Privacy Preserving Data Mining

Andrei Karpau
Institute of Technology Blanchardstown
Dublin 16

January 4, 2017

Abstract

Currently more and more data is analysed by different data mining techniques. Thus, the privacy of the data becomes nowadays a very important problem. A lot of research is made in the area of privacy preserving data mining. This paper aims to make an overview of the techniques used on different stages of working with the data: data collection, data publishing and output of data mining results. The paper takes a look into the threads of possible data disclosure and the methods for keeping the privacy of the information.

Keywords: privacy preserving data mining, anonymization, k-anonymity, data perturbation, randomization, hiding association rules

1 Introduction

Nowadays, there is a huge amount of data produced in the world. It comes from different sources and is produced by systems, sensors, and users. This volume of data growth exponentially, so more and more data is produced, aggregated and stored by different systems (Moniruzzaman and Hossain, 2013, p. 2). Then, it is often used for different business needs, such as market analysis, customer relationship management, market segmentation, medical diagnosis or even national secure. So, different data mining techniques can be applied for extraction of interesting knowledge and patterns from this data (Mining et al., 2006, pp. 4-5). In order to deal with the vast amount of data, the massive collection of it is done by automatic tools and then it is saved to the companies databases. Therefore, the private information can be often unintentionally stored and then used in further data mining process.

The privacy problem becomes increasingly important in data mining during the last years. More and more data is collected and analysed by different companies for their business needs. From one side, they need to deal with the data that is fully representative and provide the correct results after analysing it. From the other side, they must ensure that the privacy of personal and sensitive business information is not violated. In order to avoid violations, various Privacy

Preserving Data Mining (PPDM) techniques should be used (Ravi and Chitra, 2015, p. 616).
Generally, there are two steps when PPDM methods can be applied (Aggarwal, 2015, pp. 663-664): data collection and publication and output privacy of data mining algorithms. On data collection phase the special software can be used for modifying the input data before storing it. It should be used in cases, when the collector is not allowed to get the private information about the participants. In this approach saved data is anonymous and can be safely published to other organizations. Despite this, it is often the case that the collector of the data is a trusted organization and is allowed to keep the private data about its contributors. A good example can be an insurance company that keeps the information about its clients and insurance events. If it needs to publish this data to another organization for analysis, it must ensure that the private information about its contributors is not violated. So, PPDM techniques can be used during the publish stage for achieving it. However, sometimes the trusted company can work with the private data on collection and analysis stages, so it has no need to publish any data to third parties, except the data mining results. Even though, no private information is available to untrusted organizations, the results provided by data mining algorithms can also violate the data privacy. For example, it can be the associations patters, that reference the small groups of people. This is another case, where privacy preserving data mining algorithms can be used.
Overall, PPDM methods make some modifications of data or data mining techniques in order to perform privacy preservation. Often, this reduces the granularity of the data, that leads to loss of the performance of data mining techniques (Ravi and Chitra, 2015, p. 616). So, the main aim of the PPDM methods is to preserve the privacy, while keeping the high level of granularity in the data and effectiveness of data mining algorithms (Aggarwal, 2015, p. 664).

1.1 Contribution and Structure

This paper addresses the privacy preserving data mining and makes the overview of it. It considers the threads of privacy disclosure on different stages of data mining process and describes the methods for dealing with it. This paper is structured as follows: Section 1 provides the introduction and the overview of privacy preserving in data mining; Section 2 describes the privacy preserving on data collection stage; Section 3 takes a look into the privacy during data publishing; Section 4 makes the overview of privacy preserving on the output of data mining algorithms; and the last Section 5 makes the final conclusions.

2 Privacy On Data Collection Stage

Nowadays, massive amounts of data are coming from different sources. Depending on business needs, this data is often collected and stored in some centralized storages. For example, in business intelligence scenarios the collected data is

usually saved into data warehouses for further analysis. However, if the collector of the data is not trusted, then the privacy preserving techniques should be applied on data collection stage. Often, the same techniques can be used also on data publishing stage, when the information collected by company should be sent to some external organizations.

The obvious solution of these problems is removing of all the information that can identify the person. Nevertheless, it is usually difficult to determine, which information should be really removed. Moreover, deleting of such data can decrease the effectiveness of further data mining (Vaidya and Clifton, 2004, p. 20). The main problem in de-identification is that often even non personal attributes can provide the valuable information in combination with another attributes or even public data. For example, Sweeney (2001, p. 21) could precisely re-identify some patients only by publicly available information.

Another possible solution is to avoid building the centralized storage of the data. Especially, it can be used when the data is stored in many different distributed storages (Vaidya and Clifton, 2004, p. 20). So, data mining algorithms can work in a distributed manner and as a result avoid the problem of collecting the data. Example of distributed techniques are proposed by Lindell et al. (2002, p. 177) or Sakuma & Arai (2010, p. 935). Both algorithms use secure multi-party computation (SMC) protocols that can be computationally inefficient on large datasets (Vaidya and Clifton, 2004, p. 25). A distributed technique that do not need to use cryptography protocols or randomization is provided by Yan et al. (2013, pp. 1-12). They suggest the method that works in distributed autonomous online-learning scenarios and does not require any modifications of original data-mining algorithms.

Another technique that is worth to note in data collection scenarios is random data perturbation. The idea behind it is that the original data can be somehow modified, so it does not represent the real, privacy-sensitive values. The main difficulty in this technique is to obtain the correct results after applying data mining algorithms on the modified data. One popular example of perturbation is data swapping. It was first suggested by Reiss (1980, p. 38) in 1980s. The data swapping is based on the idea that the values are exchanged between the rows, while keeping the statistics of the dataset relevant. The further research has shown the it can be effective for privacy preserving scenarios and is computationally effective (Moore, 1996, pp. 1, 25).

Randomization is another popular data perturbation approach. The main idea of the method is adding the additional noise to original dataset (Aggarwal and Philip, 2008, p. 12). The noise should be pretty large, so it is impossible to recover the private information from the data set. However, it is still important that the data set stays representative for data mining algorithms. Thus, the randomization method mainly has 3 important steps behind it: at first the noise is added to the original data and the modified dataset is released together with the distributed function used for producing the noise; then the original distribution is reconstructed using the modified dataset and noise distribution; finally, data mining techniques are applied to the reconstructed distributions (Aggarwal, 2015, p. 665). One of the drawbacks of this approach is that not

all data mining algorithms are capable to work with probability distributions. Some of the methods require to have the original data, so they cannot be used together with randomization. Moreover, it was shown that it is sometimes possible to derive the private information from the randomized dataset (Huang et al., 2005, p. 37).

3 Privacy On Data Publishing Stage

In the publishing scenario the collected data is stored in the trusted storage. However, it is needed to make it available to third party organization for further data mining. The private information inside the data set should be preserved, while not decreasing the analytical capabilities of the dataset. This business case is more or less similar to the data collecting problem. Some techniques such as randomization can be used on both stages. There are some other approaches that help to ensure the privacy during data publishing. One of the oldest is k-anonymization. It states that there are at least k data entities with the same data on the specific attribute (Willemsen, 2016, p. 2). So, if the attribute has k-anonymity it is not possible to distinguish the record by this attribute from at least k-1 other records. These model addresses the problem of re-identifying individuals by combinations of non-private attributes. For example, even if the private data is removed from the data set, the combination of age and Zip-code can sometimes precisely point to only one person. K-anonymity ensures that there are at least k different persons that are pointed by each value of attribute and highly decreases the possibility of using the combinations of the attributes for gathering the private information. For reaching the k-anonymity property usually two techniques are used: suppression and generalization (Aggarwal and Philip, 2008, p. 20). In suppression the value of attribute is removed completely. It is often used if the value references only one person in the dataset. In case of generalization the values in the column are exchanged by the ranges, so the granularity of the data is decreased. These methods allow to ensure the defined level of privacy in the data, while decreasing the background information available for data mining algorithms.
Even though, the k-anonymity is a useful technique for privacy preserving in the data set, it still has some vulnerabilities. Among them are Homogeneity Attack and the Background Knowledge Attack (Machanavajjhala et al., 2007, pp. 3-4). These attacks are based on the analyzation of the combinations of the attributes that can form the distinct groups for valuable attributes and on some overall knowledge related to them. L-diversity is a model that helps to overcome these vulnerabilities of the k-anonymity method. It ensures that each group in the dataset contains l different values.

4 Privacy On Data Mining Output

The last stage of the data analysis process is applying the data mining algorithms on the data set. The output of such algorithms can also disclose the private information, which exists in the training data. One of the possible solutions of this problem is the use of privacy preserving methods on the initial data. These methods can be applied on data collection or data publication stage and, as a result, provide the data set with hidden or removed private information. So, the final analysis does not disclose any private information, when working with such data.

However, many privacy preserving methods, that are applied on data collection or data publishing stages, reduce the informativeness of the data and can deteriorate the effectiveness of final data mining. So, when working with original data, the important private information can often be disclosed by association rules provided by data mining methods.

There are many different techniques for hiding private information in association rule. Generally, they can be divided into two groups (Aggarwal and Philip, 2008, p. 33): distortion and blocking. Distortion based technique somehow modifies the values of the given transaction (Verykios, 2013, pp. 28-29). For example, it can be the modification of values from 1 to 0 and back for binary attributes. The blocking approaches use hidden symbol to show the absence of a specific value (Verykios, 2013, pp. 28-29).

Another area of data mining where the final model can disclose private information is classifications. Especially, it can be a case for algorithms, where association rules are used as a subroutine. In such a cases the same techniques, which are used for hiding the association rules can be used. In other approaches data is modified such a way that the accuracy of the model provided by data mining algorithms is slightly reduced, while keeping the higher level of privacy. The measure of parsimonious downgrading is introduced for determining if the loss of the functionality of the model is worth the extra confidentiality (Chang and Moskowitz, 1998, p. 82). There are many techniques proposed for different classifiers. For example, the recent research provides the methodology for hiding rules in decision tree model (Kalles et al., 2016, pp. 1-8).

5 Conclusion

In recent years, a lot of research in the area of PPDM has been done. Generally, it addresses different techniques of keeping the confidence of the information on different stages of data mining process depending on business needs. This paper makes an overview of privacy problems during data collection, data publishing and output of final results. It investigates the threads of possible privacy disclosures and the techniques for dealing with it. Among them are data perturbation, randomization, data suppression and data generalization. For association rules hiding paper describes distortion and blocking methods. The paper also addresses the k-anonymization and l-diversity models that are used for ensuring

the level of privacy in the data set.
Generally, there are a lot of different methodologies for preserving privacy in the dataset. Many of them can provide a sufficient level of privacy of the information. However, they often lead to reduce of informativeness of the dataset or to the decrease of the accuracy of data mining models. Moreover, some of the techniques can be computationally expensive. So, it is often important to determine how much the data preserving methods influence the final data mining results.

References

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

Aggarwal, C. C. and Philip, S. Y. (2008). A general survey of privacy-preserving data mining models and algorithms. In *Privacy-preserving data mining*, pages 11–52. Springer.

Chang, L. and Moskowitz, I. S. (1998). Parsimonious downgrading and decision trees applied to the inference problem. In *Proceedings of the 1998 workshop on New security paradigms*, pages 82–89. ACM.

Huang, Z., Du, W., and Chen, B. (2005). Deriving private information from randomized data. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 37–48. ACM.

Kalles, D., Verykios, V. S., Feretzakis, G., and Papagelis, A. (2016). Data set operations to hide decision tree rules. In *Proceedings of the 1st International Workshop on AI for Privacy and Security*. ACM.

Lindell, Y. and Pinkas, B. (2002). Privacy preserving data mining. *Journal of cryptology*, 15(3):177–206.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3.

Mining, D., Han, I., and Kamber, M. (2006). Data mining concepts and techniques. *Morgan Kaufinann*.

Moniruzzaman, A. and Hossain, S. A. (2013). Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *arXiv preprint arXiv:1307.0191*.

Moore, R. (1996). Controlled data swapping techniques for masking public use microdata sets, us bureau of the census. *Unpublished manuscript*.

Ravi, A. and Chitra, S. (2015). Privacy preserving data mining. *Research Journal of Applied Sciences, Engineering and Technology*, 9(8):616–621.

Reiss, S. P. (1980). Practical data-swapping: The first steps. In *Practical Data-Swapping: The First Steps*, pages 38–38. IEEE.

Sakuma, J. and Arai, H. (2010). Online prediction with privacy. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 935–942.

Sweeney, L. (2001). *Computational Disclosure Control A Primer on Data Privacy Protection*. PhD thesis, Massachusetts Institute of Technology.

Vaidya, J. and Clifton, C. (2004). Privacy-preserving data mining: Why, how, and when. *IEEE Security & Privacy*, 2(6):19–27.

Verykios, V. S. (2013). Association rule hiding methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):28–36.

Willemsen, M. (2016). Anonymizing unstructured data to prevent privacy leaks during data mining.

Yan, F., Sundaram, S., Vishwanathan, S., and Qi, Y. (2013). Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2483–2493.