DOI : 10.15613/sijrs/2018/v5i1/188667
ISSN (Print): 2349-8919
ScieXplore: International Journal of Research in Science, Vol 5(1), 31–34, January–June 2018 ISSN (Online): 2350-0999

An Empirical Study on Preserving Sensitive Knowledge in Data Mining

S. Dhanalakshmi*

Department of Software Systems, Sri Krishna Arts and Science College, Coimbatore – 641008, Tamil Nadu, India; lakshmimkv@gmail.com

Abstract

Protection of data in data mining involves preserving sensitive data and sensitive knowledge. Preserving sensitive data focus on protecting the data when it is shared for mining with third parties. In some situations even the result of data mining techniques might reveal sensitive information. This area of research which focuses on output preservation of data mining techniques is termed as Preserving Sensitive Knowledge. The sensitive results of data mining are preserved by reducing the efficiency of the result. In this paper, an empirical study on existing sensitive knowledge preservation approaches is discussed. The approaches include rule hiding in association rule mining and classification technique. The output data protection also includes the restriction in query processing. The article highlights the merits and demerits on each approach.

Keywords: Association Rules, Privacy Preservation, Rule Hiding, Sensitive Knowledge

1. Introduction

Large volume of data is available in information industry. The data will be useful only if it is transformed into useful information. Data mining techniques like clustering, classification, regression and association rules are often used to analyse large amount of data and used to retrieve useful information. The retrieved information can be applied in many areas like medical, market analysis, fraud detection etc.

Collaborative data mining analyse huge volume of data from different parties to retrieve useful information for decision-making. For example, organizations are ready to collaborate and share their data to retrieve useful information which can be mutually used for their business decision- making. The organizations are ready for collaborative data mining, but they don't want their business strategies to be revealed during the mining process.

The privacy preserving data mining techniques focus on preserving the sensitive data and sensitive knowledge. The results of both sensitive data preservation and sensitive knowledge preservation techniques should ensure that the data mining results generated after applying the preservation techniques should give valid outputs when it is evaluated. It should also assure that the privacy is also maintained.

2.Preservation Techniques

2.1 Sensitive Input Data

The aim of sensitive data protection methods is to preserve the sensitive data before it is used in mining. This type of preservation falls in to three main categories – randomization, encryption based and anonymization based techniques[1]. Randomization methods mask the sensitive data value by adding noise. The randomization based privacy preservation uses additive[2], multiplicative[3] and random response techniques[4] to mask the sensitive input data. All these techniques use some form of transformation to the original data. The main focus of preservation techniques is to ensure privacy and also should yield valid data mining results.

*Author for correspondence

2.2 Sensitive Knowledge

Sensitive knowledge preservation methods achieve protection of sensitive data mining results by degrading the extracting knowledge. Preserving the sensitive knowledge is equally important to that of preserving the sensitive data because the sensitive output of data mining techniques can also breach the privacy.

Many approaches are devised in the research area of privacy preserving data mining to protect the sensitive data and sensitive knowledge. This article focuses on the approaches used in preserving sensitive knowledge.

3. Sensitive Knowledge Preservation Approaches

Sensitive knowledge can be extracted using data mining techniques such as:

- Association Rules.
- Classification Models.
- Clusters.

Lot of research work is carried on the sensitive knowledge preservation using the association rule mining and classification model prediction. The aggregate pattern generated using these techniques can infer sensitive information.

Output data privacy also involves preserving privacy during query processing. Privacy preserved query processing goal is to execute queries over multiple sources of data without extracting any extra information from any data sources.

Therefore, existing approaches used for preserving the sensitive data mining end results include Rule Hiding and Query Processing which is shown in Figure 1.

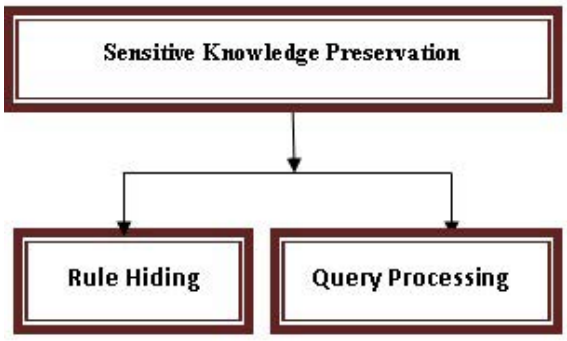

Figure 1. Approaches used for sensitive knowledge preservation.

3.1 Rule Hiding

Rule mining is used to generate patterns for large populations of data. Data mining techniques Association rules and Classification rules are used for generating patterns.

Privacy preservation in rule mining highlights that inference rules can be used to infer sensitive knowledge of individuals. The research work carried out in rule hiding mainly focus on Association and classification rule hiding as shown in Figure 2.

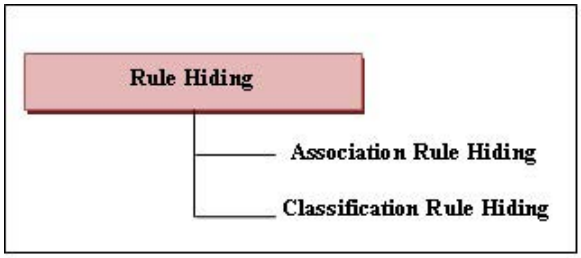

Figure 2. Rule hiding.

3.1.1 Association Rule Hiding

Association rule data mining technique is used to generate association rules from huge datasets. The extracted association rules are used by the organization to improvise their business growth. If the generated frequent item set or generated rules contains sensitive knowledge, it has to be hidden before the data is shared or published.

The privacy preservation techniques use approaches to restrict the frequent item set or association rule generated from unauthorized access[5].

The process in association hiding restricts the disclosure of sensitive knowledge by reducing the support and confidence of item sets.

If the frequent rule generated contains strong support and confidence more than the sensitive threshold level, the rule hiding process should be applied.

The association rule hiding process includes pattern generation specification for finding sensitive rules, data sanitization and measuring the side effects of sanitization[6].

In the pattern generation, frequent item sets are extracted from original dataset using association rule mining algorithms. The specification step includes in predicting the sensitive rules. If the frequent item set contain sensitive rules the data sanitization is applied to the dataset to hide the generated sensitive pattern. Then the data utility and privacy level is measured by applying the association rule mining techniques to distorted dataset[6].

The research work in association rule hiding mainly focuses on two approaches. The first approach finds the sensitive frequent items in the rule, then attempts to change

the transaction by transaction until the confidence and support of the rule is reduced below the minimum support of confidence level. This is achieved either by adding new items to transactions or by deleting items from transaction. Another approach of association rule hiding uses cryptography based method which discloses only the date mining results. It hides the sensitive patterns by sanitizing or restricting patterns[7].

3.1.2 Classification Rule Hiding

The classification rule hiding preservation approaches inference the disclosure of sensitive classification rules. The classification rules are generated for large population of datasets. The generated rules are presented to the data owner; if the discovered classification rules are found to be sensitive it is hidden. The pattern is reconstructed with new dataset which contains only classification rules that are non sensitive[8].

The main advantage of rule hiding approaches is it preserves the sensitive rules. Disadvantage of rule hiding is effectiveness of the knowledge extracted is downgraded to certain level. The possible attack to retrieve the sensitive rule is background knowledge attack.

3.2 Query Processing – Preservation

The query results disclosure control can be categorized into query disclosure and query auditing approaches as shown in Figure 3. The query result disclosure control preserve end result either by denying the output or by modifying the results.

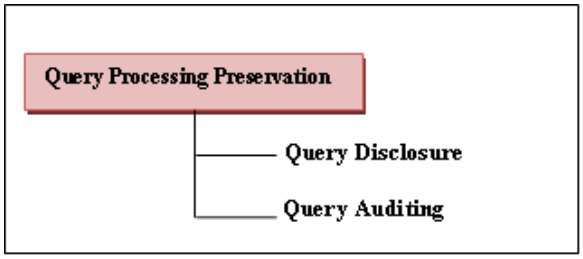

Figure 3. Query result disclosure control.

3.2.1 Query Disclosure

Query disclosure methods preserve the sensitive aggregate queries results either by restricting the output of the queries generated or by sanitizing the generated output of the queries.

The approach used to restrict the sensitive query results first calculate the query output, if the query reveals sensitive knowledge the noisy version is returned to the user to preserve the query results.

The approach sanitizes the result by adding noise according to the sensitivity of the query function[9].

3.2.2 Query Auditing

The focus of query auditing method is to evaluate the queries and to check whether the results of query will reveal sensitive information from the data source. Such sensitive query which violates the privacy is denied. The query auditing can be done either as online or offline auditing. In online auditing the process is to check whether to answer the query or deny if it leaks privacy. In offline auditing the queries are posed to the dataset the auditor has to check whether to answer or ignore the queries. Various approached has been followed in query auditing to reply or deny the query results[10].

The main advantage of query disclosure approaches is it preserves the sensitive knowledge. Disadvantage of query disclosure or query auditing is too much denials to the queries lead to less utility of the database. The denials to the queries can also have the possibility for leakage of information.

4. Conclusion

The research focus in privacy preserving data mining can be categorized in to input data preservation and output knowledge preservation. In this paper a survey on output data preservation techniques is highlighted. Sensitive knowledge preservation approaches in rule hiding using association rule mining and classification is addressed. The process of rule hiding approaches is addressed. The article also focuses on query results disclosure control techniques which restrict the leakage of sensitive information through query results. The paper is concluded with the advantages and disadvantages of each approaches used in sensitive knowledge preservation. The possible attacks on the rule hiding and query disclosure control are also addressed.

5. References

1. Aggarwal CC, Yu PS. A general survey of privacy-preserving data mining models and algorithms. Privacy-preserving Data Mining; 2008. p. 11–52. https://doi.org/10.1007/978-0-387-70992-5_2
2. Agrawal D, Aggarwal CC. On the design and quantification of privacy preserving data mining algorithms. Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems; 2001. p. 247–55. https://doi.org/10.1145/375551.375602

3. Chen K, Liu L. A survey of multiplicative perturbation for privacy-preserving data mining. Privacy-Preserving Data Mining. 2008. p. 157–81. https://doi.org/10.1007/978-0-387-70992-5_7
4. Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association. 1965; 60(309):63–9. https://doi.org/10.1080/01621459.1965.10480775
5. Refaat M, Aboelseoud H, Shafee K, Badr M. Privacy preserving association rule hiding techniques: Current research challenges. International Journal of Computer Applications. 2016; 136(6):11–7. https://doi.org/10.5120/ijca2016908446
6. Telikani A, Shahbahrami A. Data sanitization in association rule mining: An analytical review. Expert Systems with Applications. 2018; 96:406–26. https://doi.org/10.1016/j.eswa.2017.10.048
7. Garg V, Singh A, Singh D. A survey of association rule hiding algorithms. 2014 Fourth Intenational Conference on Communication Systems and Network Technologies, IEEE; 2014. p. 404–7. PMid: 25657953 PMCid: PMC4311352. https://doi.org/10.1109/CSNT.2014.86
8. Natwichai J, Li X, Orlowska M. Hiding classification rules for data sharing with privacy preservation. International Conference on Data Warehousing and Knowledge Discovery; Springer; Berlin, Heidelberg. 2005 Aug. p. 468–77). https://doi.org/10.1007/11546849_46
9. Dwork C, Kenthapadi K, Mcsherry F, Mironov I, Naor M. Our data, ourselves: Privacy via distributed noise generation. Advances in Cryptology-EUROCRYPT; 2006. p. 486–503. https://doi.org/10.1007/11761679_29
10. Thong T, Buttyan L. Query auditing for protecting max/min values of sensitive attributes in statistical databases. Trust, Privacy and Security in Digital Business. 2012. p. 192–206. https://doi.org/10.1007/978-3-642-32287-7_17