International Journal of Current Trends in Engineering & Technology
www.ijctet.org, ISSN: 2395-3152
Volume: 04, Issue: 01 (January- February, 2018)

A Comparative Study of Data Suppression Technique for Privacy of Individuals

Ravindra Tiwari, Dr. Priti Maheshwari, Dr. Binod Kumar
Dept. Computer Science AISECT University Bhopal
9907494354

Abstract:- *As number of data mining application is increasing for different purposes like extraction of information and patterns, so sensitive information need to be secured. So privacy preserving mining helps in this field by using the anonymizing the selected portion of the dataset which includes sensitive or private data. In this paper k-anonymity, super class substitution and item removal methods of privacy preserving mining are compared for finding best suppression method. Here each method secured all type of data which include either numeric or text column. Each method was compared on same set of dataset. Results show that class substitution method was better on various evaluation parameters.*

KEYWORDS: - Distributed Data, Data Mining, Encryption, Effective Pruning, Super class substitution.

I. INTRODUCTION

Knowledge Discovery in Databases (KDDs) is the way toward recognizing substantial, novel, valuable, and justifiable examples from huge informational indexes. Information Data mining is the basic tool of knowledge discovery database calculation which provides initial investigation, different models and develops identifiable examples. But the major drawback of this procedure was that it extracts some of sensitive information from the dataset which may harmful for individual, industry, or organization, etc. This sensitive information related to medical reports, personal document, some action of any class or community, etc. This can be understand as let a medical hospital maintain the data of patients for internal assessment which include patient personal information, disease, etc. So if outsider want to know that what kind of treatment that person was taken than utilization of data mining will easily help him to identify patient with all records of prescription. Here privacy needs to be maintained for the same so that reverse mining procedure need to be applied for the dataset. To avoid such conditions, security controls were proclaimed in various countries. The data proprietor is required to disregard recognizing data so that to ensure, with high likelihood, that private information about individuals can't be derived from the educational accumulation that is released for examination or sent to another data proprietor [12]. Privacy mining deals with the tradeoff between the practicality of the mining technique and insurance of the subjects, going for constraining the security introduction with unimportant effect on mining comes to fruition.

Annonymity Tables in [3] considers the issue of discharging tables from a social database containing individual records, while guaranteeing singular protection and keeping up information honesty. Mysterious Connections and Onion Routing [6] give unknown associations that are emphatically impervious to both eavesdropping and activity examination. Database Security – Concepts, Approaches and Challenges in [2] gives a total answer for information security must meet the basic fundamentals. This paper causes us to study the most significant ideas hidden although of database security and condense the most surely understood systems. Data sharing crosswise over Private Databases [7] implicitly accept that the information in every database can be uncovered to alternate databases.

II. RELATED WORK

R. Agrawal and R. Srikant [1] use ARM (Association Rule Mining) approach on huge database. This paper exhibit two calculation in light of association rules that find connection between things. Despite the fact that resulting parameters values were reduced with increment in database. One more point is that it doesn't consider thing quantity data.

T. Calders and S. Verwer [2] uses Naive Bayes approach for order of substantial database. Here researcher cluster dataset based on highly frequent sensitive thing sets. Here separation was done based on sexual orientation, race, and so forth which are regular class of the general population. So partition done on this premise is against law, which should overwhelm in the dataset. Albeit numeric esteems introduce in the dataset stay same as past, so it requires being irritated as it contains numerous sensitive relations.

F. Kamiran and T. Calders [3] display another approach of arrangement of database based on non discriminating thing sets. So nearness of discriminating thing in dataset for arrangement isn't required. Here direct evacuation of sensitive data was performing. This is conceivable by testing in the dataset, here examining influence information to free from segregation. Here discriminating models are not taken for assessment that no data was mined from operated information. So classifying on the basis of discriminating issues is not a moral view.

In [8] multilevel protection is give by the researcher, essential idea create in this paper was to build separated irrelative duplicate of the dataset for various client. Here client are separate into their trust level so base on the trust level dataset irritation rate get increment. Here paper settle one issue of database remaking by brushing the distinctive level irritated duplicate at that point recover into single unique database. So to conquer this

issue irritation of next level is done in bothered duplicate of past one. Along these lines if lower trust client get consolidate and attempt to recover unique dataset then just a single higher annoyed duplicate can be recover. The conveyance of the passages in such a framework looks like corner-waves started from the lower right corner.

In [9, 13] paper cover another issue for the direct indirect segregation avoidance in the dataset. Here it will gather segregate thing set which help in delivering the association manage for recognizing the immediate or indirect principles. At that point shroud the standards which are over the edge an incentive by changing over the X→Y to X→Y' where X is an arrangement of segregating thing this tend to conceal the data which will create just those rules that not give any discriminating guideline. Here Y is change to Y' implies an opposite esteem was pass at few sessions.

III. PROPOSED WORK

In this algorithm three approaches are compared, although each has their own steps for perturbing the dataset. But comparing is done in this section for finding best method of perturbing the dataset which will reduce the risk while size of dataset remains same.

DATA SET

As original dataset need to be perturbed so it should be read from the stored file format to the matrix in which each row and column is same as the dataset but here it can be read and written much easier way. Some time dataset is represented as the sequence of data where pattern need to be found in order to generate the set while text on the basis of the pattern one can judge that this data is of same column and same row.

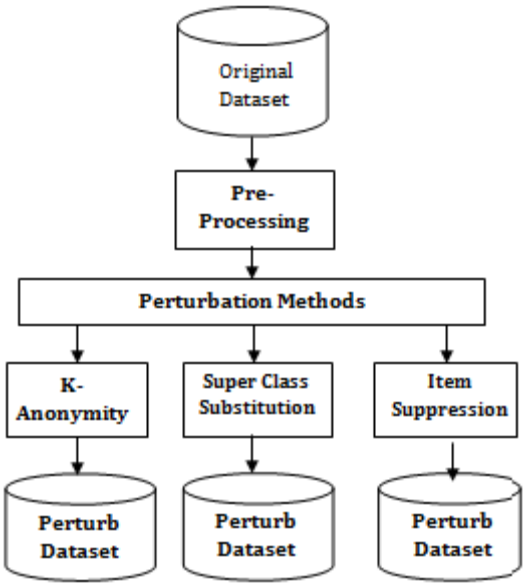

Fig. 1 Flowchart for Privacy Preserving Mining of multiple attributes using Perturbation.

PRE-PROCESSING

As the dataset obtained from the above steps contain many unnecessary information which one need to be removed for making proper operation on those sets. This can be comprehended as let the identity number be the same as it is in the original set so to put this segment in the preprocessed dataset isn't important and it can be expelled from the above arrangement of vectors, while if to hide data of the pin code of the individual then one needs to roll out improvements from the first, subsequently this sort of numeric information which should be hideaway was handled by our strategy.

Super Class Substitution

In this step whole multi attributes are replace by its hierarchy value in the super modularity tree, while replacing it is required to balance the dataset utility and risk by making required changes. This was done in [Base Paper]. This replacement is so designed that utility of the data get increase while risk remain below under some threshold value. Here replacement of above content is done for making two level of anonymity of data, this required to find single row pattern where elements are replace by super class. If pattern present in more than two sessions than super class substitution is not required. By doing this as compare

K-Anonymity

K-Anonymity protects against identity disclosure when the attacker knows the subset of the population represented in the dataset, knows the true attributes of all individuals in the population (including those in the dataset), and knows what data was published for each entry in the dataset. K-Anonymity is a privacy preserving method for limiting disclosure of private information in data mining. The process of anonym zing a database table typically involves generalizing table entries and, consequently, it incurs loss of relevant information. This motivates the search for anonymity algorithms that achieve the required level of anonymity while incurring a minimal loss of information. In case of k-anonymity algorithm two level of anonymity was achieved by introducing one more fake row of the session whose similar copy is not available in the dataset. Here this method increases the dataset size as the number of privacy level get increases.

Suppression

In this method sensitive item in the dataset get removed by just replacing blank in the cell of the row. Here this reduces whole risk while utilization of the dataset also gets highly reduced. In this work suppression of single row elements are removed while those session who have achieved two level of anonymity are remain same so algorithm not required increasing the dataset size as compared to k-anonymity algorithm.

IV. EXPERIMENT AND RESULT

Dataset

In order to analyze proposed algorithm, it is in need of the dataset. So college admission dataset is use that has following attribute {branch, course, gender, pin code, etc.}. Here student information is pin code, gender, branch while sensitive items are important for the admission dataset owner. So for the privacy preservation both things need hide. So in order to provide protection against the private data of the customer one concept of super modularity has been include which make multiple copy of the same student with different values.

Evaluation Parameters

Risk: - In this parameter the sum of information is done where highest subclass get higher value of risk. Each set of attribute have different set of subclass so risk of sharing information vary as per value pass in the perturbed dataset.

$$R = \frac{R(i, j)}{j}$$

Utility: - In this parameter the sum of information is done where highest subclass get higher value of utility. Each set of attribute have different set of subclass so utility of sharing information vary as per value pass in the perturbed dataset.

$$U = \log \frac{U(i, j)}{j}$$

Results

Dataset Percentage	Risk Value		
	K-Anonymity	Suppression	Super Class Substitution
20	8.3292×10^3	8.2687×10^3	8.064×10^3
30	1.2496x104	1.2413x104	1.2132×10^4
40	1.6662×10^4	1.6560×10^4	1.6214×10^4
50	2.0829×10^4	2.0711×10^4	2.0312×10^4
60	2.4996×10^4	2.4863×10^4	2.4416×10^4

Table 1 Comparison of risk value of various data suppression methods.

From table 1 it is obtained that the risk value of the dataset was quit low in super class substitution as compare to k-anonymity. While lowest value of risk was present in item suppression method. In other words previous work has reduced the risk value but to less extent. It was obtained that Item removal have reduce risk as compare to super class suppression.

Dataset Percentage	Utility Value		
	K-Anonymity	Suppression	Super Class Substitution
20	1.3919×10^3	586.8622	2.4964×10^3
30	2.3544×10^3	1.4414×10^3	4.0661×10^3
40	3.4515×10^3	2.5351×10^3	5.7726×10^3
50	4.7041×10^3	3.8877×10^3	7.6272×10^3
60	6.0443×10^3	5.3612×10^3	9.5510×10^3

Table 2 Comparison of Utility value of various data suppression methods.

From table 2 it is obtained that the utility value of the dataset was quit high in super class substitution as compare to k-anonymity. While lowest value of utility was present in item suppression method. In other words previous work has reduced the utility value but to less extent. It was obtained that Item removal have reduce risk as compare to super class suppression.

Dataset Percentage	Dataset size		
	K-Anonymity	Suppression	Super Class Substitution
20	3587	2000	2000
30	5181	3000	3000
40	6990	4000	4000
50	8107	5000	5000
60	9481	6000	6000

Table 3 Comparison of various data suppression methods.

From table 3 it is obtained that the dataset size value was quit low in super class substitution as compare to k-anonymity. While lowest value of dataset is same for both in item suppression method and super class substitution. In other words k-anonymity work has increase the dataset size value but to less extent.

V. CONCLUSION

Mining of individual information from the raw dataset leads to retrieve information. So protection of this information from the mining algorithm is required. Privacy preserving mining is applied to provide protection of that information in the dataset. In this paper three data suppression methods are compared on same set of data. Results are compared on various evaluation parameters and it was obtained that super class substitution was better than other methods K-anonymity and suppression.

REFERENCES

[1]. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases, pp. 487-499, 1994.
[2]. T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277-292, 2010.
[3]. F. Kamiran and T. Calders, "Classification with no Discrimination by Preferential Sampling," Proc. 19th Machine Learning Conf. Belgium and The Netherlands, pp 1-6, 2010.
[4]. Huhtala, Y., Karkkainen, J., Porkka, P., and Toivonen, H., (1999), TANE: An Efficient Algorithm for discovering Functional and Approximate Dependencies, The Computer Journal, V.42, No.20, pp.100-107.
[5]. Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, "Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases" in IEEE SYSTEMS JOURNAL, VOL. 7, NO. 3, SEPTEMBER 2013.
[6]. Shyue-liang Wang, Jenn-Shing Tsai and Been-Chian Chien, "Mining Approximate Dependencies Using Partitions on Similarity-relation-based Fuzzy Databases", IEEE International Conference on Systems, Man and Cybernetics (SMC) 1999.
[7]. Yao, H., Hamilton, H., and Butz, C., FD_Mine: Discovering Functional dependencies in a Database Using Equivalences, Canada, and IEEE ICDM 2002.
[8]. Wyss. C., Giannella, C., and Robertson, E. (2001), FastFDs: A Heuristic-Driven, Depth-First Algorithm for Mining Functional Dependencies from Relation Instances, Springer Berlin Heidelberg 2001.
[9]. Russell, Stuart J. and Norvig, Peter. Artificial Intelligence: A Modern Approach. Prentice Hall, 1995.
[10]. Mannila, H. (2000), Theoretical Frameworks for Data Mining, ACM SIGKDD Explorations, V.1, No.2, pp.30-32.
[11]. Stephane Lopes, Jean-Marc Petit, and Lotfi Lakhal, "Efficient Discovery of Functional Dependencies and Armstrong Relations", Springer 2000.
[12]. Thasneem M, S.Ramesh, Dr. T. Senthil Prakash. "An Effective Attack Analysis and Defense in Web Traffic Using Only Timing Information". International Journal of Scientific Research & Engineering Trends Volume 3, Issue 3, May-2017, ISSN (Online): 2395-566X, www.ijsret.com
[13]. Heikki Mannila and Kari-Jouko R¨aih¨a. Design by example: An application of Armstrong relations. Journal of Computer and System Sciences, 33(2):126{141, 1986.