International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Survey on Security of Association Rules Using RDT in Distributed Network

Anand Bhabat[1], **R. A. Satao**[2]

[1]Pune University, Smt. Kashibai Navale College of Engineering, Vadgaon(BK), Pune411041, India

[2]Pune University, Smt. Kashibai Navale College of Engineering, Vadgaon(BK), Pune411041, India

Abstract: ***In many data mining methodologies as the data become more and more to process then they are unable to work in a single machine. So as a solution to this distributed paradigm is a suitable scenario. By creating random decision trees to distribute the data in the network it becomes easy to handle the scenario. But in the distributed network maintaining privacy of the data becomes a challenging job. Most of the association rules are becoming so heavy for large datasets. So proposed system introduces an idea of performing horizontal and vertical association using Apriori and Éclat algorithm respectively in the distributed paradigm. For maintaining the privacy of the data we are using the most secured reverse circle cipher cryptography technique.***

Keywords: RDT framework, TF-IDF, Shannon info gain, Apriori mining association rule, Éclat mining association rule, Reverse circle cipher for privacy preserving.

1. Introduction

In the era of modernization as the industries, organization grows rapidly their adoption for information systems also take a leap growth with daily business approach. The advancement of database tools has also expressively increased privacy concerns among paradigm. The existing database makes it possible to gather and store a large amount of person-specific data. With these the database become more vulnerable to threats, security breaches although it gain the advantage of efficiency and productivity. The information breaches from such databases may have serious impact on business. So new access control mechanism for databases and especially for web bases have become a need of today's world. Thus applicable access control approach is to be used to allow access control only to authorized person.

In order to provide privacy over consumers data, organization are maintaining privacy relevance technology .Although the term privacy is synonym of confidentiality it is far different from one another. In other words confidentiality refers to access control mechanism that allows only the authorized user to access the database. Recently encryption technique has gain lot of importance in outsourcing data and maintaining over a large data set. However, data integrity jointly specifies access control and semantic integrity over database. It verifies whether a subject is right to access a dataset and semantic integrity check whether the changes made are appropriate or not. Access control mechanism focuses, that privacy protection cannot be easily handle by traditional access control structure, since traditional access control models does not focus on privacy policies that are concerned with which data object is used for which purpose(s) rather it focus on which user is carrying which action on which data object. Keeping these challenges in mind, a model was proposed that plays a major role in data preserving privacy and appropriate metadata model that support preserving privacy.

An early effort has been made in the area of access control paradigms that focus on relational database. It has been observe that relational database is a high-level model that specify the logical structure of data and made the development of simple declarative languages for specifying an access control policies possible over a network. Earlier proposed model [1], enforce some fundamental principle that set access control model for database that adopt operating system. First principle focus on access control mechanism for database that should be express in terms of relational database, secondly name base access control implements object that are specified by their name, whereas content base access control supports the dataset. Also work in database access control has been carried out and grouped into the areas of discretionary and mandatory [2]. It allows a database administrator to grant and invalidate access privileges that typically refer to entire tables or view and later DBA may specify that others are authorized to grant to access control.

2. Literature Review

This while building any data mining model system assumes that the underlying data should be freely accessible. One of the main challenges in data mining system is its privacy. The basic care should be taken while building data mining system is to protect the input data, yet allow the data miners to extract the useful knowledge models. Hence the privacy and security of such data restrict the use of centralized data. Privacy preserving data mining has been emerged as important method to solve this problem [3].

RDT is described as random decision tree used for the classification purpose. The function of RDT is to predict the value of target variable based on the several input variable. Decision tree also described as a combination of mathematical and computational technique for to aid the description, categorization and generalization of a given set of data. A RDT is a advanced of Bays optimal classifier. Use of multiple RDT offers many advantages over traditional classification approach. RDT is a general approach in which same code works for classification, regression, ranking and multi-label classification. Thus with the same techniques, we can solve four typical learning problems in the same framework. RDT is based on two stages, training and classification and s structure of a random tree is constructed completely independent of the training data.

Number of approaches has been proposed to preser5ve the data privacy using RDT, where firstly data is modified by using different data modification and perturbation-based approaches and then decision tree mining is applied to modified or sanitized dataset. A crypto graphical approach offers a great security and privacy for data mining system but it suffers due to its poor performance. As the size of data increases the performance of system degrades. The statistical approach has been used to mine decision trees [2], association rules [6, 8, 11, 12], and clustering [13], and is popular mainly because of its high performance. So survey tells that statistical approach having higher edge over another because of its great performance as increase in data size.

Another cryptographic approach [10] is based on ID3 decision tree where the training set is distributed between two parties. According to Breiman [14] the complexity of tree has a vital effect on the performance of the system .Performance of the system is depends on the following factors.

- Total number of leaves.
- Tree depth.
- Number of attributes used.

However, the problem of privacy-preserving decision tree mining has yet to be thoroughly understood. In fact, the privacy-preserving decision tree mining method explored in [3] was recently showed to be completely broken, meaning that an adversary can recover the original data from the unsettled one. The reason for the attack to be so powerful was that the adopted data perturbation technique, called *noise adding*. In spite of attack as described above on framework [3] have u useful features: The anxious data can be analyzed by the data miners by using conventional data mining algorithms. On one hand, an owner can protect its data by releasing only the perturbed version and a miner equipped with a specification of the perturbation technique can derive a decision tree that is quite accurate, compared to the one derived from the original data. We believe that this feature makes the framework a good-fit for many real-life applications.

[15] This paper makes several contributions towards privacy-preserving decision tree mining. What is perhaps the most important is that the framework introduced in [3] can be rescued by a new data perturbation technique based on *random substitutions*. This perturbation technique is similar to the randomization techniques used in the context of statistical disclosure control, but is based on a different privacy measure called $\rho1$-to-$\rho2$ privacy breaching and a special type of perturbation matrix called the γ-diagonal matrix. This utilization of both $\rho1$-to-$\rho2$ privacy breaching [8] and γ-diagonal matrix [6] seems to be new. This is because both [8] and [6] were explored in the context of privacy-preserving association rule mining. Further, it was even explicitly considered as an open problem in [6] to extend the results therein to other data mining tasks such as decision tree mining. For instance, the integration of the perturbation matrix of is non-trivial, because system needs to make it work with continuous-valued attributes. As a consequence, system needs to analyze the effect of the dimension size of the matrix with respect to the accuracy of decision trees and the performance of the system. To this end, they introduce a novel error-reduction technique for our data reconstruction, so that it not only prevents a critical problem caused by a large perturbation matrix, but also guarantees a strictly better accuracy.

[16] Represents K- Anonymity based privacy preserving system where a record from a dataset cannot be distinguished from at least k-1 records whose data is also in the dataset. It preserves the accuracy and privacy of the system to good instance. An Anonymization based approach is proposed by the [17] which hides individual’s sensitive data from owner’s record. Here K-anonymity is used for generalization and suppression for data hiding. Background Knowledge and Homogeneity attacks of K-Anonymity Algorithm do not preserve sensitivity of an individual.

[18] tried to preserve data privacy by adding random noise, while making sure that the random noise still preserves the “signal” from the data so that the patterns can still be accurately estimated. A randomization-based Techniques are used to generate random matrices .[19] This approach works with pseudo-data rather than with modifications of original data, this helps in better preservation of privacy than techniques which simply use modifications of the original data. The use of pseudo-data no longer necessitates the redesign of data mining algorithms, since they have the same format as the original data.

Here in [20] author makes use of Anonymizing Demonstrator. The main purpose behind making demonstrator is performing anonymization with user friendly interfaces. Furthermore swapping and recording can be applied to enhance the utility. [21] Propose to add specific noise to the numeric attributes after exploring the decision tree of the original data. Proper data is not revealed to the second party during the mining process. It works on numerical data only; need to work on data quality and security level measurement.

[22] tries to tackled the problem of classification & introduced a generalized privacy preserving variant of the ID3 algorithm for vertically partitioned data distributed over two or more parties . From this they conclude that Required to find a tight upper bound on the complexity .[23] Here a new perturbation based technique proposed a modified C4.5 decision tree classifier where they says that system need various ways to build classifiers which can be used to classify the perturbed data set.

[24] gives a new perturbation and randomization based approach via data set complementation . They suggest that system requires extra storage for storing perturbed and complement of sample data set and fails if all training data sets were leaked.

3. Conclusion

In the proposed approach of mining association rules system efficiently enhance the feature of Éclat algorithm with comparative powerset.Comparitive powerset extract the maximum frequent itemsets from important words which are been decided by TF-IDF and Shannon information gain. Proposed system enforces the powerset with multi recursion methodology to get as maximum as possible of intersection transactions. This method actually enhances the Éclat algorithm to create frequent itemsets on intersection and thereby to reduce the space and time complexity efficiently. System efficiently takes comparatively less processing time to get the rules for the given minimum support than the other mining algorithms like Apriori and Éclat in single machine as the our approach uses RDT for distributed system using efficient Depth first algorithms.

References

[1] E.B. Fernandez, R.C. Summers, and C. Wood, Database Security and Integrity. Addison-Wesley, Feb. 1981.

[2] R. Ramakrishnan and J. Gehrke. Database Management Systems. McGraw-Hill, 3rd edition, 2003.

[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *ACM SIGMODInternational Conference on Management of Data*, pages 439–450. ACM, 2000.

[4] Dakshi Agrawal and Charu C. Aggrawal. On the design and quantification of privacy preserving data mining algorithms. In *ACM Symposium on Principles of Database Systems*, 2001.

[5] Shipra Agrawal and Jayant R. Haritsa. A framework for high-accuracy privacy preserving mining. In *IEEE International Conference on Data Engineering*, 2005.

[6] J. Vaidya, C. Clifton, and M. Zhu, Privacy-Preserving Data Mining.ser. Advances in Information Security first ed., vol. 19, Springer-Verlag, 2005.

[7] Cynthia Dwork and Kobbi Nissim. Privacy–preserving data mining on vertically partitioned databases. Microsoft Research, 2004.

[8] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaching in privacy preserving data mining. In *ACM*.

[9] W. Fan, H. Wang, P.S. Yu, and S. Ma, "Is Random Model Better? On Its Accuracy and Efficiency," Proc. Third IEEE Int'l Conf. Data Mining (ICDM '03), pp. 51-58, 2003.

[10] Y. Lindell and B. Pinkas. Privacy preserving data mining. In M. Bellare, editor, *Advances in Cryptology – Crypto 2000*, pages 36–54. Springer, 2000. Lecture Notesin Computer Science No. 1880.

[11] A. Evmievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *International Conference on Knowledge Discovery and - Data Mining*, 2002.

[12] Shariq J. Rizvi and Jayant R. Haritsa. Maintaining data privacy in association rule mining. In *International - Conference on Very Large Data Bases*, 2002.

[13] Srujana Merugu and Joydeep Ghosh. Privacy-preserving distributed clustering using generative models. In *IEEE International Conference on Data Mining*, 2003.

[14] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth Int. Group, 1984.

[15] Jim Dowd, Shouhuai Xu, and Weining Zhang "Privacy-Preserving Decision Tree Mining Based on Random Substitutions"

[16] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", in proceedings of Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2002.

[17] Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira, "Efficient Multi-Dimensional Suppression for K-Anonymity", in proceedings of IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 3. (March 2010), pp. 334-347, IEEE 2010.

[18] H. Kargupta and S. Datta, Q. Wang and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", in proceedings of the Third IEEE International Conference on Data Mining, IEEE 2003.

[19] C. Aggarwal , P.S. Yu, "A condensation approach to privacy preserving data mining", in proceedings of International Conference on Extending Database Technology (EDBT), pp. 183–199, 2004. 746.

[20] Martin Beck and Michael Marh¨ofer," Privacy-Preserving Data Mining Demonstrator", in proceedings of 16th International Conference on Intelligence in Next Generation Networks, IEEE 2012.

[21] Mohammad Ali Kadampur, Somayajulu D.V.L.N." A Noise Addition Scheme in Decision Tree for Privacy Preserving Data Mining", IEEE 2010.

[22] J. Vaidya and C. Clifton. Privacy-preserving decision trees over vertically partitioned data. In *Proceedings of the 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security*, Storrs, Connecticut, 2005. Springer. L. Liu, M. Kantarcioglu, and B. Thuraisingham, "Privacy Preserving Decision Tree Mining from Perturbed Data," Proc. 42nd Hawaii Int'l Conf.

[23] Li Liu Global Information Security eBay Inc," Privacy Preserving Decision Tree Mining from Perturbed Data" ,Proceedings of the 42nd Hawaii International Conference on System Sciences – 2009

[24] B.OBULESU1, D.SIREESHA2," PRIVACY PRESERVING DECISION TREE LEARNING USING UNREALIZED DATA SETS" , International Conference on Computer, Control and Cognitive Sciences, 1 st September, 2013, Bangalore, I.