ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Data Perturbation Techniques For Privacy Preservation In Association Rule Mining

[1]Dr.T.Ravi and [2]R. Prasanna Kumar

[1]Principal, Srinivasa Institute of Engineering and technology, Chennai, Tamil Nadu, India 600056
[2]Associate Professor, Department of Computer Science and Engineering, T.J.S Engineering College, Puduvoyal, Chennai, TamilNadu,India 601206

ARTICLE INFO

Article history:
Received 16 April 2015
Accepted 12 June 2015
Available online 1 July 2015

Keywords:
Association rule mining, Non synthetic, Perturbation, Privacy, Synthetic

ABSTRACT

In recent, data mining is becoming a popular analysis tool to extract knowledge from collection of large amount of data. The protection of the confidentiality of sensitive information in a database becomes a critical issue when releasing data to outside parties. Association analysis is a powerful and popular tool for discovering relationships hidden in large data sets. These process increases the legal responsibility of the parties. So, it is severe to reliably protect their data due to legal and customer concerns. In this paper, a review of the state-of-the-art methods of data perturbation techniques for privacy preservation is presented.

To Cite This Article: Dr.T.Ravi and R. Prasanna Kumar., Data Perturbation Techniques For Privacy Preservation In Association Rule Mining. ***Aust. J. Basic & Appl. Sci.,*** *9(20): 220-227, 2015*

INTRODUCTION

Data mining technology aims to find useful patterns from large amount of data. These patterns represent knowledge and are expressed in decision trees, clusters or association rules.

Recent advances in privacy preserving algorithms Agrawal (Agrawal, R., *et al*., 1993) put the sensitive and confidential information that resides in large data stores at risk. Providing solutions to privacy and security problems combines several techniques and mechanisms. An organization may have data at different sensitivity levels. This data is made available only to those with appropriate rights.

The knowledge discovered Verykios (Aris Gkoulalas–Divanis;Vassilios S. Verykios, 2010) by various data mining techniques may contain private information about individual. Disclosure of any private information may cause threat to security. For example, in banking database, it is useful to share information about account details but at the same time it is required to preserve holder's identity. Here individual privacy must be maintained. Some private information could be easily discovered by this kind of tools. Another example is Health care database which is used to analyze patience's behavior represented in terms of association rules. In health care database, instead of data related to individuals, the sensitive information or knowledge derived from data is required to be protected. The sharing of data and or knowledge may come at a cost to privacy, primarily due to two main reasons: 1.if the data refers to individuals then its disclosure can violate the privacy 2.if the data regards to business information.

Large numbers of research papers are available in this field, each tackling the problem of privacy preservation of data in different angle using different techniques. Most of the methods result in information misplacement and side-effects.

MATERIALS AND METHODS

The architecture for data perturbation for privacy preserving in Association Rule Mining (ARM) is given in the Figure.1

Corresponding Author: Dr.T.Ravi, Principal, Srinivasa Institute of Engineering and technology, Chennai, Tamil Nadu, India 600056

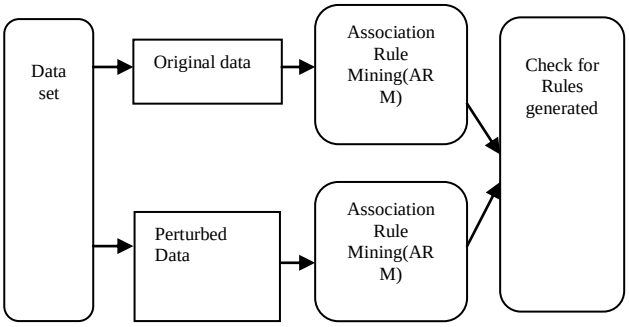

Fig.1: proposed architecture for privacy preservation

Association Rule Mining:

Association rules are an important class of regularities within data which have been extensively studied by the data mining community. The problem of mining association rules can be stated as follows: Given I = {i1 , i2 , ... , im } is a set of items, T = {t1, t2 , ... , tn} is a set of transactions, each of which contains items of the itemset I . Each transaction t_i is a set of items

An association rule is an implication of the form: $X \rightarrow Y$, where $X, Y \subset I$ and $X \cap Y = \emptyset$. X (or Y) is a set of items, called itemset. In the rule $X \rightarrow Y$, X is called the antecedent, Y is the consequent. It is obvious that the value of the antecedent implies the value of the consequent. The antecedent, also called the "left handside" of a rule, can consist either of a single item or of a whole set of items. This applies for the consequent, also called the "right hand side", as well. Often, a compromise has to be made between discovering all itemsets and computation time. Generally, only those item sets that fulfill a certain support requirement are taken into consideration. Support and confidence are the two most important quality measures for evaluating the interestingness of a rule.

The support of the rule $X \rightarrow Y$ is the percentage of transactions in T that contain $X \cap Y$. It determines how frequent the rule is applicable to the transaction set T. The support of a rule is represented by the formula transactions containing X which also contain Y. It is given by

$$\text{Support}(X \rightarrow Y) = \frac{X \cap Y}{N}$$

where $|X \cap Y|$ is the number of transactions that contain all the items of the rule and n is the total number of transactions.

Confidence is a very important measure to determine whether a rule is interesting or not. The process of mining association rules consists of two main steps. The first step is, identifying all the item sets contained in the data that are adequate for mining association rules. These combinations have to show at least a certain frequency and are thus called frequent item sets. The second step generates rules out of the discovered frequent item sets. All rules that has confidence greater than minimum confidence are regarded as interesting.

The confidence of a rule describes the percentage of

$$\text{Confidence } (X \rightarrow Y) = \frac{X \cap Y}{X}$$

Apriori Algorithm:

Apriori is a algorithm proposed by R. Agrawal et .al (2000) for mining frequent item sets for Boolean association rule. The name of algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties, as we shall see following. Apriori employs an iterative approach known as level wise search, where k item set are used to explore (k+1) item sets. There are two steps in each iteration. The first step generates a set of candidate item sets. Then, in the second step we count the occurrence of each candidate setting database and prunes all disqualified candidates.

Apriori uses two pruning technique, first on the bases of support count (should be greater than user specified support threshold) and second for an item set to be frequent , all its subset should be in last frequent item set The iterations begin with size 2 item sets and the size is incremented after each iteration.

The algorithm is based on the closure property of frequent item sets: if a set of items is frequent, then all its proper subsets are also frequent.

Algorithm_apriori(I, Min_sup, Min_con)
Initialize: k := 1, C1 = all the 1- item sets;
read the database to count the support of C1 to determine L1.
L1 := {frequent 1- item sets};
k:=2; //k represents the pass number//
while (Lk-1 ≠ ∅) do
begin
Ck := gen_candidate_itemsets with the given Lk-1
prune(Ck)
for all transactions t ∈T do
increment the count of all candidates in CK that are
contained in t;

Lk := All candidates in Ck with minimum support ;
k := k + 1;
end

Privacy Preserving In Association Rule Mining (Pparm):

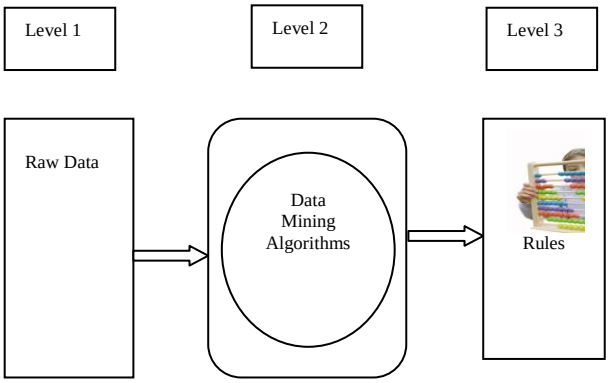

Fig. 2: Levels of PPARM

Level 1 applies different techniques to raw data for protecting the privacy of individuals, by preventing data miners from getting sensitive data or sensitive knowledge.

Clifton (Clifton, C. and D. Marks" 1996) presented a number of ideas to protect the privacy of individuals at Level 1. These include the following:

- Limiting access
- Fuzz the data
- Eliminate unnecessary data
- Augment the data
- Audit

Level 2, privacy-preserving techniques are embedded in the data mining which results in the masking of sensitive rules.

R.Agrawal *et al* (2000), applied techniques to impose constraints during the mining process to limit the number of rules to what they call "interesting rules".

Level 3, applies different techniques to the output of data mining algorithms or techniques for privacy preservation.

Output of data mining algorithms and techniques is shared. Privacy at this level provides more security since no raw data or databases are shared here.

Data Perturbation Techniques:

A random noise term is used for the perturbation of the original values of the sensitive data. Categorical as well as quantitative data are applicable for perturbation. Most applications of data perturbation, deals with numerical confidential variables. Data perturbation is non-reversible, unless the specific parameters are known hence, the user can be told exactly what type of perturbation was performed on the data. The work in Dr.T.Ravi *et al* (2014) represents the non-synthetic data perturbation in association rule mining where lost and camouflaged rules are identified for a dataset of varying confidence and a constant support.

Data perturbation approaches can be grouped into two main Categories

1. Probability distribution approach - approach replaces the data with another sample from the same (or estimated) distribution C.K. Liew *et al* (1985) or by the distribution itself E. Lefons (1983),

2. Value distortion approach - perturbs data elements or attributes directly by either additive noise, multiplicative noise, or some other randomization procedures. N.R. Adam *et al*. (1989).

Kargupta *et al*. (2003) questioned the use of random additive noise and pointed out that additive noise can be easily filtered out in many cases that may lead to compromising the privacy.

Two basic forms of multiplicative noise have been discussed in J.J. Kim (2003).

1. One is to multiply each data element by a random number that has a truncated Gaussian distribution with mean one and small variance.

2. Next method is to take a logarithmic transformation of the data first, add predefined multivariate Gaussian noise, and take the antilog of the noise-added data.

Additive and multiplicative perturbation usually deals with numeric data.

Perturbation for categorical data was initially considered in S. Warner (1965), where a randomized response method was developed for the purpose of data collection through interviews. The work in A. Evfimievski (2002) considered categorical data perturbation in the context of association rule mining.

Non-synthetic Data Perturbation (NDP):

Non-synthetic univariate:

Synthetic data results in information loss, because the perturbed data may be very "different" from the original values. In fact, information loss is based on the distance between the original and perturbed values. The additive perturbation is a

technique for privacy-preserving data mining in which noise is added to the data K. Muralidhar (2008) in order to mask the attribute values of records.Non- synthetic perturbed data maintains the mean vector and covariance matrix of the masked data to be exactly the same as the original data while offering a selectable degree of similarity between original and perturbed data. The noise added is sufficiently large so that individual record values cannot be recovered. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions algorithm Non synthetic(A,B,C,x_i,s_i) begin

let $x_i=\{x_1,x_2,x_3\ldots x_n\}$, $s_i=\{s_1,s_2,s_3\ldots s_n\}$ be set of confidential and non confidential values for all A do

compute masking of x_i with s_i using A,B,C $y_i=Ax_i+(1-A)Bs_i+sqrt((1-A^2)(1-B^2))C$ return y_i end

We begin with simple case of a single confidential variable X and single non confidential variable S. For simplicity and without loss of generality, we will assume the mean of X and S equal 0.The parameter A is the "similarity" parameter. When A=0,X and Y are most dissimilar. Thus, the parameter allows the data provider to control the level of similarity between the original and perturbed data. From K. Muralidhar *et al* (2008) we project the values can be perturbed from the formula as follows:

Where A is the similarity parameter is the correlation between x and s and C is normally distributed with mean 0 and unit variance When $0<A<1$, the values of the A represents the extent to which the perturbed value is a function of the original values. Large (small) values of A indicate the original values are significant (non-significant) component of the perturbed value. Conversely, as the value of 'A' approaches zero (one), the level of perturbation increases (decreases).

Non Synthetic Multivariate:

Let X (= X1,…, XK) represent a set of K confidential variables, let S (= S1, …, SL) represent a set of L non- confidential variables, and let Y (= Y1, …, Yk) represent the set of K masked variables. Let n represent the number of records in the data set. Let ΣXX , ΣSS , and ΣYY represent the covariance matrix of X, S, and Y, respectively. Let ΣXS and ΣYS represent the covariance between (X and S) and (Y and S), respectively. Let X , S , and Y be the mean vector of X, S and Y, respectively.

Let α be a matrix of size (K × K) representing the multipliers of X and let β be a matrix of size (K × L) representing the multipliers of S. let e_i be covariance matrix of noisealgorithm non synthetic (γ,xi,yi,α,β,e_i)

begin

let $x_i=\{x_1,x_2,x_3\ldots x_n\}$, $s_i=\{s_1,s_2,s_3\ldots s_n\}$ be set of confidential and non confidential values

for all x_i with s_i

compute $x_i\alpha^T$ and $s_i\beta^T$ with the covariance matrix of α and β

compute γ by the mean of x and s

compute noise e with mean 0 and some variance

$y=\gamma+x_i\alpha^T+s_i\beta^T+e_i$, i = 1, …, n

return y

end

Synthetic Data Perturbation (SDP):

It's been a conjecture that, rather than adding noise, multiplying noise might better protect the confidentiality. To further assure confidentiality swapping T. Dalenius (1982), K. Muralidhar (2006) of the perturbed values can be done.

Synthetic Multiplicative Perturbation:

Let $x_{i\ j}$ be the value for the i^{th} person's j^{th} characteristic, i = 1, 2,... n; j = 1, 2,... . p. We will denote the noise e_{i1}, e_{i2} . . . . e_{ip} corresponding to x_{i1}, x_{i2} . . . , x_{ip}. We let where ej is a random variable following a normal distribution with mean μ_j and variance σ_j algorithm synthetic(x,e)

begin

let $x_i=\{x_1,x_2,x_3\ldots x_n\}$ be confidential values

for all x_i

compute e by normal distribution with 0 mean and some variance

$y_{i\,j}=x_{i\,j}\,e_{i\,j}$

return y_{ij}

end

Synthetic Logarithmic Transformation Perturbation:

We define x_{ij}, V (Y) =Σ,

Let $y_{ij}=\log x_i+e_i$

$z_i=\text{Antilog}(y_i)$

where Σ is the variance/covariance matrix of variables x1, x2, . . . xp . We generate the random numbers following a multivariate normal distribution , where c is a positive number N (0, cΣ) between zero and one. We denote the noise variables e1, e2, . . . ep

algorithm logarithmic (x,e)

begin

let $x_i=\{x_1,x_2,x_3\ldots x_n\}$ be confidential values

for all x_i

generate e by normal distribution with 0 mean and some variance

compute y_{ij} by taking log for x_i and adding e

compute antilog for y_{ij}

return z_i

end

Random perturbation:

Johnson–lindenstrauss lemma:

Concerning low-distortion embeddings of points from high-dimensional into low-dimensional Euclidean space. The lemma states that

a small set of points in a high-dimensional space can be embedded into a space of much lower dimension in such a way that distances between the points are nearly preserved.

Given $0 < \varepsilon < 1$, a set X of m points in $\mathbf{R}^N$, and a number $n > 8 \ln(m) / \varepsilon^2$, there is a linear map $f : \mathbf{R}^N \rightarrow \mathbf{R}^n$ such that

$(1-\varepsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\varepsilon)\|u-v$

for all $u, v \in X$.

Synthetic Random Data Perturbation (SRDP):

Let $X \varepsilon R^{n \times m}$ where X is the dataset, m is the datapoints, n-dimensional space and R be k×n(k<n) random matrix where the elements are randomly distributed with mean 0 and some small variance (r)

algorithm random(R,X,r,k)

begin

Let X be a dataset with m points with n-dimensional spaces and R be a random matrix

Generate a random matrix k×n (k<n) of elements with mean=0 and some variance

Compute $y = \frac{1}{\sqrt{kvar\,(r)}} RX$

return y

end

Experimental Analysis:

In general, decreasing the support and confidence level of the frequently occurring item below minimum support and minimum confidence hides a rule. This can be achieved by masking the values of frequently occurring sensitive data items such that the item support goes below minimum support. We worked with Apriori association rule mining algorithm and examined their performance in order to analyses their impact on the original database. We worked with three datasets such as Measure of Birth and Death from U.S Department of Health & Human Services, Census Income, skin segmentation from UCI Machine Learning Repository. The dataset of Measures of Birth and Death consists of twenty one quantitative and six categorical attributes. skin segmentation dataset consists of 245057 instances and 4 attributes. Census Income consists of 48842 instances and 14 attributes. Experiments are conducted for 5000 transactions.

Figure 1 shows the total no of rules in perturbed approach for varying confidence of 20, 30, 40, 50, 60, 70, 80, 90, 100 and a constant support of 10, the scenario depicts that when ever there is a change in the similarity parameter the no of rules being generated varies accordingly but when A=0.7 and 0.9 the rules being generated are exactly same for a constant confidence of 50, this establishes a relationship between similarity parameter and no of rules being generated. This shows the performance metric of the no of rules can be changed in with similarity parameter, in case of non synthetic data perturbation.

Figure 2 shows the total no of rules in non-perturbed approach for varying confidence and a constant support of 10., the scenario depicts the no of rules being generate are said to be constant as the values are not perturbed, in case of non synthetic data perturbation.

Figure 3 shows the total no of lost rules approach for varying confidence of and a constant support of 10, the scenario shows when the similarity variable is near to the original values, the no of lost rules becomes null as the values are most similar, in case of non synthetic data perturbation.

Figure 4 shows the total no of camouflaged rules approach for varying confidence of and a constant support of 10, the scenario shows when the values are dissimilar the camouflaged rules gets increased, in case of non synthetic data perturbation.

Figure 5 shows the total no of rules approach for varying confidence and constant support of 10 for original and perturbed, the scenario depicts the original and perturbed generates most similar rules for the case of Non synthetic multivariate perturbation

Figure 6 shows the total no of rules approach for varying confidence and constant support of 10 for original and perturbed, the scenario depicts the original and the perturbed generates most similar rules for the case of synthetic multiplicative perturbation.

Figure 7 shows the total no of rules approach for varying confidence and constant support of 10 for original and perturbed, the scenario depicts the original and perturbed generates most similar rules for the case of logarithmic transformations.

Figure 8 shows the total no of rules approach for varying confidence and constant support of 10 for original and perturbed, the scenario depicts the original and perturbed generates most similar rules for the case of Random perturbation.

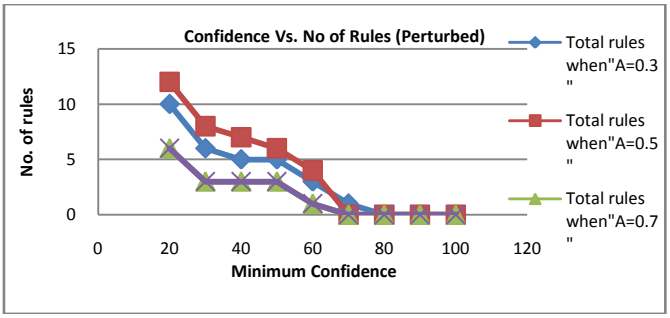

Fig. 3: Minimum Confidence vs. No. of Rule (Perturbed)

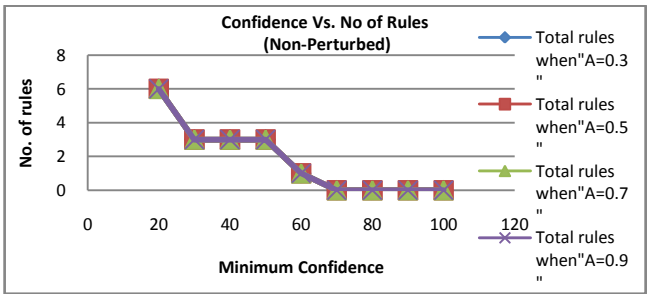

Fig. 4: Minimum Confidence vs. No. of Rules (non-pert)

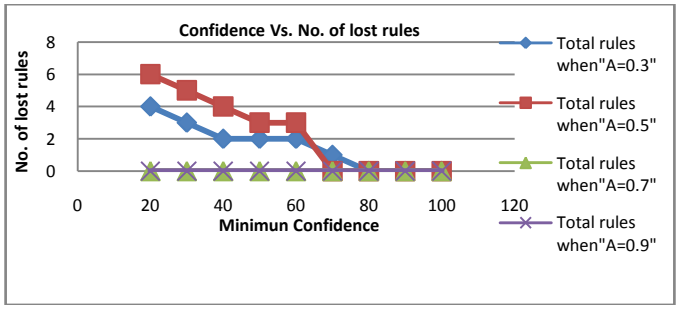

Fig. 5: Minimum Confidence vs. No. of lost rules

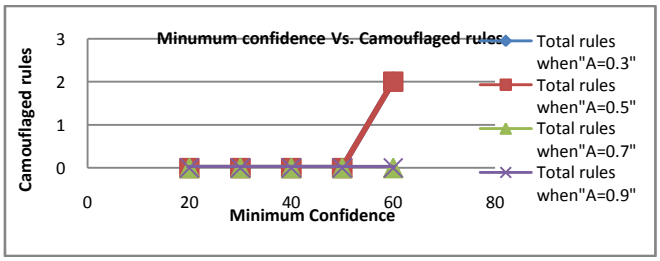

Fig. 6: Minimum Confidence vs. No of Camouflaged rules

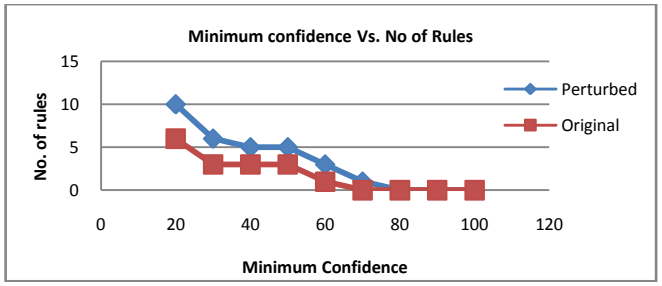

Fig. 7: Minimum Confidence vs. No of rules

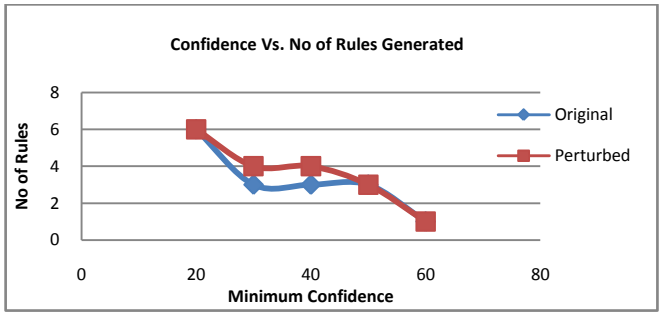

Fig. 8: Minimum Confidence Vs. No of rules

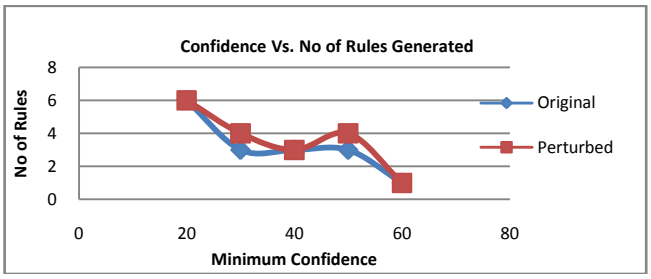

Fig. 9: Minimum Confidence Vs. No of rules

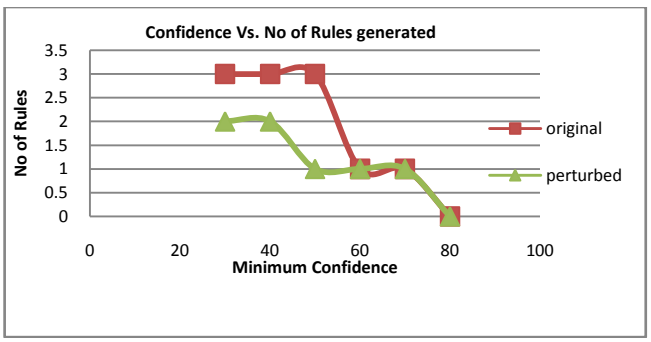

Fig. 10: Minimum Confidence Vs. No of rules

Conclusions and Future Directions:

We have proposed techniques for generating non synthetic and synthetic perturbed data for privacy preservation in Association Rule Mining. We used different datasets for generating perturbed datasets and these perturbed data were given as input to apriori algorithm and the association rules were generated. We conclude that higher similarity measures where perturbation interval is very less generates rules that are similar to the original ones. A lower similarity value which generates perturbed data with more perturbation intervals generates more false rules as in case of non synthetic data perturbation. Rules generated in synthetic multiplicative perturbation, logarithmic transformations andRandom perturbation produces most similar no of rules both in original and perturbed data values.

The algorithm for privacy preservation are limited to binary data, which can be extended to quantitative data, that can be implemented in the cloud environment preserving privacy in large

datasets. Hybrid techniques can be implemented to reduce the side effects of rule hiding. The measure of the rules are subjected to only support and confidence, different measures are to be constructed to make the privacy preservation to be more effective.

REFERENCES

Aris Gkoulalas–Divanis;Vassilios S. Verykios, 2010. "Association Rule Hiding For Data Mining" Springer, DOI 10.1007/978-1-4419-6569-1, Springer Science + Business Media, LLC

Evfimievski, A., R. Srikant, R. Agrawal and J. Gehrke, 2002. "Privacy Preserving Mining of Association Rules," Proc. Eighth ACMSIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD'02).

Clifton, C. and D. Marks" 1996, Security and privacy implications of data mining", In Workshop on Data Mining and Knowledge Discovery, pp: 15-19.

Liew, C.K., U.J. Choi and C.J. Liew, 1985. "A Data Distortion by Probability Distribution," ACM Trans. Database Systems (TODS), 10(3): 395-411.

Lefons, E., A. Silvestri and F. Tangorra, 1983. "An Analytic Approach to Statistical Databases," Proc. Ninth Int'l Conf. Very Large Data Bases, pp: 260-274.

Kargupta, H., S. Datta, Q. Wang, and K. Sivakumar, 2003. "On the Privacy Preserving Properties of Random Data Perturbation Techniques," Proc. IEEE Int'l Conf. Data Mining, pp: 99-106.

Kim, J.J. and W.E. Winkler, 2003. "Multiplicative Noise for Masking Continuous Data," Technical Report Statistics #2003-01, Statistical Research Division, US Bureau of the Census, Washington D.C.

Muralidhar, K. and R. Sarathy, 2006. "Data Shuffling- A New Masking Approach for Numerical Data," Management Science, 52(5): 658-670.

Muralidhar, K. and R. Sarathy, 2008. "Generating Sufficiency- based Non-Synthetic Perturbed Data," Management Science, Transaction on Data Privacy, pp: 17-33.

Adam, N.R. and J.C. Worthmann, 1989. "Security-Control Methods for Statistical Databases: A Comparative Study," ACM Computing Surveys (CSUR), 21(4): 515-556.

Agrawal, R., T. Imieliński, A. Swami, 1993. "Mining association rules between sets of items in large databases". "Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93". pp: 207.

Agrawal, R. and R. Srikant, 2000. "Privacy-Preserving Data Mining", SIGMOD, pp: 161-172.

Warner, S., 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," J. Am. Statistical Assoc., 60: 63-69.

Dalenius, T. and S.P. Resiss, 1982. Data Swapping: A technique for Disclosure Control", Journal of Statistical Planning and Inference, 6: 73-85.

Dr.T.Ravi, R. Prasanna Kumar, Komal Kumar, 2014."A Non Synthetic Data Perturbation Technique for Privacy preservation in Association Rule Mining", In the International Journal of Applied Engineering Research, 9(24): 24311-24320