SCITECH
RESEARCH ORGANISATION|

Volume 7, Issue 1
January 02, 2018

Journal of Information Sciences and Computing Technologies
www.scitecresearch.com/journals

Concealment Conserving the Data Mining of Groups & Individual

Abu Sarwar Zamani [1], Md. Mobin Akhtar [2], Danish Ahamad [3]
[1] *Lecturer, Dept. of Computer Science, College of Science & Humanity, Al Quwaiyah, Shaqra University, Kingdom of Saudi Arabia, sarwar_zamani@yahoo.com*
[2] *Lecturer, College of Computing and Information Technology, Shaqra University, Kingdom of Saudi Arabia, jmi.mobin@gmail.com*
[3] *College of Science and Arts, Sajir, Shaqra University, Kingdom of Saudi Arabia, danish.ahamad@gmail.com*

Abstract

We present an overview of privacy preserving data mining, one of the most popular directions in the data mining research community. In the first part of the chapter, we presented 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, in the context of traditional (relational) datasets. Following that, in the second part of the chapter, we focused our attention on one of the most recent as well as prominent directions in privacy preserving data mining: the mining of user mobility data. Although still in its infancy, privacy preserving data mining of mobility data has attracted a lot of research attention and already counts a number of methodologies both with respect to sensitive data protection and to sensitive knowledge hiding. Finally, in the end of the chapter, we provided some roadmap along the field of privacy preserving mobility data mining as well as the area of privacy preserving data mining at large.

Keywords: Traditional Datasets; Mobility Data.

1. Introduction

The significant advances in data collection and data storage technologies have provided the means for the inexpensive storage of enormous amounts of data in data warehouses that reside in companies and public organizations. Despite the benefit of using this data per se (e.g. for maintaining up to date profiles of the customers and record of their recent or historical purchases, maintaining an inventory of the available products, as well as their quantities and price, etc), the mining of these datasets with the existing data mining tools can reveal invaluable knowledge that was unknown to the data holder beforehand. The extracted knowledge patterns can provide insight to the data holders and at the same time can be invaluable in tasks such as decision making and strategic planning. Moreover, private companies are often willing to collaborate with other entities who conduct similar business, towards the mutual benefit of their businesses. Significant knowledge patterns can be derived and shared among the collaborative partners with respect to the collective 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 Institute of Health (NIH) endorses research that leads to significant findings which improve human health and provides a set of guidelines which 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 has to be shared with other (possibly untrusted) entities. Public agencies for example collect data for different purposes like population surveys, epidemiological and clinical studies, as well as various other social and economic experiments to answer a variety of problems that disturb the society as a whole. The sharing of data and/or knowledge may come at a cost to privacy, primarily due to two reasons: (a) if the data refers to individuals (e.g. as in customers' market basket data, medical data, preferences data and the like) then the disclosure of this data or any knowledge extracted from the data can potentially violate the privacy of the individuals if their identity is revealed to untrusted third parties, and (b) if the data regards 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 business competitors and thus can cause the data holder to lose business over his/her peers. The aforementioned privacy issues in the course of data mining are amplified due to the fact that untrusted entities (adversaries and data terrorists) may utilize other external and publicly available sources of information (e.g. the yellow pages, public reports) in conjunction with the released data or knowledge, in order to reveal even more protected sensitive information.

2. Main Pierce of the Chapter

Privacy preserving data mining is a new research area inspired by the need of scientists to analyze, interrogate and utilize row collected data without harming the privacy of the subjects contained in the data itself. In the sequel we give an overview of privacy preserving data mining approaches proposed for the protection of sensitive traditional forms of data like textual data. We have selected for presentation in this section techniques classified as perturbative, non-perturbative and secure multiparty computation. The second part in the main thrust is devoted to techniques related to protecting sensitive patterns from mining. In this part we focus our attention on two paradigms, the so-called association rule hiding and classification rule hiding. The third and last part is cornerstone to the significance of our book chapter since it is related to addressing state-of-the-art issues in privacy preserving data mining like privacy aware mobility data mining. The approaches presented in this part include approaches like data perturbation and obfuscation, secure multipart computation approaches and sequential pattern hiding approaches.

3. Protecting Traditional Sensitive Data during Mining

A wide range of methodologies have been proposed in the 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 all these privacy preserving methodologies is to ensure that the distorted (also known as sanitized) dataset (a) properly shields all the sensitive information that was contained in the original dataset, (b) has similar properties (e.g. first/second order statistics, etc) to the original dataset – 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) when mined.

3.1 Perturbative Approaches

In data perturbation approaches the attribute values of the original dataset are modified in a way that the released values are inaccurate. Several data perturbation approaches have been proposed in the research literature; the most prevalent ones can be partitioned under the following directions: (a) the addition of noise based on an underlying distribution (see Brand, 2002 for a detailed presentation of the methodologies in this direction), (b) the use of micro aggregation, in which the data records are partitioned into groups of either fixed.

3.2 Non-Perturbative Approaches

In non- perturbative approaches the attribute values of the original data are altered in a way that affects the precision in which they are released in the sanitized dataset. The most prevalent non perturbative methodologies are sampling and global recoding. For example, in a categorical dataset attribute marital status can take the value “married” for one record and “divorced” for another record of the original dataset, while it can be substituted with “been married” in both records in the sanitized counterpart. Probably the most important approach that ecompases global recoding is K-anonymity (see Samarati, 2001 and Samarati & Sweeney, 1998). The K-anonymity principle requires that every record in the sanitized dataset is indistinguishable from at least K-1 other records with respect to a set of identifying variables that formulate the quasi-identifier. Several algorithms have been proposed to enforce K-anonymity, while several of its variations have been explored.

3.3 Secure Multiparty Computation Approaches

The two previous categories of approaches 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 it thus solved by using cryptographic approaches.

4. Protecting Sensitive Patterns from Mining

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 mining the data. Similarly to the methodologies that

we have presented so far for protecting the sensitive data prior to its mining, the methodologies of this category also modify the original dataset but in such a way that certain sensitive knowledge patterns are suppressed, when mining the data. In what follows, we discuss methodologies that have been proposed for the hiding of sensitive knowledge in the context of association and classification rule mining.

4.1 Association Rule hiding

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 to 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 and thus no other, non-sensitive rules are lost due to side-effects of the sanitization process. On the other hand, 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 to conceal all the sensitive knowledge, as indicated by the data holder. The problem of association rule hiding has been studied along three directions, namely (a) heuristic approaches, (b) border-based approaches, and (c) exact approaches. The first class of approaches collects time and memory efficient algorithms that heuristically select a portion of the transactions 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 take locally best decisions when performing knowledge hiding, which may not always be (and usually are not) globally best.

4.2 Classification Rule hiding

Privacy-aware classification has been studied to a substantially lower extent than privacy preserving association rule mining. Similarly 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 holding 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. A system that is based on former category of approaches was proposed by Moskowitz & Chang (2000). On the other hand, reconstruction based reconstruction based approaches target at reconstructing the dataset by using only those transactions of the original dataset that support the non-sensitive classification rules.

5. Privacy Aware Mobility Data Mining

The remarkable advances in telecommunications and in location tracking technologies, such as GPS, GSM and UMTS, have made possible the tracking of mobile devices (and thus their human companions) at an accuracy of a few meters, at an affordable cost. From this perspective, we have nowadays the means of collecting, storing and processing mobility data of unprecedented quantity, quality and timeliness. The movement traces, left by the mobile devices of the users, are an excellent source of information that can aid towards decision making in mobility-related issues, such as urban planning, traffic analysis, forecasting of traffic-related phenomena, and timely detection of problems that emerge from users' movement behavior. On the other hand, it becomes evident that on the wrong hands this type of emergent knowledge may lead to an abuse scenario, as the mobility data may reveal highly sensitive personal information. Some examples of misuse include, but are not limited to, user tailing, surveillance or even unsolicited advertising. The existing so far methodologies can be partitioned in two broad categories: (a) methodologies that protect the sensitive data related to user mobility prior to the course of data mining, and (b) methodologies that hide sensitive knowledge patterns that summarize user mobility, which are identified as a result of the application of data mining. The first category of approaches collects data perturbation and obfuscation methodologies that distort the original dataset to facilitate privacy-aware data publication, as well as distributed privacy-aware methodologies for secure multiparty computation. On the other hand, the second category of approaches treats the mobility data as sequential data and applies a sequential pattern hiding strategy to prevent the disclosure of the sensitive sequential patterns in the course of sequential pattern mining. After the application of these approaches, only the non-sensitive patterns, summarizing user' movement behavior, survive the mining process, while the sensitive ones are suppressed in the data mining result. In what follows, we present in detail some of the approaches that have been proposed along each of these three categories.

6. Data Perturbation and Obfuscation

Data perturbation and obfuscation approaches aim at sanitizing a dataset containing user mobility data, in such a way that an adversary can no longer match the recorded movement of each user to a particular individual (thus reveal the identity of the user based on his or her recorded movement in the sanitized dataset). In what follows, we

consider that user mobility is captured as a set of trajectories (one per user) that depict the locations and times in the course of his or her history of movement. We assume that these location/ time recordings occur at a reasonably high rate that allows the tracking of user movement in the original dataset. For example, an adversary could use these recordings to track the user down to his/her house or place of work, even if the user trajectory was not accompanied by an explicit user identifier, such as the user id, the social security number or even the name of the user.

7. Secure Multiparty Computation

Secure multiparty computation has also been studied in the context of user mobility (and more generally on spatiotemporal) data. Inan & Saygin (2006) were the first authors to propose a privacy-aware methodology that clusters a set of spatiotemporal datasets, owned by different parties. To perform clustering, a similarity measure is necessary in order to quantify the proximity between two objects (e.g. the user trajectories), such that in the computed clustering solution, the co-clustered objects are more similar to one another than to objects belonging in different clusters. a secure protocol that can be employed to enable the pair wise secure computation of trajectory similarity among all the trajectories of the different parties, thus building a global matrix of trajectory similarity. By using this matrix, a trusted third party can perform the clustering on behalf of the users and communicate the computed clustering results back to the collaborating parties. The proposed privacy preserving protocol supports all the necessary basic operations for the computation of trajectory similarity based on widely adopted trajectory comparison functions: (a) Euclidean distance, (b) longest common subsequence, (c) dynamic time warping, and (d) edit distance.

8. Sequential Pattern hiding

The extraction of frequent patterns from mobility data has primarily concentrated on the sequential nature of such datasets by extracting frequent subsequences of user mobility (e.g. Cao, et al. 2005, Giannotti, et al. 2006). Giannotti, et al. (2007) proposed the integration of spatial and temporal information in the extracted mobility patterns by temporally annotating the extracted sequences, depicting frequent movement, with the transition times from one element (place of interest) to another. In a similar manner, the approaches that have been proposed for the hiding of frequent mobility patterns consider knowledge hiding in the form of sequential pattern hiding.

9. Conclusion

This line of research can be primarily attributed to the growing concern of individuals, organizations and the government regarding the violation of privacy in the mining of their data by the existing data mining technology. In the first part of the chapter, we presented 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, in the context of traditional (relational) datasets. In the second part of the chapter, we focused our attention on one of the most recent as well as prominent directions in privacy preserving data mining: the mining of user mobility data. As a result, a whole new body of research was introduced to allow for the mining of data, while at the same time prohibiting the leakage of any private and sensitive information. The authors focus their attention on very recently investigated methodologies for the offering of privacy during the mining of user mobility data. In the end of the chapter, they provide a roadmap along with potential future research directions both with respect to the field of privacy-aware mobility data mining and to privacy preserving data mining at large.

10. Future Work

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 come into existence, as well as new application areas and 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, as well as privacy-aware stream data mining are among the most recent and prominent directions of privacy preserving data mining. in these applications privacy is a major concern and thus novel privacy preserving methodologies will have to be proposed to protect those patterns that are sensitive with respect to the privacy of individuals. In what follows, we briefly present some future research directions both with respect to the field of privacy-aware mobility data mining and to privacy preserving data mining at large.

References

[1] Abul, O., Atzori, M., Bonchi, F., & Giannotti, F. (2007a). Hiding Sequences. In *IEEE International Conference on Data Engineering Workshop* (pp. 147-156), Istanbul, Turkey.

[2] Abul, O., Atzori, M., Bonchi, F., & Giannotti, F. (2007b). Hiding Sensitive Trajectory Patterns. In *IEEE International Conference on Data Mining Workshops* (pp. 693-698), Omaha, NE.

[3] Abul, O., Bonchi, F., & Nanni, M. (2008). Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. In *International Conference on Data Engineering* (pp. 376-385), Cancun, Mexico.

[4] Aggarwal, C. C., & Yu, P. S. (Eds.). (2008). *Privacy Preserving Data Mining: Models and Algorithms*. New York: Springer-Verlag.

[5] Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In *ACM SIGMOD International Conference on Management of Data* (pp. 207-216), Washington, DC.

[6] Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In *International Conference on Very Large Data Bases* (pp.487-499), San Francisco, CA.

[7] Agrawal, R., & Srikant, R. (2000). Privacy Preserving Data Mining. In *ACM SIGMOD Conference on Management of Data,* (pp. 439-450), Dallas, TX.

[8] Atallah, M., Bertino, E., Elmagarmid, A. K., Ibrahim, M., & Verykios, V. S. (1999). Disclosure Limitation of Sensitive Rules. In *IEEE Workshop on Knowledge and Data Engineering Exchange* (pp. 45-52), Chicago, IL.

[9] Brand, R. (2002). Microdata Protection Through Noise Addition. In *Inference Control in Statistical Databases* [). New York: Springer-Verlag.]. *Theory into Practice*, *2316*, 97–116.

[10] Cao, H., Mamoulis, N., & Cheung, D. W. (2006). Discovery of Collocation Episodes in Spatiotemporal Data. In *International Conference on Data Mining* (pp. 823-827), Hong Kong, China.

[11] Chang, L., & Moskowitz, I. S. (1998). Parsimonious Downgrading and Decision Trees Applied to the Inference Problem. In *Workshop on New Security Paradigms* (pp. 82-89), Charlottesville, VA.

[12] Chen, K., & Liu, L. (2005). Privacy Preserving Data Classification with Rotation Perturbation. In *IEEE International Conference on Data Mining* (pp. 589-592), Houston, TX.

[13] Clifton, C. (2000). Using Sample Size to Limit Exposure to Data Mining. *International Journal of Computer Security*, *8*(4), 281–307.

[14] Clifton, C., Kantarcioglou, M., Lin, X., & Zhu, M. (2002). Tools for Privacy Preserving Distributed Data Mining. *ACM SIGKDD Explorations*, *4*(2), 28–34. doi:10.1145/772862.772867

[15] Dasseni, E., Verykios, V. S., Elmagarmid, A. K., & Bertino, E. (2001). Hiding Association Rules by Using Confidence and Support. In *International Workshop on Information Hiding* (pp. 369-383), Pittsburgh, PA.

[16] Defays, D., & Nanopoulos, P. (1993). Panels of Enterprises and Confidentiality: The Small Aggregates Method. In *Symposium on Design and Analysis of Longitudinal Surveys* (pp. 195-204). Ottawa, Canada: Statistics Canada.

[17] Domingo-Ferrer, J., & Mateo-Sanz, J. M. (2002). Practical Data-Oriented Microaggregation for Statistical Disclosure Control. *IEEE Transactions on Knowledge and Data Engineering*, *14*(1), 189–201. doi:10.1109/69.979982

[18] Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, Continuous and Heterogeneous K-anonymity Through Microaggregation. *Data Mining and Knowledge Discovery*, *11*(2), 195–212. doi:10.1007/s10618-005-0007-5

[19] Giannotti, F., Nanni, M., & Pedreschi, D. (2006). Efficient Mining of Temporally Annotated Sequences. In *SIAM International Conference on Data Mining*, Bethesda, MD, (pp. 346-357).

Authors' Biography

Abu Sarwar Zamani is working as a Senior Lecturer in Shaqra University, Kingdom of Saudi Arabia. He is doing PhD form Pacific University, Udaipur, India. He has received his Master of Science in Computer Science from Jamia Hamdard (Hamdard University), New Delhi, India in the year of 2007, and later he did Master of Philosophy in Computer Science from Vinayak Mission University, Chennai, India in 2009.He is a Member of International Association of Computer Science and Information Technology- IACSIT, (Singapore, Member No:80341225), Member of International Journal of Computer Science & Emerging Technology IJCSET (UK), Program Committee Member of Academy & Industry Research Collaboration Center (AIRCC), Member of International Association of Engineers-IAE (Hong Kong, Member No:113797) and Member of IEEE. He has actively attended and published various research papers in National as well as International.

Md. Mobin Akhtar has received his Master of Science Technology from Jamia Millia Islamia University, Delhi India in 2008 with distinction marks. He has done his B.Sc. (Hons.) in Mathematics in 2003 from Vinoba Bhave University, India. He is currently working as Lecturer in College of Computing and Information Technology, Shaqra, Shaqra University, Kingdom of Saudi Arabia. His research interests Data Mining, Cloud Computing, and Big Data. Before starting his teaching career, He spent two years at S.G Innovative Noida (India) working on the XSLT, Regulation Expression and Qube Activity tools. His teaching abilities include innovative skills and extensive use of technology in teaching.

Danish Ahamad has received his Master of Computer Application (MCA) from Shobhit University, Meerut, India in 2006 with distinction marks. He has done his B.Sc.2003 from Chaudhary Charan Singh University, Meerut, India. He is currently working as Lecturer in College of Science and Arts, Sajir, Shaqra University, Kingdom of Saudi Arabia. His research interests Database Security, Cloud Computing, and Network Security. Before starting his teaching career, He spent four years at MIS & Analytics Training Institute (MAATI) as a Teacher. His teaching abilities include innovative skills and extensive use of technology in teaching.