ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 3, Issue 4, April 2017

Privacy Preserving based MASK Algorithm in Big Data

Raghavendran.G.[1]
[1]Senior Grade Lecturer, Department of Electrical and Electronics,
DACG Government Polytechnic, Ratnagiri Road, Chikkamagaluru-577101

ABSTRACT

The amount of data in our world has been exploding. Obtaining valuable information from mass data is a precondition for decision making, which drew more attention of scholars and engineers. Data mining is a powerful data analysis tool, by which rules, patterns and knowledge from large datasets can be extracted. However, people's privacy cannot be effectively protected during the data mining process. The data mining process inevitably causes the exposure of sensitive data, thus resulting in the leakage of privacy. Therefore, how to protect information privacy and security in big data analysis has become a major challenge. In order to deal with the relationship between data privacy preservation and data analysis, privacy-preserving data mining technology has emerged.Association rules mining has been playing an important role in the field of data mining. Since correlation or related links among items can be found by association rules mining, it has been widely applied in decision support by governments, enterprises, and individuals. When it comes to the data mining algorithms for privacy preserving association rules, the low execution efficiency is still a common problem. This work presents a new algorithm, which improves the time efficiency from two aspects (Data Perturbation and Query Restriction) and has a better degree of privacy protection.

I. INTRODUCTION

1 Introduction to Data Mining and Big Data Analytics

1.1 Privacy Preserving Data Mining

Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Mining encompasses various algorithms such as clustering, classification, association rule mining and sequence detection. Traditionally all these algorithms have been developed within a centralized model, with all data being gathered into a central site, and algorithms being run against that data.

Data mining is the process of gathering information about the user specific data, also called knowledge discovery. The problem with data mining output is that it also discloses some information, which is considered to be private and personal. Effortless access to such personal data causes a peril to individual privacy. Official statistics, Health information, and E-commerce are some key concern for privacy.

Privacy Preserving Data Mining technique (PPDM) gives novel way to solve this problem. The main purpose of privacy preserving data mining is to design competent frameworks and algorithms that can extract relevant knowledge from a large amount of data without revealing of any sensitive information. It protects sensitive information by providing sanitized database of original database on the internet or a process is used in such a way that private data and private knowledge remain private even after the mining process. It is PPDM due to which the benefits of data mining be enjoyed, without compromising the privacy of concerned individuals.

PPDM Techniques can be classified over five dimensions.

The first dimension is related to distribution of data i.e. Centralized or Distributed.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 3, Issue 4, April 2017

The second dimension refers to the modification of original values of data that are to be released for data mining task. Modification is carried out using perturbation, blocking, aggregation, merging, swapping or sampling or any combination of these.
The third dimension is that of data mining algorithms. The data mining algorithm are applied on the transformed data to get useful nuggets of information that were hidden previously.
The fourth dimension refers to whether the raw data or aggregated data should be hidden.
The fifth and the final dimension refer to the techniques that are used for protecting privacy.
Based on these dimensions, different PPDM techniques may be classified into following five categories;

1. Anonymization based PPDM
2. Randomized Response based PPDM
3. Condensation approach based PPDM
4. Cryptography based PPDM
5. Perturbation based PPDM

Most current implementations are based any one of the above techniques. The proposed project implements the Privacy Preserving Association Rule Mining is based on Perturbation technique and Data Distribution approach

2 Big Data Analytics

Big Data is the ocean of information we swim in every day – vast zetabytes of data flowing from our computers, mobile devices, and machine sensors. Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.
The three Vs of *volume, velocity and variety* are commonly used to characterize different aspects of big data.
Volume: Enterprises are awash with ever-growing data of all types, easily amassing terabytes even petabytes of information.
Turn 12 terabytes of Tweets created each day into improved product sentiment analysis
Convert 350 billion annual meter readings to better predict power consumption
Velocity: Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.
Scrutinize 5 million trade events created each day to identify potential fraud
Analyze 500 million daily call detail records in real-time to predict customer churn faster
Variety: Big data is any type of data - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together.
Monitor 100's of live video feeds from surveillance cameras to target points of interest
Exploit the 80% data growth in images, video and documents to improve customer satisfaction

Big data, a collection of data sets, is so large and complex that it is beyond the ability of typical database software tools to capture, store, manage, and process the data within a tolerable elapsed time. The existing frequent pattern mining algorithms are not suitable to mine a big data, for the size of a big data is so big that it is inefficient to scan the data multiple time. Moreover, when processing a big data, it requires massively parallel software running on hundreds or even thousands of cheap computers that act as a cluster, instead of autonomously algorithm running on one expensive supercomputer.

3 MapReduce

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model.

Programs written in this functional style are automatically parallelized and executed on a large cluster of computing machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 3, Issue 4, April 2017

machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system.

II. EXISTING SYSTEMS

1 Gap in Existing Systems/Methodologies

As an important research branch of data mining, Association rules mining has made significant progress in privacy protection area. MASK (Mining Associations with Secrecy Konstraints) was first put forward to mine Association rules with privacy protection. Using the methods of randomized disturbance and reconstruction of distribution, privacy-preserving data mining for association rules was achieved in MASK algorithm. But the low time-efficiency of reconstructing original support limited its practical applications. In order to improve the time-efficiency, EMASK (Efficient Mask) was proposed by Agrawal. But the EMASK did not break the exponential complexity of reconstructing original support. A new optimization called MMASK (Modified MASK), was given by Andruszkiewicz, which decreased time complexity than EMASK.. The above algorithms were only based on the data perturbation approach, which led to low privacy-preserving degree. Moreover, the time complexity still had room for improvement.

2 Problem Definition

Data mining is a powerful data analysis tool, by which rules, patterns and knowledge from large datasets can be extracted. However, people's privacy cannot be effectively protected during the data mining process. The mining process inevitably causes the exposure of sensitive data, thus resulting in the leakage of patient privacy. Therefore, how to protect information privacy and security in big data analysis has become a major challenge.

In order to deal with the relationship between data privacy preservation and data analysis, an efficient and secure privacy-preserving data mining technology need to be developed. [2] proposed a secure hash message authentication code. A secure hash message authentication code to avoid certificate revocation list checking is proposed for vehicular ad hoc networks (VANETs). The group signature scheme is widely used in VANETs for secure communication, the existing systems based on group signature scheme provides verification delay in certificate revocation list checking. In order to overcome this delay this paper uses a Hash message authentication code (HMAC). It is used to avoid time consuming CRL checking and it also ensures the integrity of messages. The Hash message authentication code and digital signature algorithm are used to make it more secure . In this scheme the group private keys are distributed by the roadside units (RSUs) and it also manages the vehicles in a localized manner. Finally, cooperative message authentication is used among entities, in which each vehicle only needs to verify a small number of messages, thus greatly alleviating the authentication burden. [3] discussed about Reconstruction of Objects with VSN. By this object reconstruction with feature distribution scheme, efficient processing has to be done on the images received from nodes to reconstruct the image and respond to user query. Object matching methods form the foundation of many state- of-the-art algorithms. Therefore, this feature distribution scheme can be directly applied to several state-of- the-art matching methods with little or no adaptation. The future challenge lies in mapping state-of-the-art matching and reconstruction methods to such a distributed framework. The reconstructed scenes can be converted into a video file format to be displayed as a video, when the user submits the query. This work can be brought into real time by implementing the code on the server side/mobile phone and communicate with several nodes to collect images/objects. This work can be tested in real time with user query results.

3 Objectives

The objectives of this project are,

To study the methodologies proposed by various research fellows on Privacy Preserving Association Rule Mining in Big Data.

Learn the tools and techniques used for Data Perturbation and Distribution for Big Data in Distributed Systems such as Hadoop, MapReduce, Data Mining Tool Weka etc.

Propose and develop a new methodology for Privacy Preserving Association Rule Mining in Big Data.

Implement the proposed new methodology using software tools and report the result

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 3, Issue 4, April 2017

III. LITERATURE SURVEY

Many people have proposed and implemented several algorithm for mining Privacy Preserving Association Rules in Big Data. Few of them are discussed below in brief.

Improved MASK Algorithm by Haoliang Lou, Yunlong Ma, Feng Zhang, Min Liu and Weiming Shen

With the arrival of the big data era, information privacy and security issues become even more crucial. The Mining Associations with Secrecy Konstraints (MASK) algorithm and its improved versions were proposed as data mining approaches for privacy preserving association rules. The MASK algorithm only adopts a data perturbation strategy, which leads to a low privacy-preserving degree. Moreover, it is difficult to apply the MASK algorithm into practices because of its long execution time. They proposed a new algorithm based on data perturbation and query restriction (DPQR) to improve the privacy-preserving degree by multi-parameters perturbation. In order to improve the time-efficiency, the calculation to obtain an inverse matrix is simplified by dividing the matrix into blocks; meanwhile, a further optimization is provided to reduce the number of scanning database by set theory. Both theoretical analyses and experiment results prove that the proposed DPQR algorithm has better performance.

Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data by Feng Zhang, Chunming Rong, Gansen Zhao, Jinxia Wu and Xiangning Wu

There are methods to perform privacy-preserving distributed association rules mining protocols with the assumption of three or more parties at a horizontally partitioned data scenario. However, the method depends on a secure multi-party summary and union computation, which cannot guarantee security while the number of participating parties is two. All final globally frequent itemsets (rules) are disclosed to all participating parties. In a two-party case, it may be argued that knowing an (a) itemset (rule) is supported globally but is not supported at one's site reveals that the other site supports that itemset (rule). This leakage cannot be avoided.

Privacy preserving distributed association rules mining protocols have been developed for horizontally partitioned data scenarios with more than two participating parties. However, they depend on a secure multi-party summary and union computation, which cannot guarantee security while the number of participating parties is two. We use commutative encryption and design a secure division computation protocol as the core techniques to implement the protocols for the privacy-preserving two-party distributed mining of association rule mining. The protocols' security and performance are analyzed.

An efficient algorithm for mining association rules from multiple databases by G. Liu and R. Geng

The mining objects of traditional mining association rules techniques mainly focus on mono-database. With the rapid development of database technologies, multi-database mining is becoming more and more important. In order to make the synthetic result of multiple data sources more accurate, the parameter C, which indicates the number of transactions in local branch database, is proposed. We design a novel and efficient multi-databases mining algorithm using the parameter C. It can reduce the network congestion and solve the new incremental addition of transaction data effectively. Experiment results demonstrate the algorithm is efficient.

Privacy Preserving Data Mining for Association Rules using Optimization for MASK scheme by Piotr Andruszkiewicz

Concerns about privacy of information provided by users and collected to discover hidden knowledge lead to inaccuracy of this data. Users are afraid of revealing sensitive data. The cause to provide wrong data may be the worry that the provided data can be misused. Considering these concerns several methods of preserving privacy for association rules mining have been proposed. The goal of preserving privacy is to encourage people to provide true information, even about sensitive values. People have different concerns about different items (attributes). This regularity can be used to obtain higher accuracy. Thus, several solutions have been proposed to take the advantage of it. Incorporating privacy in association rule mining in MASK (Mining Associations with Secrecy Konstraints) scheme results in time cost.

Several optimizations of MASK were proposed, but no optimization could have broken the exponential complexity of reconstructing the original support of a set based on the distorted database in general case. He presented a new optimization, called MMASK (Modified MASK), which breaks the exponential complexity and achieves better results for higher values of privacy. Effectiveness and accuracy of MMASK have been tested on synthetic and real data sets and compared to Apriori and original MASK scheme.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 3, Issue 4, April 2017

Addressing Efficiency Concerns in Privacy-Preserving Mining by Shipra Agrawal, Vijay Krishnan and Jayant R. Haritsa

Their study was carried out in the context of extracting *association rules* from large historical databases, a popular mining process that identifies interesting correlations between database attributes. For this framework, they presented a scheme called MASK (Mining Associations with Secrecy Konstraints), based on a simple probabilistic distortion of user data, employing random numbers generated from a pre-defined distribution function. It is this distorted information that is eventually supplied to the data miner, along with a description of the distortion procedure. A special feature of MASK is that the distortion process can be implemented at the data source itself, that is, at the *user machine*. This increases the confidence of the user in providing accurate information since she does not have to trust a third-party to distort the data before it is acquired by the service provider.

They addressed the runtime efficiency issue in privacy-preserving association rule mining by improving the MASK scheme. The improvement is achieved through changes in both the distortion process and the mining process of MASK, resulting in a new algorithm that they refer to as EMASK (Efficient MASK). They demonstrated that it is possible to bring the efficiency to well within an order of magnitude with respect to direct mining, while retaining satisfactory privacy and accuracy levels. Their new design is validated against a variety of synthetic and real datasets.

IV. METHODOLOGY

Step 1: To study the methodologies proposed by various research fellows on Privacy Preserving Association Rule Mining in Big Data.

Study and understand various technical papers published by research fellows on Privacy Preserving Association Rule Mining in Big Data.

List the gaps in these techniques and identify any one gap that can be filled to improve the

Step 2: Learn the tools and techniques used for Privacy Preserving Association Rule Mining in Big Data.

Download and install the software tools required and Learn to use them by studying the tutorials and user manuals, and by developing sample applications.

Step 3: Propose and develop a new methodology for Privacy Preserving Association Rule Mining in Big Data.

Using the knowledge gained from Objective 1 and using different innovation techniques develop a new Privacy Preserving Association Rule Mining in Big Data.

Develop a technical paper proposing this new methodology.

Step 4: Implement the proposed new methodology using software tools and report the result.

Generate the Bit Mapping for the Big Transaction Data by assigning 1 for the attributes present in the transaction and 0 for the attributes where were not.

Distort the data set using the Data Perturbation Technique.

Split the perturbed dataset and distribute to n node using Query Restriction Strategy as shown in Figure 1.

Generate the frequent n-itemsets by Reconstructing n-itemset support from distributed distorted dataset using MASK algorithm.

Step 5: Experiment the implementation with different input parameters and report result.

Different Distortion Parameter P – Experiment the algorithm with different Minimum Support and report result.

Different Minimum Support – Experiment the algorithm with different Minimum Support and report result.

Algorithm Tuning – With the results found from above experiment, tune the algorithm for optimum performance

V. RESULTS

File Name: BigTranactionData.dat
1 2 3 4 5 2 4 5 6 7 8 9 10 2 5 7 3 4

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 3, Issue 4, April 2017

1 4 6 8 9 8 9
1. Each row represents a transaction. The row number is considered as "Transaction ID" 2. Each number in the row represents the Item ID of the items bought in that transaction

Distorting and Recovery of the Big Transactional Data

Item ID/Transaction ID	1 2 3 4 5 6	
1 2 3 4 5 6 7 8 9 10	1 0 0 0 1 0 1 1 1 0 0 0 1 0 0 1 0 0 1 1 0 1 1 0 1 1 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0 0 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 0 0	Key Generation Key = (0 0 1 1 0 1)
1 2 3 4 5 … …	1 0 1 1 1 1 … … … …	Distorted Data = Original Data XOR Key E.g. For Item ID 1 1 0 0 0 1 0 XOR 0 0 1 1 0 1 --- 1 0 1 1 1 1 // Distorted Data ---
1 2 3 4 5 … …	1 0 0 0 1 0 … … … …	Restoration of Data = Distroted Data XOR Key E.g. For Item ID 1 1 0 1 1 1 1 XOR 0 0 1 1 0 1 --- 1 0 0 0 1 0 // Restored Data ---
Note: Using the Distortion and Restoration Algorithm is the Same. Since it is binary operation, the performance of the system is improved. The Distortion key shall be kept secret and shall be encoded in the application.		

Finding frequent n-itemset from the recovered data

Item ID/Transaction ID	1 2 3 4 5 6	Item Count	Description
1 2 3 4 5 6 7 8 9 10	1 0 0 0 1 0 1 1 1 0 0 0 1 0 0 1 0 0 1 1 0 1 1 0 1 1 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0 0 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 0 0	2 3 2 4 3 2 2 3 3 1	E.g. Item ID 1 has been bought 2 times in the whole transactions. The transaction ID in which the Item ID 1 has bought are 1 and 5. If Minimum Support Count = 3 then the frequent 1-Itemset are Item ID 2, 4, 5, 8 and 9 Since 1, 3, 6, 7 and 10 are infrequent, as per Apriori Principle all supersets having these items are also infrequent. So for finding 2-Itemsets using only frequent 1-Itemsets in the next iteration

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 3, Issue 4, April 2017

2 4 (Binary AND) 2 5 2 8 2 9 4 5 4 8 4 9 5 8 5 9 8 9	1 1 0 0 0 0 1 1 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 1	2 3 1 1 2 2 2 1 1 3	The 2-Itemsets are found using Binay AND operation on the data of the items participating as set. Now only Itemset (2, 5) and (8, 9) are frequent.
2 5 8 2 5 9 5 8 9	0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0	1 1 1	Since none of the itemsets count is >= given minimum support count of 3, no itesets are selected as frequent.
Output	Min Sup = 3	2 Iterations	Frequent Items = {2, 4, 5, 8, 9, (2, 5), (8, 9) }

VI. FUTURE ENHANCEMENT

The limitations of this proposed algorithm are

1. The proposed approach is more suitable for Boolean data, and it cannot deal with numerical data or other types of data.
2. If the data partitioning among the computing nodes is properly done then there may be chances that few nodes are overly loaded and some are less loaded. Additional pre-processing is required to find the optimum split of data.

As mentioned in the limitations, the proposed algorithm is applicable only for Boolean data. This algorithm can be extended to work for other kind of data types which will increase the applicability of this algorithm for practical purposes.

There is lot of scope for improvement in the data partitioning scheme where optimum partition should be found out with little computation overhead.

REFERENCES

[1]. Haoliang Lou, Yunlong Ma, Feng Zhang, Min Liu and Weiming Shen, "Data Mining for Privacy Preserving Association Rules Based on Improved MASK Algorithm", Computer Supported Cooperative Work in Design (CSCWD), Proceedings of the 2014 IEEE 18th International Conference, 21 - 23 May 2014, pp. 265 – 270.

[2]. Christo Ananth, M.Danya Priyadharshini, "A Secure Hash Message Authentication Code to avoid Certificate Revocation list Checking in Vehicular Adhoc networks", International Journal of Applied Engineering Research (IJAER), Volume 10, Special Issue 2, 2015,(1250-1254).

[3]. Christo Ananth, M.Priscilla, B.Nandhini, S.Manju, S.Shafiqa Shalaysha, "Reconstruction of Objects with VSN", International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST), Vol. 1, Issue 1, April 2015, pp:17-20.

[4]. P. Andruszkiewicz, Optimization for MASK scheme in privacy preserving data mining for association rules, International Conference on Rough Sets and Intelligent Systems Paradigms, Warsaw, 2007, pp. 465-474.

[5]. S. Agrawal, V. Krishnan, and J. R. Haritsa, On addressing efficiency concerns in privacy-preserving mining, International Conference on Database Systems for Advanced Applications, Jeju Island, 2004, pp. 113-114.