International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 1 | Issue – 5

Privacy Preserving Approaches for High Dimensional Data

Tata Gayathri
Assistant Professor, Department of CSE,
Shri Vishnu engineering college for women,
Bhimavaram, Andhra Pradesh, India

N Durga
Assistant Professor, Department of CSE,
Shri Vishnu engineering college for women,
Bhimavaram, Andhra Pradesh, India

ABSTRACT

This paper proposes a model for hiding sensitive association rules for Privacy preserving in high dimensional data. Privacy preservation is a big challenge in data mining. The protection of sensitive information becomes a critical issue when releasing data to outside parties. Association rule mining could be very useful in such situations. It could be used to identify all the possible ways by which 'non-confidential' data can reveal 'confidential' data, which is commonly known as 'inference problem'. This issue is solved using Association Rule Hiding (ARH) techniques in Privacy Preserving Data Mining (PPDM). Association rule hiding aims to conceal these association rules so that no sensitive information can be mined from the database.

Keywords*: Bigdata, Association Rule Mining, Association Rule Hiding*

1. INTRODUCTION

Privacy preserving is important in wherein data mining turns into a cooperative assignment among members. Privacy preserving data mining is an important topic on which lot of researchers going on last years. There are many approaches to hide association rule. In this paper Efficient Heuristic approach method is proposed which is more effective to hide association rule. The objective of this algorithm is to extract relevant knowledge from large amount of data, while protecting at the time sensitive information. The proposed method focused on hiding set of frequent items containing highly sensitive knowledge that only remove information from transactional database with no hiding failure.

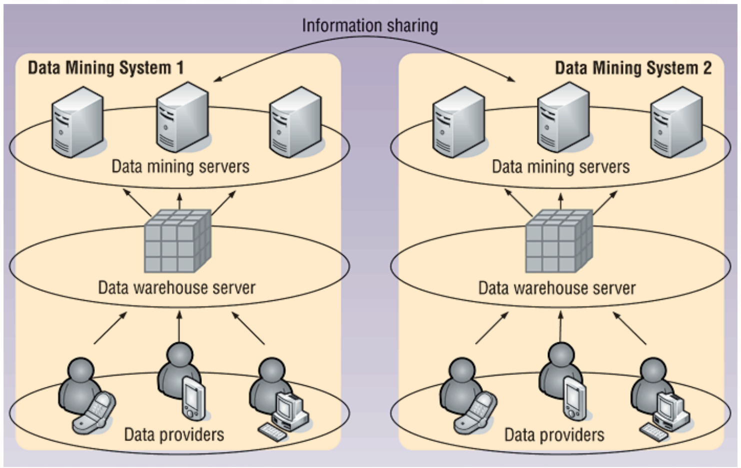

Figure 1: Architecture for Privacy Preserving system

Advances in computer networks and data acquisition techniques have enabled the collection and storage of huge amounts of data. This data is of no use until it is analyzed and and then analyzed to find patterns. To get more precise data patterns, organizations share their data, which can compromise the privacy of users and their data. There are many techniques are developed to ensure security and privacy of data. In this lines several cryptographic techniques are such as homomorphic encryption, secure computation, verifiable computation and threshold cryptographic techniques. As a solution to the privacy issues in distributed data-mining, privacy-preserving data mining was introduced by Agarwal et al [1] and Lindell and Pinkas [2]. Privacy-preserving distributed data-mining is the cooperative computation of data that is distributed among multiple parties without revealing any of their private data items. A. Data Privacy and Security The privacy of data is suitably defined as the appropriate use of data. Securing sensitive data is usually known as data security and usually referred to as the availability, confidentiality and integrity of data. Data security guarantees that the data is correct, dependable and accessible when those with permitted access require it. Organizations want to endorse a policy of data security for the single purpose of guarantying data privacy or the privacy of their consumers'data, particularly when it is in use. One strategy for protecting the privacy of the individual records is to perturb the original data. Data perturbation procedures are statistically based strategies that try to ensure secret data by adding random noise to private, numerical attributes, thereby shielding the original data. B. Privacy Preserving Data Mining Consider a circumstance in which more than two parties having sensitive information intend to processes a calculation on the mix of their inputs without uncovering any undesirable data. In the ideal circumstance each participant sends their inputs to the classified party, who next processes the capacity and sends the right results to alternate party without losing security of individual inputs. In this way we can preserve privacy even in the presence of adversarial participants that attempt to gather information about the inputs of their parties. After Lindell et.al proposal on concept of secure computation in the field of data mining, since then, privacy preserving distributed data mining has attracted much attention and many secure protocols have been proposed for specific data mining algorithms.

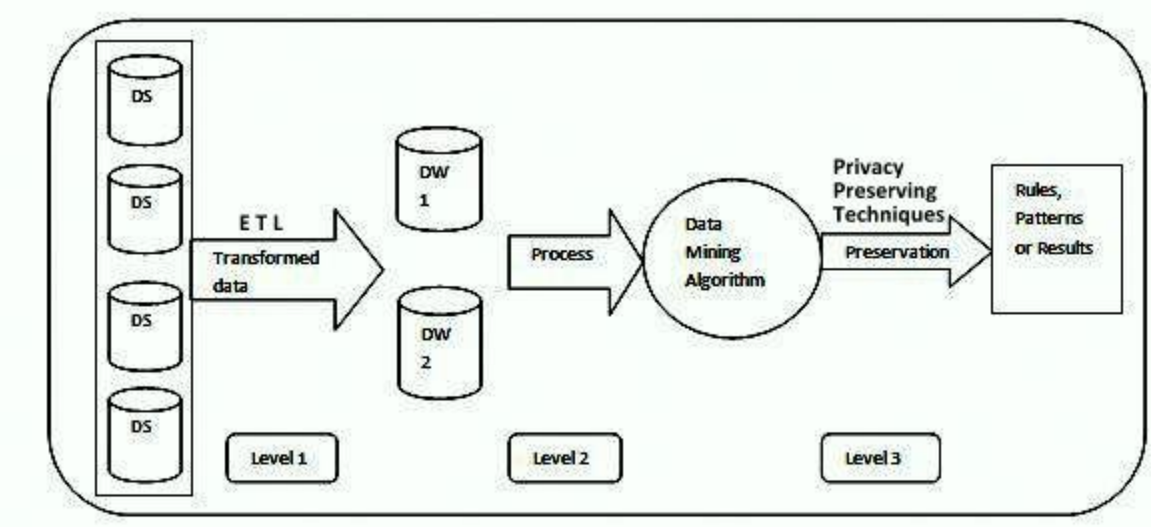

Figure 2: Overview of Privacy Preserving Distributed Data Mining protocols

II. Tools for Privacy Preserving Distributed Data Mining

Secure Sum

Given a number of values belonging to n entities

We need to compute ∑ xi for i= 1 to n

Such that each entity ONLY knows its input and the result of the computation (The aggregate sum of the data)

III. PRIVACY PRESERVINGIN BIG DATA

Data is currently one of the most important assets for companies in every field. The continuous growth in the importance and volume of data has created a new problem: it cannot be handled by traditional analysis techniques. This problem was, therefore, solved through the creation of a new paradigm: Big Data. However, Big Data originated new issues related not only to the volume or the variety of the data, but also to data security and privacy. In order to obtain a full perspective of the problem, we decided to carry out an investigation with the objective of highlighting the main issues regarding Big Data security, and also the solutions proposed by the scientific community to solve them. In this paper, we explain the results obtained after applying a systematic mapping study to security in the Big Data ecosystem. It is almost impossible to carry out detailed research into the entire topic of security, and the outcome of this research is, therefore, a big picture of the main problems related to security in a Big Data system, along with the principal solutions to them proposed by the research community

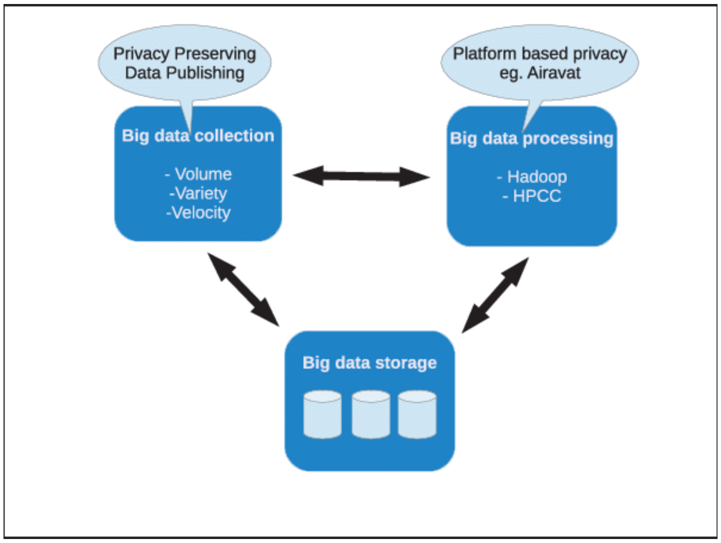

Figure 3: Privacy Preserving In Big Data

IV. ASSOCIATION RULE

Let I = {i1,…., in} be a set of items. Let D be a database which contains set of transactions. Each transaction t _ D is an item set such that t is a proper subset of I. As transaction t supports X, a set of items in I, if X is a proper subset of t. Assume that the items in a transaction or an item set are sorted in lexicographic order. An association rule is an implication of the form X_Y, where X and Y are subsets of I and X_Y= Ø. The support of rule X_Y can be calculated by the following equation: Support(X_Y) = |X_Y| / |D|, where |X_Y| denotes the number of transactions containing the itemset XY in the database, |D| denotes the number of the transactions in the database D. The confidence of rule is computed by Confidence(X_Y) = |X_Y|/|X|, where |X| is number of transactions in database D that contains itemset X. A rule X_Y is strong if support(X_Y) _ min_support and confidence(X_Y) _ min_confidence, where min_support and min_confidence are two given minimum thresholds. Association rule mining algorithms calculate the support and confidence of the rules. The rules having support and confidence higher than the user specified minimum support and confidence are retrieved. Association rule hiding algorithms prevents the sensitive rules from being revealed out. The problem can be declared as follows “Database D, minimum confidence, minimum support are given and a set R of rules are mined from database D. A subset SR of R is denoted as set of sensitive association rules.SR is to be hidden. The objective is to modify D into a database D’ from which no association rule in SR will be mined and all non sensitive rules in R could still be mined from D.

IV. APPROACHES OF ASSOCIATION RULE HIDING ALGORITHMS

Association rule hiding algorithms can be divided into three distinct approaches. They are heuristic approaches, border-revision approaches and exact approaches.

➢ Heuristic Approach

Heuristic approaches can be further categorized into distortion based schemes and blocking based schemes. To hide sensitive item sets, distortion based scheme changes certain items in selected transactions from 1’s to 0’s and vice versa. Blocking based scheme replaces certain items in selected transactions with unknowns. These approaches have been getting focus of attention for majority of the researchers due to their efficiency, scalability and quick responses.

➢ Border Revision Approach

Border revision approach modifies borders in the lattice of the frequent and infrequent item sets to hide sensitive association rules. This approach tracks the border of the non sensitive frequent item sets and greedily applies data modification that may have minimal impact on the quality to accommodate the hiding sensitive rules. Researchers proposed many border revision approach algorithms such as BBA (Border Based Approach), Max– Min1 and MaxMin2 to hide sensitive association rules. The algorithms uses different techniques such as deleting specific sensitive items and also attempt to minimize the number of non sensitive item sets that may be lost while sanitization is performed over the original database in order to protect sensitive rules.

➢ Exact Approach

Third class of approach is non heuristic algorithm called exact, which conceive hiding process as constraint satisfaction problem. These problems are solved by integer programming. This approach can be concerned as descendant of border based methodology.

V. Association Rule Hiding Framework

In order to hide an association rule, X → Y, we can either decrease its support or its confidence to be smaller than user-specified minimum support transaction (MST) and minimum confidence transaction (MCT). To decrease the confidence of a rule, we can either (1) increase the support o of X, the left hand side of the rule, but not support of X → Y, or (2) decrease the support of the item set X →Y .For the second case, if we only decrease the support of Y, the right hand side of the rule, it would reduce the confidence faster than simply reducing the support of X → Y. To decrease support of an item, we will modify one item at a time by changing from 1 to 0 or from 0 to 1 in a selected transaction.

Based on these two concepts, we propose a new association rule hiding algorithm for hiding sensitive items in association rules. In our algorithm, a rule X → Y is hidden by decreasing the support value of X →Y and increasing the support value of X. That can increase and decrease the support of the LHS and RHS item of the rule correspondingly. This algorithm first tries to hide the rules in which item to be hidden i.e., X is in right hand side and then tries to hide the rules in which X is in left hand side. For this algorithm t is a transaction, T is a set of transactions, R is used for rule, RHS (R) is Right Hand Side of rule R, LHS (R) is the left hand side of the rule R, Confidence (R) is the confidence of the rule R, a set of items H to be hidden.

ALGORITHM:

INPUT: A source database D, A minimum support min_support (MST), a minimum confidence min_confidence (MCT), a set of hidden items X.

OUTPUT: The sanitized database D, where rules containing X on Left Hand Side (LHS) or Right Hand Side (RHS) will be hidden.

Steps of algorithm:

```
1. Begin
2. Generate all possible rule from given items X;
3. Compute confidence of all the rules for each hidden item H, compute confidence of rule R.
4. For each rule R in which H is in RHS 4.1 If confidence (R) < MCT, then Go to next 2-itemset;
Else go to step 5
 5. Decrease Support of RHS item H.
 5.1 Find T=t in D fully support R;
 5.2 While (T is not empty)
5.3 Choose the first transaction t from T;
5.4 Modify t by putting 0 instead of 1 for RHS item;
5.5 Remove and save the first transaction t from T; End While
6. Compute confidence of R;
7. If T is empty, then H cannot be hidden;
8. For each rule R in which is in LHS
9. Increase Support of LHS;
10. Find T=t in D| t does not support R;
11. While (T is not empty)
12. Modify t by putting 1 instead of 0 for LHS item;
13. Remove and save the first transaction t from T; End While
14. Compute confidence of R;
15. If T is empty, then H cannot be hidden; End For; End Else; End For;
16. Output update D, as the transformed D;
```

The framework of these approach is shown in figure

VI. SECURITY ANALYSIS

The total computation cost of the clustering is depends on the initial clusters and the number of iterations required for finding final clusters. A.

Privacy Theorem

The privacy of the secret data can be acheived stated earlier is fulfilled.

Proof: As we have seen, the chosen codeword C, can be reconstructed by specifying any of its N components. In [n, k, d] MDS code, message symbols are of any of k symbols are taken. Even out of n, if (k − 1) servers are compromised even though secret cannot be reconstructed. This way we can acheive the privacy preserving of the data. Less than k symbols or an unauthorized set recovering probability of the secret is equal to same as the that of the exhaustive search, which is 1 q .

Theorem: The PPDM protocol is efficient and ideal.

Proof: Initially, we distribute the secret data to each servers is given exactly one share. Also, the chosen secret data sets and the generated shares space is Fq. Shares are distributed uniquely and randomly to the servers efficiently. So, the proposed algorithm is ideal and efficient.

CONCLUSION

Privacy becomes an important factor in data mining so that the sensitive information is not revealed after mining. But the data quality is important such that no false information is provided and the privacy is not jeopardized. Association rule is one category of data mining technique. Other data mining techniques should also be considered for securing both data and knowledge.

REFERENCES

[1] Agrawal, R., and Srikant, R. (2000). Privacy Preserving Data Mining.ACM SIGMOD International Conference on Management of Data, SIGMOD00, Dallas, USA.439-450.

[2] Lindell Y, B Pinkas, Privacy preserving data mining, in CRYPTO 00: Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology (Springer, London, 2000), pp. 3654,2000.

[3] Lindell, Y, Pinkas, B. (2002). Privacy Preserving Data Mining. Journal of Cryptology, 15 (3), 177-206. (An extended abstract appeared in Advances in Cryptology, CRYPTO00. 36-54.)

[4] G. Aggarwal, N. Mishra, and B. Pinkas. Secure computation of the k-th ranked element. In Proc. Advances in Cryptology EUROCRYPT 2004, volume 3027 of LNCS, pages 4055. Springer, 2004.

[5] G. R. Blakley, Safeguarding cryptographic keys, AFIPS, Vol. 48 (1979), pp. 313-317.

[6] Shamir, A. 1979. How to share a secret.Comm. ACM 22, 612613.

[7] A.Naidu.T, P.Paul, V.Ch.Venkaiah., Computationally perfect secret sharing schemes Based on MDS codes. International Journal of Trust Management in Computing and Communications (IJTMCC), Vol. 2, No.4, pp 353-378, 2014.

[8] M. Mignotte. How to share a secret.In T. Beth, editor, CryptographyProceedings of the Work-shop on Cryptography, Burg Feuerstein, 1982, Vol.149, LNCS, pp. 371-375.Springer-Verlag, 1983.

[9] Asmuth, C., Bloom, J.: A modular approach to key safeguarding. IEEE Transactions on Information Theory IT-29(2),pp. 208-210 (1983).

[10] Josef Pieprzyk, and Xian-Mo Zhang, Ideal threshold schemes from MDS codes, ICSC, vol. 2587, pp. 253 - 263, 2003.

[11] Pinkas, B.: Cryptographic techniques for privacy-preserving data mining. SIGKDD Explor.Newslett. 4(2), 1219 (2002),

[12] E. Bertino, I.N. Fovino, L.P. Provenza. A Framework for Evaluating Privacy Preserving Data Mining Algorithms. Data Mining and Knowledge Discovery, 11 (2): pp. 121-154, 2005.

[13] Verykios, S., Bertino, E., Fovino, I., Provenza, L., Saygin, Y., Theodoridis, Y.: State of the-art in Privacy Preserving Data Mining. ACM SIGMOD Record 33(1), 5057 (2004).

[14] V Baby and Subhash N Chandra. Privacy-Preserving Distributed Data Mining Techniques: A Survey. International Journal of Computer Applications 143(10):37-41, June 2016.

[15] J.Brickell and V.Shmatikov. Privacy-preserving classifier learning.In Proc. 13th International Conference on Financial Cryptography and Data Security, 2009.

[16] Jagannathan, G., Wright, R.N.: Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In: KDD, pp. 593599 (2005).

[17] Bunn, P., Ostrovsky, R.: Secure two-party k-means clustering. In: ACM Conference on Computer and Communications Security, pp. 486-497 (2007).

[18] Jha, S., Kruger, L., McDaniel, P.: Privacy Preserving Clustering. In: di Vimercati, S.d.C., Syverson, P.F., Gollmann, D. (eds.) ESORICS 2005. LNCS, vol. 3679, pp. 397417.Springer, Heidelberg (2005).

[19] Upmanyu, M., Namboodiri, A.M., Srinathan, K., Jawahar, C.V.: Efficient Privacy Preserving K-Means Clustering. PAISI 2010. LNCS, vol. 6122, pp. 154166. Springer, Heidelberg (2010).

[20] Doganay, M.C., Pedersen, T.B., Saygin, Y., Savas, E., Ltributed privacy preserving k-means clustering with additive secret sharing. In: 2008 International Workshop on Privacy and Anonymity in Information Society, Nantes, France, pp. 311 (2008).

[21] GeoffroyCouteau, Efficient Secure Comparison Protocols - Cryptology ePrint Archive eprint.iacr.org/2016/544.pdf.