ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 8, February 2015

A Privacy Preservation Technique Using Machine Learning Technique

Kunwar Singh kushwah, Abhay Panwar

Abstract*— In this work, a new approach for privacy preserving association rule mining is presented. This approach provides excellent accuracy in reconstructing frequent item sets with no influence on support of item sets. At present, the application of this approach is limited to a local environment i.e., used within the organization. This work can be extended further to deal with a distributed environment. Data is distributed among numerous system side environments, so for data mining, mining algorithm is required by which a global consequence is generated which provide knowledge from the distributed database without violation of privacy. To Aim Apply data mining algorithm on distributed data and extract knowledge exclusive of violation of privacy.

***Index Terms*— Privacy Preservation Technique, Machine Learning Technique, Association Rules.**

I. INTRODUCTION

As the growing utilize of data mining, huge volumes of detailed personal data are regularly collected and analyzed. Such data include shopping habits, criminal records, tourist record, environmental record ,medical history, credit records, and among others. On the one hand, such data is an important asset to business organizations and governments both to decision making processes and to provide social benefit, such as medical research, crime reduction, national security, tourism, etc. On the other hand, analyzing such data opens new threats to privacy and autonomy of the individual if not done properly. Thus, an interesting new direction of data mining research, known as privacy preserving data mining has been working on privacy and knowledge discovery. The aim is the extraction of relevant knowledge from large collection of data which is present on distributed form, while protecting private information simultaneously. Association rules are one of the most common tasks in data mining.

In this research, we will propose a novel data perturbation technique on the base of Distributed association rules mining. Machine learning technique has extensive function to determine motivating associations among item set. The approach comprises of two phases. In the first phase, the each local system side compute and supply the corresponding association rules function to a global system side. The second phase which is Global system side where data miner generates global valid association rules function. By using these global association rule function, a mining result is obtain. currently, terrorist organizations have found a cost-effective resource to advance their courses by posting high-impact Web sites on the Internet. Several problems prevent effective and efficient knowledge discovery: the dynamic and hidden character of terrorist Web Sites, information overload, and language barrier problems. Data mining technology has emerged as a means of identifying patterns and trends from large quantities of data. Data mining and data warehousing go hand-in-hand: most tools operate by gathering all data into a central site, then running an algorithm against that data. However, privacy concerns can prevent building a centralized warehouse – data may be distributed among several custodians, none of which are allowed to transfer their data to another site. This research addresses the problem of computing association rules within such a scenario. We will assume homogeneous databases: All sites have the same schema, but each site has information on different entities. The goal is to produce distributed association rule that hold globally, while limiting the information shared about each site.

II. RELATED WORK

Submit the hiding of classification rules is our main focus. Most of the initial work on this field addresses the problem of individual privacy. However, over the past few years interest has increased towards dealing with the problem of hiding sensitive patterns. In [4], a reconstruction algorithm is proposed for classification rules hiding. They proposed an algorithm to preserve the privacy of the classification rules by using reconstruction technique for categorical datasets. In which, only non-sensitive rules of the dataset are used to build a decision tree. Finally, the new dataset which contains only non-sensitive classification rules is reconstructed from the tree. Privacy preserving distributed data mining has increase extensive attractiveness between researchers in the precedent decade. The Privacy preserving distributed data mining algorithms can be classified into two categories depending on the way data is distributed across multiple sites horizontal data partitioning and vertical data partitioning.

Privacy preservation

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 8, February 2015

This refers to the privacy preservation technique used for the selective modification of the data. The techniques used are:

• Heuristic-based techniques modify selected values i.e. changing some data values in a given dataset from an original value to another value.

• Reconstruction-based techniques where the original distribution of the data is reconstructed. These algorithms are implemented by perturbing the data first and then reconstructing the distributions.

• Cryptography-based techniques like secure multiparty computation where a computation is secure if at the end of the computation, no party knows anything except its own input and the results.

Vikas Ashok at al [3]In this research, they take a completely different approach each data owner derives association rules locally, sanitizes them if necessary, and sends them to a third-party data miner. The data miner collects local rules from all data owners, regenerates an estimate of global data, and performs global data mining. They suggest schemes to reduced he generation of spurious rules, a possible outcome of data generation from rules. The proposed method is illustrated using an example of association rule data mining. They are currently in the process of formalizing some of the underlying techniques and to make them more efficient.

D.Karthikeswarant at al[4]This research presents a novel based approach that strategically modifies a few transactions in the database. It modifies support or confidence values for hiding sensitive rules without producing many side effects. Nevertheless, undesired side effects such as nonsensitive rules falsely hidden and spurious rules falsely generated, may be produced in the rule hiding process.

TAMIR TASSA at al[5] they consider the problem of computing efficient anonymizations of partitioned databases. Given a database that is partitioned between several sites, either horizontally or vertically, they devise secure distributed algorithms that allow the different sites to obtain a k-anonymized and diverse view of the union of their databases, without disclosing sensitive information. Our algorithms are based on the sequential algorithm[Goldberger and Tassa 2010] that offers anonymizations with utility that is significantly better than other anonymization algorithms, and in particular those that were implemented so far in the distributed setting. Algorithms can apply to different generalization techniques and utility measures and to any number of sites. While previous distributed algorithms depend on costly cryptographic primitives, the cryptographic assumptions of our solution are surprisingly minimal.

ArashGhorbanniaDelavar at al[6] In this research a novel framework for integrated resource systems was provided which was used in proposed phases. A novel algorithm, Apriori Checksum, was used in this novel proposed framework. By this technique duplicate data can be optimized compared with previous methods or even able to reach main result without repeating. A special condition is provided in novel ERPAC framework which can decrease repetitions in distributed database of Literacy Movement Organization IRAN and finally with providing a new technical approach in proposed Checksum algorithm, the comparison was made after choosing the biggest candidate key as a recursive set to increase the level of customers' satisfaction and competitive service providers, increase time access to information and also productivity in comparison to previous studied framework.

III. PROPOSED METHODOLOGY

Using this approach we will make Knowledge supervision scheme for Researching on Sensitive data, environment and tourism data, medical data, criminal data. By using our technique, data owners can share their data with data miners to find accurate mining result without any concern about violating data privacy. Extract Knowledge from Sensitive data, environmental and tourism data which are on distributed system. Provide privacy on distributed data in the process of data mining. Design Distributed association rules mining algorithm based privacy preserving method for distributed data.

Transaction data, such as shopping basket data, web queries, movie ratings and click streams are extremely useful for association rule mining, recommendation systems and user behavior prediction ect. Publication of these data is more and more necessary. Each transaction consists of some items which are similar to attributes in relational database. As the same case with relational database, besides the identifier items, combination of several items may be linked to a specific individual or reveal individual's sensitive information. So publishing transaction data directly may lead to privacy breach and incur some new challenges in privacy-preserving transaction data publishing. In recent years, many people engaged in researching on privacy-preserving for transaction data publication, and have obtained some achievements However; there is no approach suitable for any privacy-preserving problems. We will propose machine learning technique techniques that have extensive function to determine motivating associations among item set. The approach comprises of two phases. In the first phase, the each local system side compute and supply the corresponding association rules function to a global

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 8, February 2015

system side. The second phase which is Global system side where data miner generates global valid association rules function. By using these global association rule function, a mining result is obtain and mining result is send to each distributed side owner. Most of works pay attention to solving the conflict between privacy-preserving and retaining information utility. In the above scope of research. We will review and compare existing approaches in terms of privacy models, privacy preservation operations, information loss metrics, and a privacy preservation approach using mining of distributed association rule algorithms for transaction data publishing we will Extract Knowledge from Sensitive data, environmental and tourism data which are on distributed system. Machine learning technique has extensive function to determine motivating associations among item set. The approach comprises of two phases. In the first phase, the each local system side compute and supply the corresponding association rules function to a global system side. The second phase which is Global system side where data miner generates global valid association rules function. By using these global association rule function, a mining result is obtain. In this research we will propose a machine learning technique based privacy preserving method for distributed data. We will propose technique provides a proper degree of privacy. However, this technique is aimed mainly to handle distributed data.

Phase 1- Local rules are generate using association rule mining technique for every distributed database.

Phase 2-The minimum support or the minimum confidence of rules is fixed.

For example Apriori algorithm is then applied to follow by filtering to obtain a set of rules, Rules approximating behavior.

Phase 3-The data miner now applies the merging algorithm on the received rules to obtain intermediate table, which contain association rule.

Phase 4-These association rule are used to generate the global data which does not show identification on any entity.

Phase 5-In step 4 data is generated with the help of mining result of each individual distributed database.

Phase 6- If any field of generalized data table is empty then filled by padding technique.

Phase 7-Apply association rule algorithm to global data which does not show identity and generate mining result.

The proposed approach able to recover all actual global rules for a given minimum support and minimum confidence.

Phase 1: receiving Request and assign Resource the proposed technique intelligent to recover all actual global rules for a given smallest support and minimum confidence.

Phase 2: receiving Request and allocate Resource

Phase 3: distribute transaction id into dissimilar nodes.

Phase 3: Generating the list of Item set and distribute the list among nodes.

Phase 4: pronouncement the number of occurrence of item set in every node

Phase 5: Sending an array consist of number of incidence of Itemset to reserve broker

Phase 6: devious the total incidence of each item set and inspection whether it is frequent or not by be suitable threshold value.

Phase 7: Resource broker send a array consist of 0's and 1's appearance which item sets are recurrent to all the nodes.

Phase 8: Nodes are producing (k+1)-item set by a frequent sequence.

Joint Secure File Sharing with distributed association rule There are a number of cases arises that if multi users wish to join every their data, to form one sequential file which is explain in figure 1. Consider the circumstances where multiple parties, every having a private data set (denoted by D1, D2- - - and Dn in that order), want to join every those data sets to form one sequential file (F) by collaboratively behavior distributed association rule. devoid of loss of overview the assumptions can be achieve by preprocessing the data sets D1, D2, - - - and Dn and such a preprocessing do not necessitate one party to send its data set to other parties.

The multi party has been complete with the assist of distributed association rule. A privacy preservation technique will help to check the authority of the user. A privacy preservation technique as well helps to share the file without disclosing any private data of every user. There is no chance for adversary to hack the data for the user.

IV. CONCLUSION

Considering the different size of quantitative attribute values and categorical attribute values in databases, we present two quantitative association rules mining methods with privacy-preserving respectively, one bases on Boolean association rules, which is suitable for the smaller size of quantitative attribute values and categorical attribute values in databases; the other one bases on partially transforming measures, which is suitable for the larger ones. To each approach, the privacy and accuracy are analyzed, and the correctness and feasibility are proven by experiments. Machine learning technique provide privacy at distributed

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 8, February 2015

mining processes mining sensitive data:-Using this approach we will make Knowledge supervision scheme for Researching on Sensitive data, environment and tourism data. machine learning technique based privacy preserving method for distributed data.

REFERENCES

[1] Arun K Pujari "Data Mining Techniques" Second edition 2010, universities (India) Private Limited 2009.Reprint 2011, 2012.

[2] S.Vijayarani, Dr.A.Tamilarasi, R.SeethaLakshmi: Privacy Preserving Data Mining Based on Association Rule- A Survey. International Conference on Communication and Computational Intelligence December, 2010.pp.99-103.

[3] Vikas Ashok, Ravi Mukkamala Virginia, USA- A Novel Approach To Privacy-Preserving Collaborative Distributed Data Mining WPES'11, October 17, 2011, Chicago, Illinois, USA 2011 ACM 978-1-4503-1002-4/11/10.

[4] D.Karthikeswarant, V.M.Sudha, V.M.Suresh A.Javed sultan a pattern based framework for privacy preservation through association rule mining IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM - 2012) March 30, 31, 2012.

[5] EHUD GUDES, TAMIR TASS, Secure Distributed Computation of Anonymized Views of Shared Databases ACM Transactions on Database Systems, Vol. 37, No. 2, Article 11, Publication date: May 2012.

[6] ArashGhorbanniaDelavar, NarjesRohani, Mehdi ZekriyapanahGashti," ERPAC: A Novel Framework for Integrated Distributed Systems Using Data Mining Mechanisms" 2nd International Conference on Software Technology and Engineering (ICSTE) - 2010.

[7] Liming Li, Qishan Zhang ,"A Privacy Preserving Clustering Technique Using Hybrid Data Transformation Method", Proceedings of IEEE International Conference on Grey Systems and Intelligent Services, Nanjing, China, November 10-12, 2009.

[8] Aris Gkoulalas - Divanis and Vassilios S. Verykios, "Association Rule Hiding for Data Mining", Advances in Database Systems, Volume 41, Springer, 2010.

[9] T. Berberoglu and M. Kaya, "Hiding Fuzzy Association Rules in Quantitative Data", The 3rd International Conference on Grid and Pervasive Computing Workshops, May 2008, pp. 387- 392.

[10] Manoj Gupta and R. C. Joshi, "Privacy Preserving Fuzzy Association Rules in Quantitative Data", International Journal of Computer Theory and Engineering, Vol. 1, No. 4, October, 2009, 382-388.