Volume 6, Issue 2, February 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Privacy Preservation Data Mining with It's Tools: A State-of-Art

Prashant Namdev
Department of Computer Science & Engineering
RGPV University, Bhopal, India

Manoj Kumar
Department of Computer Science & Engineering
RGPV University, Bhopal, India

Abstract— Privacy refer as a crucial properties of an information system, where systems desired to share information among distinct, not trusted entities, the protection of sensible information plays a vital role. Therefore privacy is becoming important concern issue in various data mining applications. Moreover latest trends displays that classical access control methods are not good enough to guarantee privacy when data mining methods are implemented in a malicious way. Privacy preserving data mining algorithms are being proposed having a motive of preventing the discovery of crucial information. Here we have discussed the usage of cryptography in that data mining for privacy preserving..

Keywords— Privacy Preservation, Data Mining, Data Mining Tools, Parameters for PPDM.

I. INTRODUCTION

Privacy preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure [1]. Most traditional data mining techniques analyze and model the dataset statistically, in aggregation, while privacy preservation is primarily concerned with protecting against disclosure of individual data records. This domain separation points to the technical feasibility of PPDM. Historically, issues related to PPDM were first studied by the national statistical agencies interested in collecting private social and economical data, such as census and tax records, and making it available for analysis by public servants, companies, and researchers. Building accurate socio-economical models is vital for business planning and public policy. Yet, there is no way of knowing in advance what models may be needed, nor is it feasible for the statistical agency to perform all data processing for everyone, playing the role of a "trusted third party." Instead, the agency provides the data in a sanitized form that allows statistical processing and protects the privacy of individual records, solving a problem known as privacy preserving data publishing. For a survey of work in statistical databases see Adam & Wortmann (1989) and Willenborg & de Waal (2001). These papers considered two fundamental problems of PPDM, privacy preserving data collection and mining a dataset partitioned across several private enterprises. Agrawal and Srikant (2000) devised a randomization algorithm that allows a large number of users to contribute their private records for efficient centralized data mining while limiting the disclosure of their values; Lindell and Pinkas (2000) invented a cryptographic protocol for decision tree construction over a dataset horizontally partitioned between two parties. These methods were subsequently refined and extended by many researchers worldwide.

Privacy preserving data mining is an important property that any mining system must satisfy. So far, if we assumed that the information in each database found in mining can be freely shared. Consider a scenario in which two or more parties having confidential databases wants to carry out the data mining algorithm on union of its databases in absence of declaring any unwanted information [2]. For example, consider separate medical institutions that wish to conduct a joint research while preserving the privacy of their patients. In this scenario it is required to protect privileged information, but it is also required to enable its use for research or for other purposes. In particular, although the parties realize that combining their data has some mutual benefit, none of them is willing to reveal its database to any other party. The common definition of privacy in the cryptographic community limits the information that is leaked by the distributed computation to be the information that can be learned from the designated output of the computation.

Although there are several variants of the definition of privacy, for the purpose of this discussion we use the definition that compares the result of the actual computation to that of an "ideal" computation: Consider first a party that is involved in the actual computation of a function (e.g. a data mining algorithm). Consider also an "ideal scenario", where in addition to the original parties there is also a "trusted party" who does not deviate from the behavior that we prescribe for him, and does not attempt to cheat. In the ideal scenario all parties send their inputs to the trusted party, who then computes the function and sends the appropriate results to the other parties. Loosely speaking, a protocol is secure if anything that an adversary can learn in the actual world it can also learn in the ideal world, namely from its own input and from the output it receives from the trusted party. In essence, this means that the protocol that is run in order to compute the function does not leak any "unnecessary" information [3].

II. PRIVACY PRESERVATION TECHNIQUE

Four techniques – sanitation, blocking, distort, and generalization – have been used to hide data items for centralized data distribution. Data sanitation is to remove or modify items in a database to reduce the support of some frequently used item sets such that sensitive patterns cannot be mined[4], [5]. The blocking approach replaces certain attributes of

© 2016, IJARCSSE All Rights Reserved

the data with a question mark. In this regard, the minimum support and confidence level will be altered into a minimum interval. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges, the confidentiality of data is expected to be protected. Data distort protects privacy for individual data records through modification of its original data, in which the original distribution of the data is reconstructed from the randomized data. These techniques aim to design distortion methods after which the true value of any individual record is difficult to ascertain, but "global" properties of the data remain largely unchanged. Generalization transforms and replaces each record value with a corresponding generalized value.

III. DATA MINING

Data mining refer as an crucial instrument to creates patterns or knowledge by data. Data mining technology are being implemented in mine frequent patterns, find associations, perform classification and prediction, etc. The data needed for data mining approach is being taken in a single database or in distributed resources [6].

Currently, the PPDM algorithms are mainly used on the tasks of classification, association rule and clustering. Association analysis involves the discovery of associated rules, showing attribute value and conditions that occur frequently in a given set of data. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Clustering Analysis concerns the problem of decomposing or partitioning a data set (usually multivariate) into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups. A majority of the PPDM algorithms used association rule method for mining data, and then clustering [7].

Data Mining Technique and Distributed data

A. The k-Nearest Neighbor Classifier:

Standard data mining algorithm K-nearest neighbour classification is an instance depends on learning algorithm which works very efficiently against the variety of problem areas. The aim of k nearest neighbor classification is to discover k nearest neighbors for a given instance, then assign a class label to the given instance according to the majority class of the k nearest neighbors. The nearest neighbors of an Instance are defined in terms of a distance function such as: The standard Euclidean distance:

Equation 1

Where r is the number of attributes in a record instance x , ai(x) indicate the ith attribute value of record instance x, and D(xi, xj) is the distance between two instances xi, xj

B. Vertically and Horizontally Data Partition:

When the input to a function is distributed among different sources, the privacy of each data source comes into question. Distribution process of the data plays an key role in defining the errors because of the reason that data could be separated into various parts either horizontally or vertically. After that the vertical partitioning of data shows that various sites or organizations collect distinct information about that similar set of entities or people, for example hospitals and other insurance companies collecting data about the set of people which can be jointly linked. So the data to be mined is the join of data at the sites. In horizontal partitioning, the organizations collect the same information about different entities or people.

As example super markets is collecting transaction information of their clients. As a result, the data to be mined is the union of the data at the sites. In this report it is supposed that all organizations or departments that to be mined have the same information (homogenous) but different entities (records or tuples), so horizontal work process is generated.

IV. DATA MINING TOOLS

Here in this section the open source data mining tools are mentioned

4.1 WEKA

WEKA is Waikato Environment for Knowledge Analysis, data mining/machine learning tool developed by Department of Computer Science, University of Waikato, New Zealand[8]. It is a collection of open source of many data mining and machine learning algorithms, including pre-processing on data, Classification and regression, clustering, association rule extraction, feature selection. It supports .arff (attribute relation file format) file format

4.2 RapidMiner

RapidMiner [9] provides data mining and machine learning procedures including: data loading and transformation (Extract, transform, load, a.k.a. ETL), data preprocessing and visualization, modeling, evaluation, and deployment. RapidMiner is written in the Java programming language. It uses learning schemes and attributes evaluators from the Weka machine learning environment and statistical modeling schemes from R-Project.

4.3 KNIME

KNIME, the Konstanz Information Miner, is an open source data analytics, reporting and integration platform. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept. A graphical user interface allows assembly of nodes for data preprocessing (ETL: Extraction, Transformation, Loading), for modeling and data analysis and visualization.

© 2016, IJARCSSE All Rights Reserved

4.4 Orange

Orange is a component-based data mining and machine learning software suite, featuring a visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It is implemented in C++ and Python.

Tool Name	Type	**Features**
RAPID MINER	Data mining, predictive analytics, statistical analysis,	• Approx. 20 more new functions for the data handling and analysis are involved in many new functions of aggregation • To operate directly File operators are used from the Rapid Miner • Intuitive GUI
ORANGE	Data mining, Machine learning, Data visualization	• Data Analytics found And Interaction Extendable Documentation • Huge toolbox included, also Scripting interface.
KNIME	Business Intelligence , Enterprise Reporting , Data mining	• Intuitive user interface and Scalability , • Structure API for installing extensions • complicated data handling, and Data visualization
WEKA	Use of machine learning technique.	• 3 algorithms for detecting the association rules • 3 graphical user interfaces • Bad documentation

V. LITERATURE REVIEW

The association rule mining (ARM) problem was first described about a decade ago [10], and was formulated in a distributed setting soon after [11], [12]. However, scalability for several dozens of computing nodes was considered satisfactory until recently. The first algorithm for the large-scale distributed ARM problem was presented in [13].

Privacy-preserving data mining has received a lot of attention in the past few years. Perturbation-based techniques have been widely discussed (see [14], [15]). However, because they disclose data source statistics, they are not fit for a distributed setting.

Cryptographically secure versions were developed for three data mining algorithms: distributed ARM (the same problem we discuss) [16], ARM in vertically partitioned data [17] – i.e., where each transaction is split among several nodes, and decision tree induction [18]. These are not scalable because, in contrast to the k-privacy model presented here, the cryptographic primitives they use are global and rigid. That is, the evaluation of every primitive requires the participation of all nodes, and if the data at even one node changes the process has to be repeated from scratch.

The first scalable algorithm for the privacy-preserving distributed ARM problem was presented in [19]. Nevertheless, the algorithm is not secure against malicious participants. The same also holds for all previous work in privacy preserving data mining; they all assume semi-honest attackers (those that must follow the protocol). Some authors refer to the work of Goldreich, Micali and Wigderson [20] – a method by which any private algorithm can be turned into one that assumes malicious participants – as a basis for expanding their work to the malicious model. However, in[21], at the first stage (commitment), each participant musts end a private share of its input to all other participants. In distributed data mining, that input is the local database. Since each share must include as much information as the database will provide for producing the mined model, the method in is not suitable for such scales of data as those Found in the data grid.

Dependent on analysis, Weka will be taken as a very related to the second to KNIME due to its several embedded properties which needed no coding or programming knowledge. In addition to the comparison of, Rapid Miner and Orange will be taken as suitable for modern users, specifically to those in complex sciences, due to the extra skills of programming which are required, and the restricted support of visualization which is offered. It may be decided on the basis of above tables that however the data mining is the fundamental concept to all the tool yet, in those Rapid miner is only the tool that is isolated of the language restrictions and has capabilities of predictive and statistical analysis, Therefore it may be used easily and applied on any of the system, however it aggregated more algorithms of the mentioned other tools.

In paper [8] it have been explained a normal method for the classification of the other software tools of data mining. It is explained a pattern for the categorization of Data Mining software along with a number of reciprocal dimensions, along with the dynamic database of the 41 of very famous tools of data mining. The proposal of business-oriented for the categorization of data mining tools is explained as based on the model type, business goal, process-dependent features, system requirements, user interface features, and vendor information. By the use of those features it had been characterized as the 41 very famous Data Mining tools. At the end, it have decided that by the help of the standard plan and a related database, the users are allowed to choose a software package of data mining, according to its capability, to satisfy the high-level of business targets.

VI. CONCLUSION

The main motive of privacy-preserving data mining is to search accurate, important and potential patterns and rules and predict classification in absence of precise access to the actual data. Here Data mining taken as the crucial frontiers

© 2016, IJARCSSE All Rights Reserved

and most promising interdisciplinary developments in Information technology. Hence, synthesizing a privacy-preserving data mining algorithm mostly needs three key indicators, like as privacy (security), accuracy and efficiency. In this paper, we have analysis various works and aspects of the same.

REFERENCES

[1] Tsiafoulis, S.G. Zorkadis, V.C., 2010, A Neural Network Clustering Based Algorithm for Privacy Preserving Data Mining, International Conference on Computational Intelligence and Security (CIS), 2010, pp: 401-405.

[2] Zhiqiang Yang ; Wright, R.N. 2005, Improved Privacy-Preserving Bayesian Network Parameter Learning on Vertically Partitione Data, 21st International Conference on Data Engineering Workshops, 2005. Pp:1196

[3] Honda, K. ; Kawano, A. ; Notsu, A. ; Ichihashi, H., 2012, A fuzzy variant of k-member clustering for collaborative filtering with data anonymization, Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on, pp: 1-6

[4] Shu Qin Ren , Khin Mi Mi Aung ; Jong Sou Park, 2010, A Privacy Enhanced Data Aggregation Model, Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on, pp: 985 – 990

[5] Mi Wen, Rongxing Lu ; Jingshen Lei ; Xiaohui Liang , 2013, ECQ: An Efficient Conjunctive Query scheme over encrypted multidimensional data in smart grid, Global Communications Conference (GLOBECOM), 2013 IEEE, 796 – 801

[6] Li, Yaping, Chen, Minghua ; Li, Qiwei ; Zhang, Wei, 2012, Enabling Multilevel Trust in Privacy Preserving Data Mining, Knowledge and Data Engineering, IEEE Transactions on (Volume:24, Issue: 9), pp: 1598 – 1612

[7] Kun Liu , Kargupta, H. ; Ryan, J., 2006, Random projection-based multiplicative data perturbation for privacy preserving distributed data mining, Knowledge and Data Engineering, IEEE Transactions on (Volume:18 , Issue: 1), pp: 92 – 106.

[8] Inan, A., Richardson, TX, Kantarcioglu, M., Bertino, E., 2009, Using Anonymized Data for Classification, IEEE 25th International Conference on Data Engineering, 2009. ICDE '09. , pp : 429-430

[9] Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing", IEEE Transactions on Knowledge & Data Engineering, vol.24, no. 3, pp. 561-574, March 2012, doi:10.1109/TKDE.2010.236

[10] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” in Proc. of ACM SIGMOD’93, Washington, D.C., 1993, pp. 207–216.

[11] R. Agrawal and J. Shafer, “Parallel mining of association rules,” IEEE Transactions on Knowledge and Data Engineering, vol. 8(6), pp. 962–969, 1996.

[12] D. Cheung, J. Han, V. Ng, A. Fu, and W. Fu, “A fast distributed algorithm for mining association rules,” in Proc. Of PDIS’96, Florida, December 1996.

[13] R. Wolff and A. Schuster, “Association rule mining in peerto-peer systems,” in Proc. ICDM‘03, November 2003.

[14] R. Agrawal and R. Srikant, “Privacy-preserving data mining,” in Proc. of ACM SIGMOD’00, Dallas, Texas, USA, May 14-19 2000, pp. 439–450.

[15] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules,” in Proc. Of ACM SIGKDD’02, Canada, July 23-26 2002, pp. 217–228.

[16] M. Kantarcioglu and C. Clifton, “Privacy-preserving distributed mining of association rules on horizontally partitioned data,” in Proc. of DMKD’02, June 2002.

[17] J. Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data,” in Proc. of ACM SIGKDD‘02, Edmonton, Alberta, Canada, July 2002.

[18] Y. Lindell and B. Pinkas, “Privacy preserving data mining,” Proc. of Crypto’00, LNCS, vol. 1880, August 2000.

[19] A. Schuster, R. Wolff, and B. Gilburd, “Privacy-preserving association rule mining in large-scale distributed systems,” in Proc. of CCGrid’04. Chicago, Illinois, USA.

[20] O. Goldreich, S. Micali, and A. Wigderson, “How to play any mental game - a completeness theorem for protocols with honest majority,” in Proc. of STOC‘87, 1987, pp. 218–229.

[21] Kabir, S.M.A, Youssef, A.M. ; Elhakeem, A.K., 2007, On Data Distortion for Privacy Preserving Data Mining, Canadian Conference on Electrical and Computer Engineering, 2007. CCECE 2007. , pp : 308 – 311.

© 2016, IJARCSSE All Rights Reserved