IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. VI (Jan – Feb. 2015), PP 70-76
www.iosrjournals.org

Analysis of Pattern Transformation Algorithms for Sensitive Knowledge Protection in Data Mining

Cynthia Selvi.P[1], Mohamed Shanavas.A.R[2]
1Associate Professor, Dept. of Computer Science, KNGA College(W), Thanjavur 613007 /Affiliated to Bharathidasan University, Tiruchirapalli, TamilNadu, India.
2Associate Professor, Dept. of Computer Science, Jamal Mohamed College, Tiruchirapalli 620 020/ Affiliated to Bharathidasan University, Tiruchirapalli, TamilNadu, India.

Abstract: *Privacy Preserving Data Mining(PPDM) is an emerging research area. Its objective is to extend the traditional data mining techniques to work with transformed data while protecting confidential information with minimized loss of privacy and information. In other words PPDM techniques aimed at modifying the original data while still being able to recover the data mining results from the modified data. This article analyses the existing PPDM techniques and presents the analysis of the proposed algorithms in the research work.*

Keywords: *Cover, privacy preservation, restrictive patterns, sanitization, sensitive transactions, victim transactions*

I. Introduction

PPDM Data mining is a growing area of study in computer science and it is applied to many fields. However, malicious usage may cause privacy problems. But it is a challenging task to perform data mining without violating privacy of data or knowledge. This necessity emerged privacy preserving data mining.

The main consideration in PPDM is two-fold: first, Input-privacy protection which modifies sensitive raw data like identifiers, names, addresses and the like in order to produce sanitized data which would be used for mining process. Second, output privacy protection which modifies sensitive raw knowledge(patterns) and releases sanitized(modified) patterns. Various approaches have been adopted to deal with PPDM and many algorithms have been proposed.

1.1. Privacy Preserving Techniques

There are three main privacy preserving techniques for data mining namely, heuristic-based techniques, cryptography-based techniques, and reconstruction-based techniques [1].

1) Heuristic-based Techniques: Heuristic-based approaches are mainly used in centralized database environment. They are similar to adaptive modification that modifies only selected values rather than all available values in order to minimize the utility loss. There are several heuristic-based approaches that hide both raw and aggregated data using different hiding techniques such as k-anonymization, data swapping, generalization, sampling, adding noises and sanitization.

2) Cryptography-based Techniques: These techniques are mainly used in distributed environment and are problem specific. They are applied on distributed data and in a situation where more than one party is involved to securely and collaboratively perform a computation task based on their private inputs.

3) Reconstruction-based Techniques: These techniques are applied on both centralized and distributed data. In order to preserve privacy, data is first perturbed and then, its distribution is reconstructed at an aggregated level in order to apply data mining. Since it is not possible to estimate the original values of individual records accurately, a reconstruction procedure is employed to estimate the distribution of the original values.

1.2. Technique Selection Criteria

In choosing different privacy preserving techniques, a number of selection criteria should be taken into account. In general, no privacy preserving algorithm can outperform others on every aspect. The work presented in [1] identifies four evaluation criteria i.e. performance, data utility, level of uncertainty, and resistance in order to select a suitable privacy preserving technique.

- **Performance –** it refers to the computational cost(the time) required to achieve privacy preserving.
- **data utility –** privacy preserving algorithms should be have minimum privacy loss, the information loss, and the loss of functionality of the data.
- **level of uncertainty –** means that no opponent or counterpart can predict the sensitive information that is hidden. In general, the algorithm with maximum uncertainty level will be preferred over other algorithms.
- **Resistance –** a given privacy preserving algorithm protect the private data against a particular data mining algorithm, but it may not provide the same protection against other data mining techniques.

1.3. Contributions of the Paper

- Various privacy preserving techniques with their characteristics have been discussed.
- A set of heuristic-based sanitization type pattern transformation algorithms for hiding sensitive frequent patterns(itemsets) have been developed using both objective and subjective measures of interestingness.
- The improvements over performance issues like parallel and incremental approaches have been addressed in order to handle large and dynamic databases.
- An approach for group-based sensitivity disclosure is proposed to introduce a balance between privacy and utility of the databases.
- The analysis over the results of the proposed algorithms is presented.

II. Literature Survey

Recently, researchers have been concentrating on the concept of protecting sensitive knowledge in association rule mining; many approaches have been proposed so far, that includes anonymization, randomization and sanitization.

In [2] an optimal algorithm that start from a fully generalized table and specialize the dataset in a minimal k-anonymous table is presented; an algorithm that uses a bottom-up technique and apriori computation is described in [3]; a top-down heuristic to make a table to be released with k-anonymity is proposed in [4]. In [5] randomization technique is applied which add noise to the original data and mask the attribute values of records.

The procedure of transforming the source database into a new database that hides some sensitive knowledge/rules is called sanitization process[6]. The underlying principle of data sanitization that reduce the support values of restrictive rules was initially introduced by Atallah et. al[7] and they proved that the optimality in sanitization process is NP-hard problem. In [8] the authors investigated confidentiality issues of a broad category of association rules and proposed some algorithms to preserve privacy of such rules above a given privacy threshold. In the same direction Saygin[9] introduced some algorithms to obscure a given set of sensitive rules by replacing known values with unknowns, while minimizing the side-effects on non-sensitive rules. Like the algorithms in [8], these algorithms are CPU-intensive and require various scans depending on the no. of association rules to be hidden. A framework for protecting restrictive patterns and sanitizing algorithms are introduced in [10] that require two scans. In [11]-[13], algorithms that focus on the heuristic based sanitization approach are proposed which overcome the limitations of the previous work to a high extent.

Cryptography based approaches are presented in[14] that proposes a transformation framework that allows to systematically transform normal computations to secure multiparty computations; the article[15] also presents four secure multiparty computation based methods that can support privacy preserving data mining.

A random matrix-based spectral filtering technique to recover the original data from the perturbed data is proposed in [16]. Two other data reconstruction methods, PCA-DR and MLE-DR have been proposed in [17]. In addition, several distribution reconstruction algorithms have been proposed [18]-[19] in correspondence with different randomization operators.

III. Preliminaries

Transactional Database: A transactional database consists of a file where each record represents a transaction that typically includes a unique identity number (trans_id) and a list of items that make up the transaction.

Association Rule: It is an expression of the form , $X \Longrightarrow Y$, where X and Y contain one or more patterns(categorical values) without common elements.

Frequent Pattern: A pattern(itemset) that forms an association rule is said to be frequent if it satisfies a prespecified minimum support threshold.

Restrictive Pattern: Pattern to be hidden from the transactional source database according to some privacy policies.

Sensitive Transaction: A transaction is said to be sensitive, if it contain atleast one restrictive pattern.

Transaction Size: The number of items which make up a transaction is the size of the transaction.

Transaction Degree: The degree of a sensitive transaction is defined as the number of restrictive patterns which it contains.

Victim item: Item that is to be deleted in the sensitive transaction of a restrictive pattern whose support count is to be decreased.

Null Transactions: A set of transactions is said to be null transactions($\sim S_T$) if they do not contain any of the patterns being examined.

Cover: The Cover[11] of an item A_k can be defined as, $C_{Ak} = \{ rp_i \mid A_k \in rp_i \subset R_P, 1 \leq i \leq |R_P|\}$
i.e., set of all restrictive patterns which contain A_k.

The item that is included in a maximum number of rp_i's is the one with maximal cover or maxCover;
i.e., maxCover = max($|C_{A1}|, |C_{A2}|, \ldots |C_{An}|$) such that $A_k \in rp_i \subset R_P$.

IV. Sanitization Technique

Given the source dataset(D), and the restrictive patterns(R_P), the goal of the sanitization process is to protect R_P against the mining techniques used to disclose them. To alleviate the complexity of the optimal sanitization, some heuristics could be used. A heuristic does not guarantee the optimal solution but usually finds a solution close to the best one in a faster response time. The sanitization process decreases the support values of restrictive patterns by removing items from sensitive transactions. This process mainly includes four subtasks:

1. identify the set of sensitive transactions for each restrictive pattern;
2. select the (partial) sensitive transactions to sanitize;
3. identify the candidate item(victim item) to be removed;
4. rewrite the modified database after removing the victim items.

Basically, all sanitizing algorithms differs only in subtasks 2 & 3. The algorithms proposed in the research work are given in Fig.1

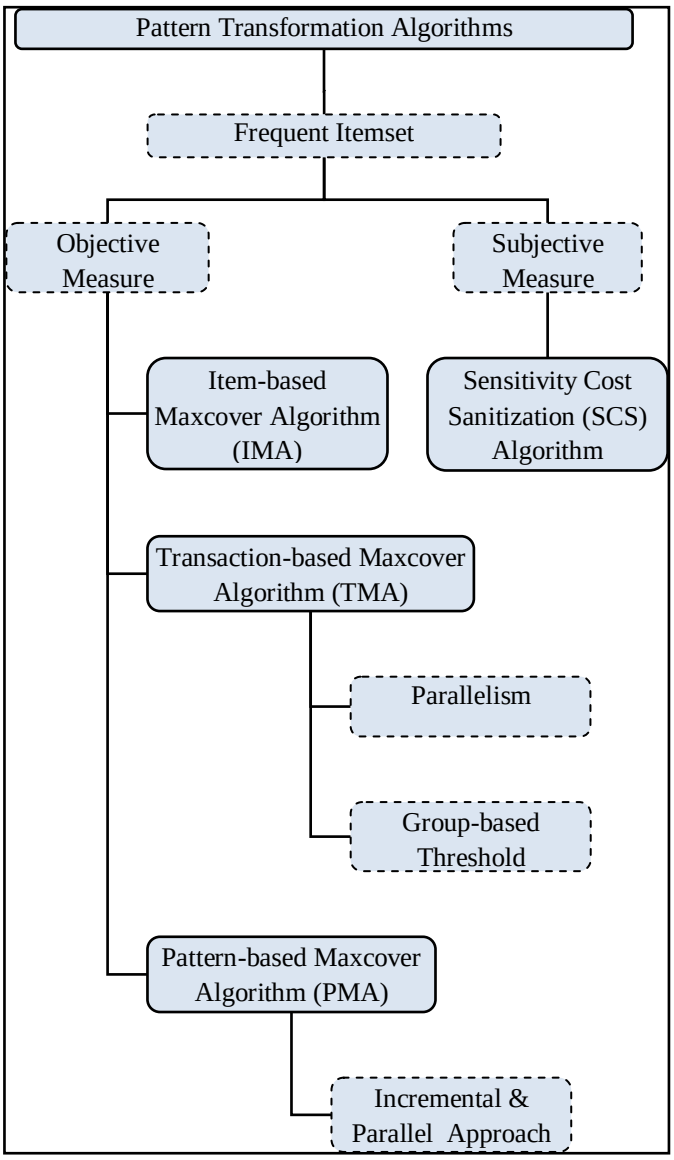


Fig.1 Taxonomy of Proposed Algorithms

4.1. Outline of the proposed algorithms

Scan Source Dataset(D);
Construct hash tables;
Find sensitive transactions(S_T);
Extract null transactions(~S_T);
Apply the heuristic & sanitize sensitive transactions(S_T');
Obtain D' ← $S_T' \cup \sim S_T$;
Release Sanitized Dataset(D').

The heuristic used in the proposed algorithms are stated below:

4.2. Heuristics of Proposed Algorithms

These algorithms are based on the support count (frequency of occurrence) and hiding is enforced by decreasing the support count according to the respective heuristic of the algorithms.

Item-based: Maxcover Algorithm(IMA)[11]:

Heuristic : Victim item is selected based on maxcover value of the items in all restrictive patterns.

Theorem: The running time of the IMA is $O[N(n_1+n_2)]$ in the worst case, where n_1 is the number of restrictive items(A_k), n_2 is the number of restrictive patterns(rp_i) and N is the number of transactions in the source dataset.

Pattern-based Maxcover Algorithm(PMA)[12]:

Heuristic: Find T = $\bigcap_{i=1}^{|RP|} t\_list(rpi)$; for every t $\in$ T, select the victim item A_k with maximal cover such that $A_k \in rp_i \subset t$ and remove; In case of tie, choose one in round robin; Continue selecting the victim items on this strategy for the left over transactions until the support count of all rp_i becomes zero.

Transaction-based Maxcover Algorithm(TMA)[13]:

Heuristic: For every transaction t of $rp_i \in R_P$, select a victim item, A_k with maximal cover within t such that $A_k \in rp_i \subset t$; In case of tie, choose one in round robin.

Sensitivity Cost Sanitization(SCS) Algorithm[20]:

This algorithm has been proposed to hide sensitive items based on the subjective measure of interestingness which would be set according to the users' belief & expectations.

Sensitivity Cost: It is a user-defined value associated with individual item based on its significance in the sensitive patterns or rules.

Heuristic: The individual items in the restrictive patterns are associated with a boolean cost vector; for every transaction t of $rp_i \in R_P$, select a victim item A_k with cost = 1 and maxcover within t such that $A_k \in rp_i \subset t$; In case of tie, choose one in round robin.

Theorem: The running time of the PMA, TMA and SCS is $O(n \times N)$ in the worst case, where n is the number of restrictive patterns(rp_i) and N is the number of transactions in the source dataset.

4.3. Performance Improvements

The following improvements have been introduced to enhance the performance of the proposed algorithms in order to handle potentially large, dynamically updated databases and also to introduce a tradeoff between the privacy and utility of the databases.

Parallel approach[21]: For very large databases a partitioned approach may be used. Here, the sanitization task is distributed among the Server and the Clients and the task is implemented as two modules namely Server Module and Client Module.

Incremental approach[22]: As the world is filled with dynamic data which grows rapidly than what we expect, some interesting new rules may be introduced in the existing databases and some existing rules may become obsolete. In order to handle such updated database, this approach performs incremental sanitization to save time and effort.

Group Privacy Threshold[23]: An approach using a privacy measure(numeric) which determines the sensitivity level of different group of associated items(patterns) especially in transactional databases is proposed in order to maintain a balance between privacy and utility of the database.

this approach, the goal is to hide a group of frequent patterns which contains highly sensitive knowledge. Such sensitive patterns that should be hidden are called restrictive patterns. Restrictive patterns can always be generated from frequent patterns.

V. Experimental Analysis

The proposed algorithms were executed for restrictive patterns (with support ranging between 0.6 and 5, confidence between 32.5 and 85.7 and length between 2 and 6) chosen under various criteria given below using the real dataset T10I4D100K[24]; The test run was made on AMD Turion II N550 Dual core processor with 2.6 GHz speed and 2GB RAM operating on 32 bit OS; The implementation of the proposed algorithm was done with windows 7 - Netbeans 6.9.1 - SQL 2005. The performance issues are studied based on the metrics suggested in [10].

Criteria-I : Restrictive Patterns (5 patterns/1K transactions) were chosen based on the following Strategies and results are shown in Fig.2 to Fig.5.

S1 – Overlapped Patterns

S2 – Mutually Exclusive Patterns

S3 – Randomly chosen Patterns

S4 – High Support Patterns

S5 – Low Support Patterns

Criteria-II : Number of Restrictive Patterns ranging between 5 and 25 (randomly chosen in 1K transactions) were selected and the results are shown in Fig.6 to Fig.9.
Criteria-III : Number of Transactions ranging between 2K and 8K (with 5 randomly chosen Restrictive Patterns) were used and the results are shown in Fig.10 to Fig.13.
The frequent patterns were obtained using Matrix Apriori[25] which use simpler data structures for implementation.

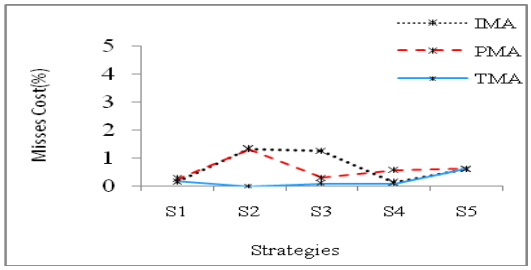
5.1. Measures of Effectiveness

Hiding Failure(HF): It is measured by the ratio of the number of restrictive patterns in the released sanitized database(D') to the ones in the given source database, which is given by HF = $\frac{|RP(D')|}{|RP(D)|}$. The proposed algorithms have 0% HF.
Misses Cost(MC): This measure deals with the legitimate patterns(non restrictive patterns) that were accidently missed. MC = $\frac{|\sim RP(D)| - |\sim RP(D')|}{|\sim RP(D)|}$. The MC is found to be very minimum for all the three proposed algorithms.
Artifactual Pattern(AP): AP occurs when D' is released with some artificially generated patterns after applying the privacy preservation approach and it is given by, AP = $\frac{|P'| - |P \cap P'|}{|P'|}$. As the proposed approach does not introduce any false drops, the AP is 0% .
Sanitization Rate(SR): It is defined as the ratio of the selectively deleted items(victim items) to the total support count of restrictive patterns(rp_i) in the source database D and is given by, SR = $\frac{|\text{victim items}|}{\text{total supCount}(rpi)}$ and it is found to be less (75%) except for the rules that are mutually exclusive.

5.2. Measures of Efficiency

Dissimilarity(dif): The dissimilarity between the original(D) and sanitized(D') databases is measured in terms of their contents which can be measured by the formula,
dif(D,D')= $\frac{1}{\sum_{i=1}^{n} fD(i)} \times \sum_{i=1}^{n}[fD(i) - fD'(i)]$, where fx(i) represents the i^{th} item in the dataset X. The proposed approach has very low percentage of dissimilarity.
CPU Time: The execution time is observed to be minimum and it interesting to note that this approach has very good scalability. The time requirement can further be reduced by adapting parallelism. However, time is not a significant criteria, as the sanitization is done offline.

5.3. Results



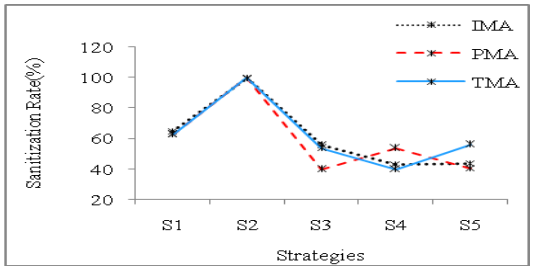
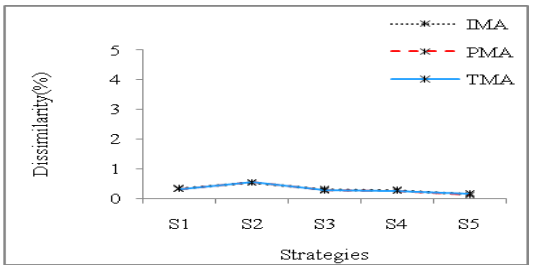


Fig.2& 3. Effectiveness measures for various strategies of rules



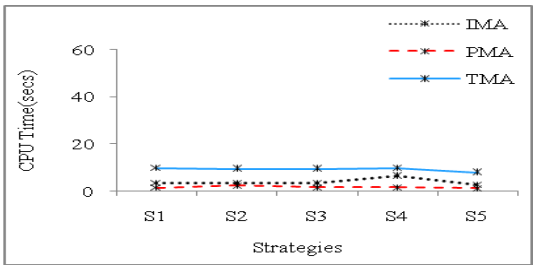
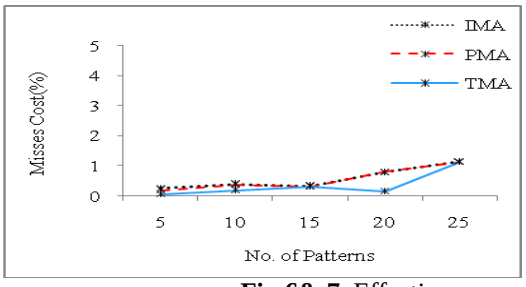


Fig.4 & 5. Efficiency measures for various strategies of rules



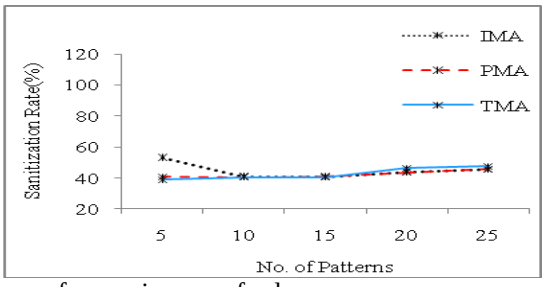
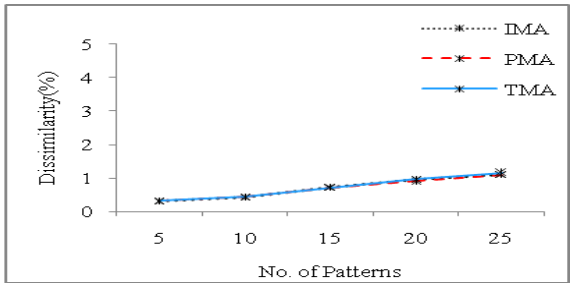


Fig.6& 7. Effectiveness measures for varying no. of rules



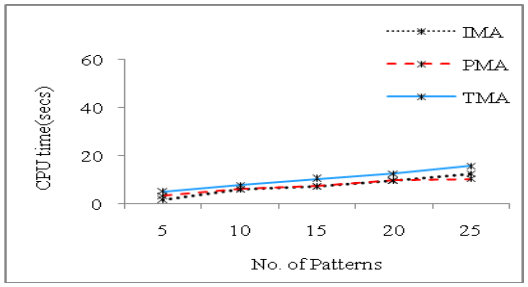
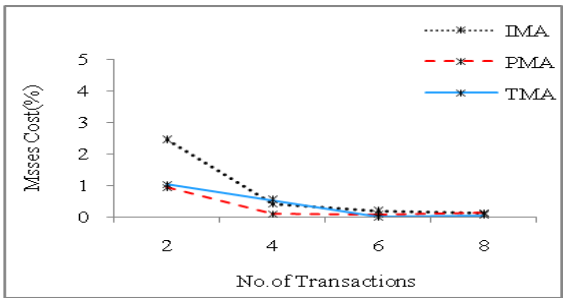


Fig.8 & 9. Efficiency measures for varying no. of rules



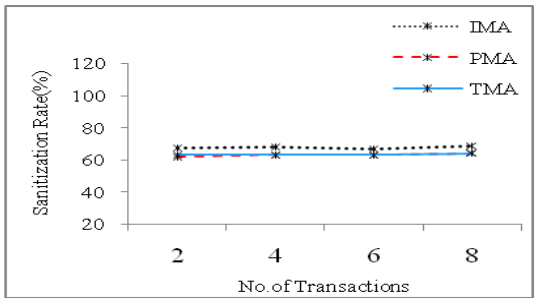
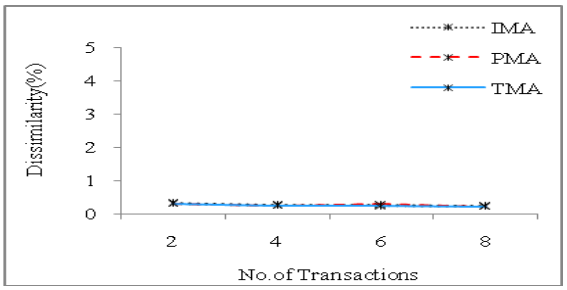


Fig.10& 11. Effectiveness measures for varying size(in K)of dataset



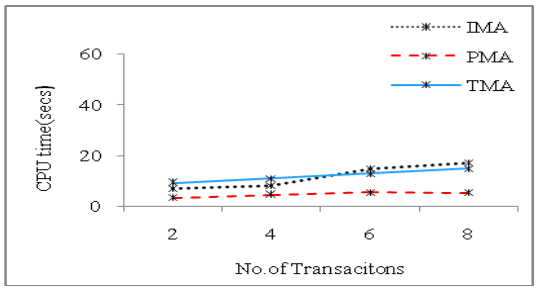


Fig.12& 13. Efficiency measures for varying size(in K)of dataset

5.4. Analysis

From the above graphs, the following observations are made:

General:

- All the three main algorithms are aimed at multiple rule hiding with no hiding failure at all and hence the privacy loss is minimized at the maximum extent.
- The legitimate rules which are accidently hidden is found to be minimum for all the three algorithms.
- This approach has reduced SR (except for the mutually exclusive patterns) which preserves the impact of the database.
- The proposed algorithms have minimum dissimilarity rate which shows that the information loss is very low and so the utility is well preserved.
- The performance(execution time) of all the three algorithms are linear.

Specific:

Among all the three algorithms, PMA is found to be better than the other two in terms of its effectiveness and efficiency under all criteria in which the algorithms are tested.

VI. Conclusion

The proposed algorithms do not require costly cryptographic operations (as mining over encrypted data) or sophisticated protocols (as secure multiparty computation based) to preserve privacy. These algorithms make use of simpler data structures and require single scan of the source database. The proposed sanitization algorithms perform multiple rule hiding with a minimal removal of items comparing to the total support count of the restrictive items. Moreover, it is proved that these algorithms have maximum uncertainty level, meaning that no counterpart or adversary can infer the hidden information even with very low thresholds and in no way reconstruction is possible. The execution time is observed to be linear. This simulation facilitates the sanitization to be done for variegated set of sensitive patterns for different collaborators.

References

[1]. V.Verykios, E.Bertino, I.N.Favino, L.P.Provenza, Y.Saygin, and Y.Theodoridis. "State-of-the-art in privacy preserving data minin", SIGMOD Record, Vol. 33, No. 1, 2004.

[2]. R.Bayardo, R.Agrawal, "Data Privacy Through Optimal k-Anonymization", In Proceedings the 21st International Conference on Data Engineering, pp.217-228, 2005.

[3]. K.Lefevre, J.Dewittd, R.Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity", In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp.49-60, 2005.

[4]. B.Fung, K.Wang, P.Yu, "Top-down Specialization for Information and Privacy Preservation", In Proceedings of the 21st IEEE International Conference on Data Engineering, pp.205-216, 2005.

[5]. R.Agrawal, R.Srikant,"Privacy-Preserving Data Mining", ACM SIGMOD Record (29): 439-450, 2000.

[6]. A.Evfimievski, R.Srikant, R.Agarwal, Gehrke, "Privacy Preserving mining of association rules", Proceedings of 8th ACM SIGKDD international conference on Knowledge discovery and data mining, Alberta, Canada, p.217-28, 2002.

[7]. M.Atallah, E.Bertino, A.Elamagarmid, M.Ibrahim and V.Verykios, "Disclosure Limitation of Sensitive Rules", In Proc. of IEEE knowledge and Data Engg.workshop, p.45-52, Chicogo, Illinois, Nov.1999.

[8]. E.Dessani, V.S.Verykios, A.K.Elamagarmid and E.Bertino, "Hiding association rules by using confidence and support", In 4th information hiding workshop, p.369-383, Pitsburg, PA, Apr 2001.

[9]. Y.Saygin, V.S.Verikios and C.Clifton, "Using unknowns to Prevent Discovery of Association Rules", SIGMOD Record, 30(4):45-54, Dec,2001.

[10]. S.R.M.Oliveria and O.R.Zaiane, "Privacy Preserving frequent itemset mining", in Proc. of the IEEE ICDM workshop on Privacy, Security and data mining, p.43-54, Malbashi city, Japan, Dec.2002.

[11]. Cynthia Selvi P, Mohamed Shanavas A.R, 'An Improved Item-based Maxcover Algorithm to Protect Sensitive Patterns in Large Databases', IOSR- Journal on Computer Engineering(JCE), Vol.14, Issue.4, Sep-Oct.2013, PP.01-05, DOI. 10.9790/ 0661-1440105,

[12]. Cynthia Selvi P, Mohamed Shanavas A.R, 'Output Privacy Protection with Pattern-Based Heuristic Algorithm', International Journal of Computer Science & Information Technology (IJCSIT), Vol.6, No.2, Apr. 2014, PP 141-147, doi.10.5121/ijcsit.2014.6210, airccse.org /journal/**ijcsit**.html

[13]. Cynthia Selvi P, Mohamed Shanavas A.R, 'Towards Information Privacy using Transaction-based Maxcover Algorithm', World Applied Sciences Journal(WASJ)-29 (Data Mining and Soft Computing Techniques): PP.06-11, 2014, DOI:10.5829/ idosi.wasj.2014.29. dmsct.2.

[14]. Wenliang Du and J.Mikhail, M.Attalah, "Secure multi-problem computation problems and their applications: A review and open problems", Tech. Report CERIAS Tech Report 2001-51, Center for Education and Research in Information Assurance and Security and Department of Computer Sciences, Purdue University, West Lafayette, IN 47906, 2001.

[15]. Chris Clifton, Murat Kantarcioglou, Xiadong Lin, and Michael Y. Zhu, "Tools for privacy preserving distributed data mining", SIGKDD Explorations **4** (2002), no. 2.

[16]. H.Kargupta, S.Datta, Q.Wang, K.Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", In Proceedings of the 3rd International Conference on DataMining, pp.99-106, 2003.

[17]. Z.Huang, W.Du, B.Chen, "Deriving Private Information from Randomized Data", In Proceedings of the ACM SIGMOD Conference on Management of Data, Baltimore, Maryland, USA, pp.37-48, 2005.

[18]. D.Agrawal, C.C.Aggarwal,"On the Design and Quantification of Privacy Preserving Data Mining Algorithms", In Proceedings of the 20th ACM SIGMOD-SIGACTSIGART Symposium on Principles of Database Systems, pp.247-255, 2001.

[19]. S.Rizvi, J.Haritsa, "Maintaining Data Privacy in Association Rule Mining", In Proceedings the 28th International Conference on Very Large Data Bases, pp.682-693, 2002.

[20]. Cynthia Selvi P, Mohamed Shanavas A.R, 'Protection by Subjective Measure', International Journal of Computer Application(IJCA), 0975 – 8887, Vol.84, No.10, Dec. 2013, PP.23-26.

[21]. Cynthia Selvi P, Mohamed Shanavas A.R, 'Introducing Parallelism in Privacy Preserving Data Mining Algorithms', International Journal of Computational Engineering Research(IJCER), Vol.04, Issue.2, Feb.2014, PP.38-41.

[22]. Cynthia Selvi P, Mohamed Shanavas A.R, 'An Incremental Sanitization Approach in Dynamic Databases', International Journal of Computer Science and Engineering Technology(IJCSET), E-ISSN.2229-3345, Vol.5, No.09, Sep.2014, PP.872-876,.

[23]. Cynthia Selvi P, Mohamed Shanavas A.R, 'Trade-off between Utility and Security using Group Privacy Threshold Sanitization', International Journal of Computer Sciences and Engineering(IJCSE), E-ISSN.2347-2693, Vol.02 , Issue.09, Sep.2014, PP-08-11.

[24]. The Dataset used in this work for experimental analysis was generated using the generator from IBM Almaden Quest research group and is publicly available from http://fimi.ua.ac.be/data/.

[25]. J.Pavon, S.Viana, S.Gomez, "Matrix Apriori: speeding up the search for frequent patterns," Proc. 24th IASTED International Conference on Databases and Applications, 2006, pp. 75-82.