section, we focus our attention on privacy preserving methodologies that protect the sensitive knowledge patterns that would otherwise be revealed after the course of the mining process. Akin to the methodologies that we have presented for protecting sensitive data prior to its mining, the approaches in this category also modify the original dataset, but thy do so in such a way that certain sensitive knowledge patterns are suppressed when mining the data. In what follows, we briefly discuss some categories of methodologies that have been proposed for hiding the sensitive knowledge in the context of association and classification rule mining.

Association Rule Hiding

The association rule mining framework, along with some computationally efficient heuristics for the generation of association rules, have been proposed in the work of Agrawal et al. [1] and Agrawal & Srikant [2]. Briefly stated, the goal of association rule mining is to produce a set of interesting and potentially useful rules, i.e., implications that hold in a dataset. The presence of a rule in a dataset is judged on its statistical significance, quantified with the aid of two measures: *confidence* and *support*. All the association rules whose confidence and support are above some user-specified thresholds are then mined. However, some of these rules may be sensitive from the owner's perspective. The association rule hiding methodologies aim to sanitize the original dataset in a way that:

- All the sensitive rules (as indicated by the data holder) that appear when mining the original dataset for association rules do not appear when mining the sanitized dataset for association rules at the same (or higher) levels of support and confidence.
- All the non-sensitive (remainder) rules can be successfully mined from the sanitized dataset at the same (or higher) levels of support and confidence.
- No rule that was not found when mining the original dataset can be found in its sanitized counterpart when mining the latter at the same (or higher) levels of support and confidence.

The first goal simply states that all the sensitive association rules are properly hidden in the sanitized dataset. The hiding of the sensitive knowledge comes at a cost of the utility of the sanitized outcome. The second and the third goals aim at minimizing this cost. Specifically, the second goal requires that only the sensitive knowledge is hidden in the sanitized dataset. Thus, no other non-sensitive rules are lost due to side-effects of the sanitization process. The third rule requires that no artifacts, i.e., false association rules, are generated by the sanitization process. To recapitulate, in association rule hiding, the sanitization process has to be accomplished in a way that minimally affects the original dataset, preserves the general patterns and trends of the dataset, and achieves concealment of all the sensitive knowledge, as indicated by the data holder.

Association rule hiding has been studied along three principal directions: (a) heuristic approaches, (b) border-based approaches, and (c) exact approaches. The first direction collects time and memory efficient algorithms that heuristically select a portion of the records of the original dataset to sanitize in order to facilitate sensitive knowledge hiding. Due to their efficiency and scalability, these approaches have been investigated by the majority of the researchers in the knowledge hiding field of privacy preserving data mining. However, as in all heuristic methodologies, the approaches of this category make locally optimal decisions when performing knowledge hiding, which may not always be (and usually are not) globally optimal. As a result, there are several cases in which these methodologies suffer from undesirable side-effects and may not identify optimal hiding solutions when such solutions exist. Heuristic approaches can rely on a **distortion**, i.e., inclusion/exclusion of items from selected transactions, or on a **blocking** scheme, i.e., replacing some of the original values in a transaction with question marks.

The second class of approaches collects methodologies that hide sensitive knowledge by modifying only a selected portion of itemsets that belong to the border in the lattice of the frequent (statistically significant) and the infrequent (statistically insignificant) patterns of the original dataset. In particular, the sensitive knowledge is hidden by enforcing the revised borders, which accommodate the hiding of sensitive itemsets, in the sanitized database. The algorithms in this class differ in the borders that they track as well as in the methodology that they apply to enforce the revised borders in the sanitized dataset. An analysis regarding the use of borders in association rule mining can be found in the work of Mannila and Toivonen [7].

Finally, the third class of approaches involves non-heuristic algorithms that conceive the knowledge hiding process as a constraints satisfaction problem (an optimization problem) that is solved through the application of integer or linear programming. This class of approaches differs from the previous two, primarily due to the fact that it collects methodologies that can guarantee optimality in the computed hiding solution (provided that an optimal hiding solution exists) or a very good approximate solution (in the case that an optimal one does not exist). However, these approaches are usually several orders of magnitude slower than the heuristics, especially due to the runtime that is required for the solution of the constraints satisfaction problem by the integer/linear programming solver.

Classification Rule Hiding

Privacy-aware classification has been studied to a substantially lower extent than privacy preserving association rule mining. Similar to association rule hiding, classification rule hiding algorithms consider a set of classification rules as sensitive and proceed to protect them from disclosure by using either **suppression-based** or **reconstruction-based** techniques. In suppression-based techniques, the confidence of a classification rule (measured in terms of the owner's belief regarding the presence of the rule when given the data) is reduced by distorting a set of attributes in the dataset that belong to transactions related to its existence. On the other hand, reconstruction-based approaches target reconstructing the dataset by using only those transactions of the original dataset that support the non-sensitive classification rules, thereby leaving the sensitive rules unsupported.

Future Trends

Data mining is a rapidly evolving field counting numerous conferences, journals, and books that are dedicated to this area of research. As new forms of data emerge, as well as new application areas and research challenges arise, it becomes evident that innovative privacy preserving data mining methodologies will also have to be proposed to keep pace with this progress. The current applications of privacy preserving data mining are numerous, spanning from the offering of privacy in the release of medical and genomic databases to the extraction of knowledge patterns that provide information related to homeland security. Mobility data mining and privacy-aware stream data mining, are among the most recent and prominent directions of privacy preserving data mining. As spatiotemporal and geo-referenced datasets grow, a novel class of applications is expected to appear that will be based on the extraction of behavioral patterns of user mobility. Clearly, in these applications, privacy is a major concern. Thus, novel privacy preserving methodologies will have to be proposed to protect those patterns that are sensitive with respect to the privacy of individuals. Another example of a research domain that is expected to receive much attention in upcoming years is privacy in the context of applications where data is released incrementally and in an unconditional rate. In this challenging area of research, privacy preserving data mining methodologies have to be designed to handle streams of data rather than datasets containing historical recordings. Finally, apart from domain-driven research, there is currently an urgent need for the development of frameworks that will unify more advanced measures for the evaluation and the comparison of different privacy preserving data mining methodologies.

Conclusion

In this article, we presented a brief introduction to the area of privacy preserving data mining, one of the most popular directions in today's data mining research community. In the first part of the article, we highlighted the necessity of this research area and its numerous applications. Then, we discussed the different categories of approaches that have been proposed for the protection of either the sensitive data itself in the course of data mining or the sensitive data mining results, after the application of data mining. Finally, we provided some roadmaps based on the most promising future directions of this area.

Key Terms

Privacy Preserving Data Mining

The area of data mining that is concerned with privacy issues related to the course of data mining, specifically (a) with the protection of privacy in data releases, (b) the preservation of privacy in the mutual mining of data among a set of collaborating parties, and (c) with the protection of sensitive knowledge patterns that can be derived due to the application of data mining tools.

Sanitization

The process of transforming the original dataset to its privacy-aware counterpart that can be safely released, as it protects the sensitive data or properly shields the sensitive knowledge from unauthorized exposure.

Rule Hiding Approaches

A category of methodologies that aims at protecting the sensitive knowledge that can be mined from a dataset in the form of sensitive association or classification rules. Rule hiding approaches primarily operate by modifying the original dataset such that the significance of the sensitive rules deteriorates in its sanitized counterpart to such an extent that they are no longer mined by the employed rule mining strategy.

Perturbation Techniques

A category of data modification approaches that protect the sensitive data contained in a dataset by modifying a carefully selected portion of attribute-values pairs of its transactions. The employed modification makes the released values inaccurate, thus protecting the sensitive data, but it also achieving preservation of the statistical properties of the dataset, e.g., the first and second order statistics, such that its mining yields accurate results.

Reconstruction Approaches

A category of sensitive knowledge hiding methodologies that operate by generating a new dataset based on a portion of the transactions of the original dataset, i.e., those transactions that support the non-sensitive knowledge. This category of approaches has been studied in the context of classification rule hiding. The transactions of the original dataset that support the non-sensitive rules are used to build a classification model from which the transactions of the sanitized dataset are generated.

Secure Multiparty Computation

A research direction within the area of privacy preserving methodologies for the protection of sensitive data. Secure multiparty computation collects distributed privacy preserving methodologies that enable a number of collaborating peers to collectively mine their data without having to reveal their datasets to each other. The approaches of this category operate by employing a family of protocols that allow the peers to exchange data in a secure manner. The security of the protocols is achieved through applying cryptographic approaches, which enables secure computation of a function based on distributed inputs.

References

1. Agrawal, R., Imielinski, T., and Swami, A. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*. 207-216.
2. Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the International Conference on Very Large Data Bases* (*VLDB*). 487-499.
3. Agrawal, R. and Srikant, R. 2000. Privacy preserving data mining. In *Proceedings of the ACM SIGMOD Conference on Management of Data*. 439-450.
4. Clifton, C., Kantarcioglou, M., Lin, X., and Zhu, M. 2002. Tools for privacy preserving distributed data mining. *ACM SIGKDD Explorations 4*, 2. 28-34.
5. Kdnuggets.com. 2009. Software for data mining, analytics, and knowledge discovery. **http://www.kdnuggets.com/software/index.html** (accessed April 1, 2009).
6. Lindell, Y. and Pinkas, B. 2000. Privacy preserving data mining. *J. Cryptology 15*, 3. 36-54.
7. Mannila, H. and Toivonen, H. 1997. Levelwise search and borders of theories in knowledge discovery. *Data Min. Knowl. Discov. 1*, 3. 241-258.
8. Pinkas, B. 2002. Cryptographic techniques for privacy preserving data mining. *ACM SIGKDD Explorations 4*, 2. 12-19.
9. Willenborg, L. and DeWaal, T. 2001. *Elements of Statistical Disclosure Control*. Springer-Verlag, Berlin, Germany.

Biography

Aris Gkoulalas-Divanis has a BS (University of Ioannina), an MS (University of Minnesota), and a PhD (University of Thessaly) in computer science. He is currently a postdoctoral research fellow in the Department of Biomedical Informatics at Vanderbilt University. He has served as a research assistant in both the University of Minnesota (2003-2005) and the University of Manchester (2006). His research interests are in the areas of privacy preserving data mining, privacy in medical records, and privacy in location-based services.

Vassilios S. Verykios received the Diploma degree in computer engineering from the University of Patras and MS and PhD degrees from Purdue University. Since 2005, he has been an assistant professor in the Department of Computer and Communication Engineering, University of Thessaly. He has also served on the faculty of Athens Information Technology Center, Hellenic Open University, and University of Patras. He has published more than 40 papers in major referred journals, conferences, and workshops.