DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4512

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

DEMONSTRATE DATA MINING PRIVACY AND EXAMINING FUTURE CHALLENGES

Ali Omidfar and Mohammad Pur Mahmood
Mazandaran University of science and technology,
Mazandaran, Iran

Abstract: Nowadays data mining discussion and available personal data has mod privacy a topic in the scientific community the methods of maintaining a privacy to the traditional way cannot support this data well and it can provide the urgent need to the maintaining data privacy. Consequently when they protect important information,the knowledge of data prevents from accessing them. In this article new research on data mining,privacy, some challenges (controversial terms),the existence and method of distortion of data to obtain association rules, privacy and privacy in data distribution are discussed!The purpose of this work is actually to provide criteria for assessing the details of privacy algorithms,includes algorithm performance,data usage,privacy and data mining problems. At the end of the discussion the development of data mining and privacy has been outlined for other areas.

Keywords: Protecting privacy, disruption of data, data distribution,data mining

1. INTRODUCTION

With the advancement of computer capabilities and data storage,new data mining algorithms have been proposed. New days there are so many information from organizations. Traditional privacy methods cannot support this information well. And it can provide the urgent need to the maintaining data privacy. Consequently when they protect information, the knowledge of data prevents from accessing them. Data mining privacy mainly consists of two aspects. First of all,how to ensure that info such as a card number,name,address and other info of a person is not disclosed in the process of using data [1].
Should the original information be reviewed or deleted from database? The purpose of this work is to prevent the privacy from receiving undesirable data. Another way how to make data usage more optimal?Exploring the sensitive applicable information should be eliminated. The basic purpose of data mining is to protect privacy,major data reformation (large) is based on several methods and development of data mining algorithms. At present privacy technology in the usage of data bases is mainly focused on data mining and data visualization. At the moment table 1 illustrates the topics that related to the research path of privacy [1].
Research topics privacy is determined by scientific applications of privacy requirementsPublic privacy methods at a low conservation level have the task of protecting data through the introduction of statistical models and possible models. Privacy in data mining is mainly used to archive privacy by specifying different data in high-level data. The relapse of data is based on providing a privacy-based approach to privacy. So the design of the privacy algorithms is a requirement [1].
Research on privacy methods focuses on data distortion, data encryption and etc. Such as classification extraction algorithms,privacy, association rule extraction,privacy, distributed privacy, and data dissemination have been developed many algorithms based on encryption methods. Like associative extracted rules in the horizontal and vertical data of the categories,clustering,decision trees and articles that work on the privacy of the flow of data are high. Agarwal etal [2] developed the K-anonymity algorithms on dissemination of data flowin privacy. This article reviews privacy algorithms.
In second section of this article,the research methods of data mining privacy algorithms are briefly reviewed. Privacy technologies are discussed in section 3 and finally the conclusion is discussed.

Table 1: Privacy research path

Available methods	**Research path**
Random swap disruption	Public privacy protection technology
Explore association rules, clustering	Data mining privacy protection technology
	Privacy protection data distribute principle

2. PRINCIPAL ALGORITHMS FOR RESEARCH OF DATA MINING PRIVACY

There are many data mining methods for privacy. Privacy classification methods based on the following aspects: Data dissemination,Data distortion,Data mining algorithms. Hidden data rules and protecting privacy,later we give a brief explanations of each of them.
Data distributed: currently, some algorithms run data mining privacy on centralized data and some others on distributed data. Distributed data includes information that is vertically divided. The records of different data bases in different sites are vertically divided. In data that is horizontally divided,each record from the databases distributes values at different sites [3].
Data distortion: this method is intended to reform original database record, before it is distributed to archive the privacy goal. The data distortion methods include deviations, obstruction or merging,exchange and sampling.

All of these methods are created by reforming the amount of one adjective or particle reforming by the amount of one adjective.

Data mining algorithms: privacy data mining algorithms includes classification,association rules, clustering and Bayesian networks [4].

Hidden data rules: This method refers to hiding the rules of the original data for the hidden rules of the original data are complex. Some people provided exploratory methods to solve this problem [3].

Privacy: In order to protect privacy we need to reform the data carefully to archive the desired data. We do this for a number of reasons: First,the information that reform based on the method of exploration and adaptation and only reforming the selected quantities (not all quantities) to minimize the data cost. Second encryption technology such as multi-part calculation is assured. In a few sections, ifany site only recognizes its input and doesn't know anything else then the calculations are assured. The third one is the data reconstruction method, in this way it is possible to reconstruct the distribution of the original data from random data [3].

3. PRIVACY TECHNOLOGY

3.1. Incorrect data techniques:

In order to protect privacy in distributed data bases, researchers have proposed effective data mining algorithms to hide original information. The purpose of privacy is as follows.

1. Hiding sensitive information that exist in the original data.
2. The original and hidden data have the same characteristics.
3. Get the accuracy of the same data as the original dataset.

Privacy data mining algorithms such as classification,discovering association rules,clustering, choosing the data needed to reform or deterge and deterge selection is a NP-hard problems. To solve these complex problems, distortion methods such as random deviation, obstruction and compression are used.

Extract association rules based on deviation.

Many statistical data have been used to judge unexpected rules in the dataset and used support and trust as standard. All association rules that are larger or equal to the definition of assurance and support for users but from the user's point of view. Some rules are more important but some others not. Associated hiding rules techniques use the following methods to detergent the original dataset.

1. All important rules can only be used for the original data mining. Also at the same time assurance and support will not be considered until the original dataset is deterged.
2. Maybe important rules cannot be extracted from the original dataset and they also cannot be extracted from the dataset in the same assurance and support.

The optimal detergent problems used to extract association rules for hiding a large item set is in up-hardID issue [6].Reference[7] we propose a detergent set for refining the original Rules. The methods used in this work are eitherto preventimportant rules generated by hiding the duplicate item set that they are derived from or to reduce the assurance of important rules by bringing it below the specified user's scope. These two methods lead to the production of three strategies, to be proposed that the probability of both quantities in the binary data bases was to correct the quantity of zero and convert it to one this flexibility in data modification has a side effect that apart from non-important association rules that are hiding,would be a non-repetitive rule.

Extract association rules using obstruction:

Another discussion about association rules has obstructed data information. In the obstructing method, the quantity of the adjective replaces each of the data with a question mark. That uses unknown quantities instead of real one. In medicine, too, incorrect quantities are used instead of the real quantities that are so common [6].It suggests a method of extracting associative rules by using reference [1], obstruction in which the minimum support is appropriately modified. And with a minimum of support, the assurance of the underlying rules is no less disrupted than the scope of assurance [7].Whether the quantities of zero and one should be converted to the question mark as the original quantity. It is a complete reference to the obstructing method, which is discussed in detail. This method of reconstructing text uses of the disruption.

Extraction of classification rules based on obstructionReference[6], provide a new case for combining the analysis of classification rules and lowering the value of saving. In this case , the data manager wants to block the values based on the class by doing this the data receiver will not be able to build a model containing useful informationfor law value data. Reducing value-effectiveness is a case for formalizing the phenomenon of data deter Genting from a data set in reducing data value. In reducing the value of money, the amount of set-up cost for this information is actually the potential that isn't sent from this amount. The main purpose for doing this is to find out if the loss of performance with undermining the value of the data deserves more or no additional trust.

3.2. Extracting distributed privacy:

In the data mining privacy, some researchers have proposed a large number of encryption approaches to solve the problem with the features described below. Two more parties extract their data on a cooperative basis,but none of them is willing to reveal their data. This is an assured multi-part collection(SMC). A problem is a distributed environment that focuses on how to transform different data mining methods into assured multi-part calculation problem such as classification of data, data clustering, extracted generalized association rules. Assured multi-part calculation methods include an assured set, assured set units, and assured set size of intersection and numerical multiplication.

Extracting the association rules of partitioned data vertically

The partitioned dataset vertically provides different features for each item on different sites. Extracting associative rules, the privacy of partitioned data is obtained vertically by finding the specific support number of an item set. If the support for an item set is calculated securely, then we can check that if supported is more than threshold. Are item sets duplicated or not? Each part involved in the calculation is a benefit to a subset of the items that is the key to calculating vector multiplication. Therefore multiplication of a point can be an assured calculation, so support can also be calculated with high confidence.

Extracting the association rules of partitioned data horizontally

Transactions are partitioned horizontally between several sites in a distributed database. The total count of total support for an item set is the sum of all local support count the X item set is vertically local support.

3.3. Reconstruct technology:

Most of the data mining methods proposed for privacy disruption or reconstruct data in the data convergence layer.Reference [9], building a classified decision tree discusses data disruption as educational data using the amount of individual records. Since the original amount of individual records cannot be accurately estimated. It is possible to consider the exact distribution of the man distribution in order to restore the main distributions of Bayes methods.

Reference[11], Improvement of the Bayes reconstruct by using the EM algorithm in distributed data.Also with more clarity, the author proves that the EM algorithm performs the most estimate as the main data on the distribution of the disruption. But it also proves that when a large amount of data is obtained, the EM algorithm can estimate the main distribution with power. Reference[10], it also shows that when the background is detected by data extraction through reconstruction, privacy estimates will decrease.

3.4. Privacy anonymity:

The anonymous publication method is used to distribute raw data. In order to achieve privacy important data isn't distributed or distributed more carefully. Recent studies focus on the technical anonymity data, it means the limitation of the risk of disclosure of privacy and data usage,which is the distribution of a selection of important data and information that may discloseimportant data,but to ensure important data and the risk of disclosing their privacy, they fall into a specific area. Anonymity data focuses on two aspects: one of this principles is to design better anonymity methods so that distributed data can not only protect privacy but also provide practical application other principles are to design more appropriate anonymity algorithms for certain anonymity principles with deep research on anonymity,how to archive the practical application of anonymous data can be another subject.

Samarattiand Slovenia Proposed on K-anonymity algorithm, in which the record in the distributed table cannot detect the other K-1 record [7]. We call the K records that cannot be identified as the term non-important features. Usually a larger value "K" is about a better degree of privacy. But missing information is increasing. In order to establish any limitation for important data, the K-anonymity algorithm is used incompletely. An attacker can use the attack protocol to identify important data or personal relationships [8], this work leads to privacy. The algorithm (a,k) anonymity [6] based on the previous algorithm, shows an improvement that not only ensures that the distribute of the K-anonymity algorithm is satisfactory, but also ensures that each record. The quantity for each adjective value in each equivalently class is not higher than "a" percent-typically, data distribution methods such as t-closeness, diversity l-k anonymity [8] and other anonymous distribution methods use generalization and accuracy techniques. In terms of dataset, if the dangers of the disclosure of all data "D" are freed by the data itself, it is less than the threshold "a", where "a" belongs to the interval {0 and 1}, that is the risk of disclosing a dataset is called "a", such as the distribute of l-diversity statistical data that assures that the disclosure risk of a distributed dataset is less than 1/L. And the dynamic data that distributes the principle of m-invariance ensures that the risk of disclosure the distributed dataset is less than 1/M.

3.5. Intelligent encryption approach to secure large data storage in cloud calculating.

Implementing cloud calculating provides several paths for web-based services to meet diverse needs. However, data security and privacy are becoming an important issue that limits many of the cloud apps. One of the major concerns about security and privacy is the fact that cloud operators are accessing sensitive information. This concern dramatically increases the anxiety of users and reduces the ability of cloud calculating in many areas, such as the financial industry and government departments. Therefore, an intelligent encryption approach should be provided that cloud service operators cannot directly obtain the details of the data. The proposed approach works in such a way as to separate the file and store the data separately in distributed cloud servers. An alternative method is to determine whether datasets are designed to be designed to shorten the operating time. The proposed scheme is called the efficient storage memory model (SA-EDS), which is mainly supported by algorithms. The algorithm is the distribution data replacement algorithm (AD2), the SED2 (SED2) algorithm and the effective data controller algorithm (EDCon) algorithm. Our tentative evaluations have assessed both performance and security performance, and empirical results show that our approach can effectively protect the underlying threats from clouds and require a reasonable computational time [9].

3.6. Assessment of privacy algorithms

An important aspect of algorithms and data mining applications is privacy and tools for development and evaluation. In order to select the appropriate evaluation criterion. But the reality is that privacy-based data mining algorithms that are under different indexes are better than other algorithms. It's really important for users to set a set of criteria to enable them to provide the most suitable algorithm for privacy and data mining problems.

Performance of the algorithm:

We can see that the complexity time algorithm is more suited than algorithms with novel complexity. A duplicate approach to assessing the time requirements will require an average of a number of operations to reduce the frequency of important information to less than the specified threshold. These quantities may not be of an absolute magnitude, but can be considered in order to perform a quick comparison between different algorithms.

Data usage:

Using privacy data is a very important issue. In order to hide important information, incorrect information must be entered into the database or obstruct the data. Although this technique doesn't change the sample data stored in the database,but this information is incomplete, they still reduce the use of data. The more database changes, the lower use of databases, through estimates of lost data are related to the specific data mining algorithms.

Degree of privacy:

The privacy policy of the protection of low-value data is up to a certain thresholdbut the hidden information reconstructed by upgrading algorithms can evaluate it. A solution can setof maximum disruption of information from

its point of view, and then achieve the degree of uncertainty by the limits of different cleaning methods we hope that an algorithm can archive the highest degree of uncertainty and is better than all other algorithms.

Severity of different data mining algorithms:

In order to provide a complete estimate of the method of cleaning we must measure the degree of difficulty of the data mining algorithms. That is different from the purifaction method and this can be called the horizontal problem parameter.

Develop aformal case that can test a transaction algorithm against a pre-selected dataset. We can prove the privacy guarantee for virtually every category of transaction algorithm.

4. CONCLUSION

Privacy technology as a growing academic research has had very broad applications in manyareas in recent years.

This article focuses on the review of the privacy technology involved in data mining. At first we introduced the privacy study and research methodology, and then introduced privacy methods such as disruption, hiding, privacy and anonymity. Because privacy technology is a combination of progress in several disciplines, there are still plenty of topics to explore. Issue like mobile data mining and data mining related to privacy in data mining, which is a promising path. With the growth of spatial and geographical data, new applications will emerge based on user behavior patterns. Another area of research on this subject is the release of data on privacy, which is an incremental increase in data. Finally, in addition to field-based research a case for estimating and comparing various types of data mining algorithms should be developed. In addition, very important and new point is the application and focus of monitoring large data in the cloud environment which is nowadays much to be considered

REFERENCES

[5] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vision,40(2):99–121, 2000.

[6] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In Proc. 22nd Intnl. Conf.Data Engg (ICDE), 2006.

[7] X. Xiao, and Y. Tao, "M-invariance: towards privacy preserving re-publication of dynamic datasets," Proceedings of the 2007 ACM SIGMOD international conference on Management of data, ACM, Year Published 10.1145/1247480.1247556.

[8] D. Agrawal, and C.C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," Proceedings of the ACM SIGACT Symposium on Principles of Database Systems, ACM New York, NY, USA, 2001, pp. 247-255

[9] Y. Li , K. Gai , L. Qiu , M. Qiu , H. Zhao , Intelligent cryptography approach for secure distributed big data storage in cloud computing, Inf. Sci. (2016) 1–13 .

[10] L. Chang, and I.S. Moskowitz Downgrading and Decision Trees Applied to the Inference Problem," Proceedings of the 1998 workshop on New security paradigms, ACM, 1998, pp. 82-89

[11] R. Agrawal, and R. Srikant, "Privacy mining," Proceedings of the 2000 ACM SIGMOD international conference on Management of data, ACM

[12] D. Agrawal, and C.C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," Proceedings of the ACM SIGACT Symposium on Principles of Database Systems, ACM New York, NY, USA, 2001, pp. 247-255.

[13] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty Fuzziness and Knowledge Based Systems, vol. 10, no. doi: 10.1142/S021848850 2001648.

[14] R. Wong, J. Li, A. Fu, and K. Wang, enhanced k-anonymity model for privacy preserv publishing," Proceedings of the 12 international conference on Knowledge discovery and data mining, ACM, 2006, pp. 754-759.

[15] X. Xiao, and Y. Tao, "M-invariance: towards privacy preserving re-publication of dynamic datasets," Proceedings of the 2007 ACM SIGMOD international conference on Management of data, ACM, Year Published 10.1145/1247480.1247556.

