International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 6, Number 1 (2016), pp. 1-10

http://www. irphouse.com

An Efficient Statistical Method for Providing Privacy and Security in Big Data

P.Shobha Rani[1], Vigneswari D[2]

[1]Associate Professor, Department of Computer Science and Engineering, R.M.D Engineering College, Chennai, India.
[2]PG Scholar, Department of Computer Science and Engineering, R.M.D Engineering College, Chennai, India.

Abstract

The growing need for computing on bigdata is getting higher, the three basic dimensions of big data are (referred as "3V" challenges) high volume, variety and velocity. The other upcoming challenge in the area of bigdata is Veracity, which means the trustworthiness of the data that is how secure the data is received, stored, processed and transmitted. Hence this Veracity is becoming a new dimension in the bigdata era. 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.

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

1. INTRODUCTION

One of the most important dimension in big data era is the Veracity – whether the data is protected securely and managed with privacy. Generally providing security to big data is a tedious process because of its large volume. Consequently the privacy and security requirements in big data is very high. 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 [1] algorithms 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 [11] 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.

2. ASSOCIATION RULE MINING

Association rules Srikant [9] 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∩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

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

2.1 Apriori Algorithm

Apriori is a algorithm proposed by R. Agrawal and R Srikant [1] 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 Agrawal [2] 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

In this paper we have proposed various techniques for synthetic data perturbation with Association rule mining of datasets. The technique follows as

i. Generating Perturbed Dataset from Original Dataset using synthetic data perturbation techniques.
ii. Mapping for Association rules of minimum confidence and support
iii. Checking for Rules Generated.

3. PRIVACY PRESERVING IN ASSOCIATION RULE MINING (PPARM)

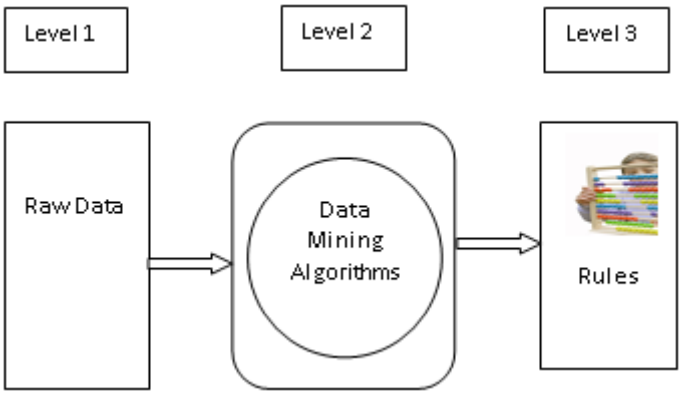

Figure 3.1 Levels of PPARM

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 [3] 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.
Srikant [4], 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.

4. SYNTHETIC DATA PERTURBATION TECHNIQUES

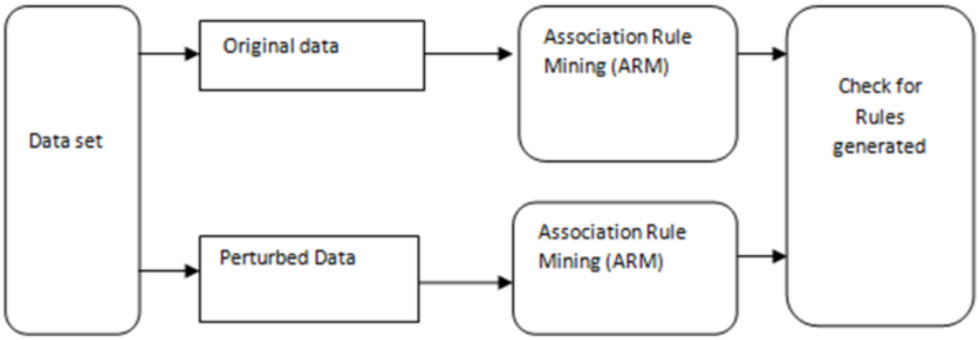

Fig 4.1 System Architecture

4.1 Multiplicative Perturbation

It's been a conjecture that, rather than adding noise, multiplying noise might better protect the confidentiality. To further assure confidentiality swapping Dalenius [5[10] of the perturbed values can be done
Two approaches for multiplicative noise kim [6]:

1. Generating random numbers which have mean one and a small variance, and multiplying the original data by the noise
2. To take logarithmic transformations [7] of the original data, generating a random number which follows mean zero and some variance, add this noise to the log value computed as above and take antilog.

4.1.1 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

4.1.2 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

4.2 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 $(1-\varepsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\varepsilon)\|u-v\|^2$ for all $u, v \in X$.

4.2.1 Synthetic Random perturbation

Kargupta[8] 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

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

return y

end

5. 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 Census Income dataset UCI Machine Learning Repository. Census Income consists of 48842 instances and 14 attributes. Experiments are conducted for 5000 transactions.

Figure 1 shows the total no of rules approach for varying confidence and constant support of 10 for original and perturbed, random noise being generated depends upon the mean and the variance for the case of Gaussian distribution, the scenario depicts the original and the perturbed values generates most similar rules for the case of synthetic multiplicative perturbation.

Figure 2 shows the total no of rules approach for varying confidence and constant support of 10 for original and perturbed, random noise being generated depends upon the mean and the variance for the case of Gaussian distribution, the scenario depicts the original and perturbed generates most similar rules for the case of logarithmic transformations.

Figure 3 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 wide deviation in rules for the case of random perturbation.

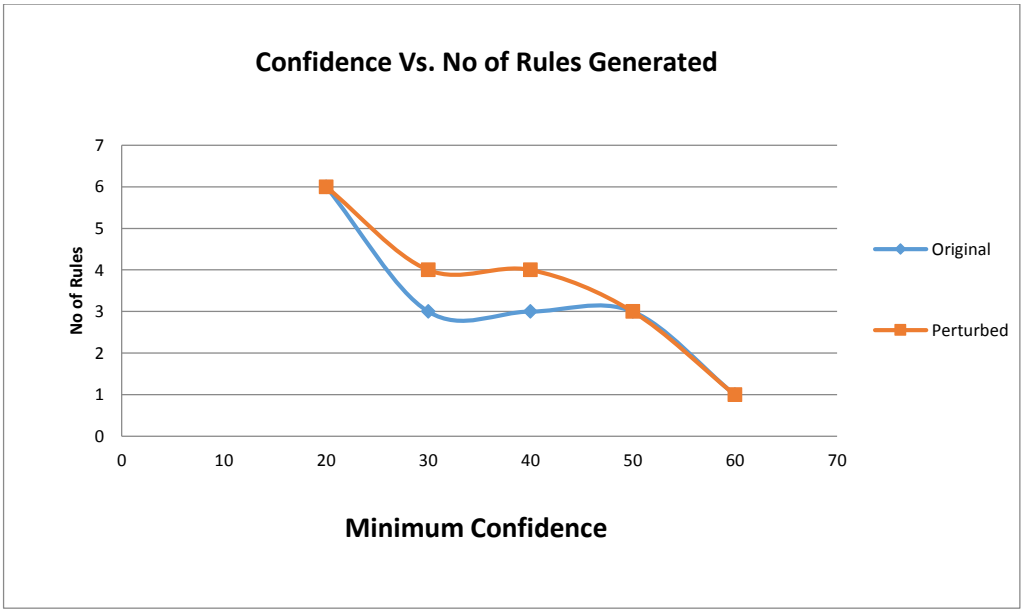

Figure 1. Minimum Confidence Vs. No of rules (synthetic multiplicative)

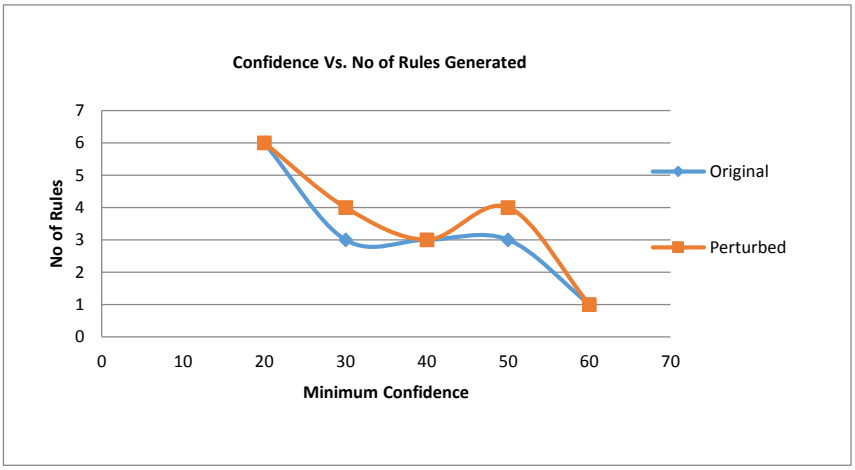

Figure 2. Minimum Confidence Vs. No of rules (logarithmic transformations**)**

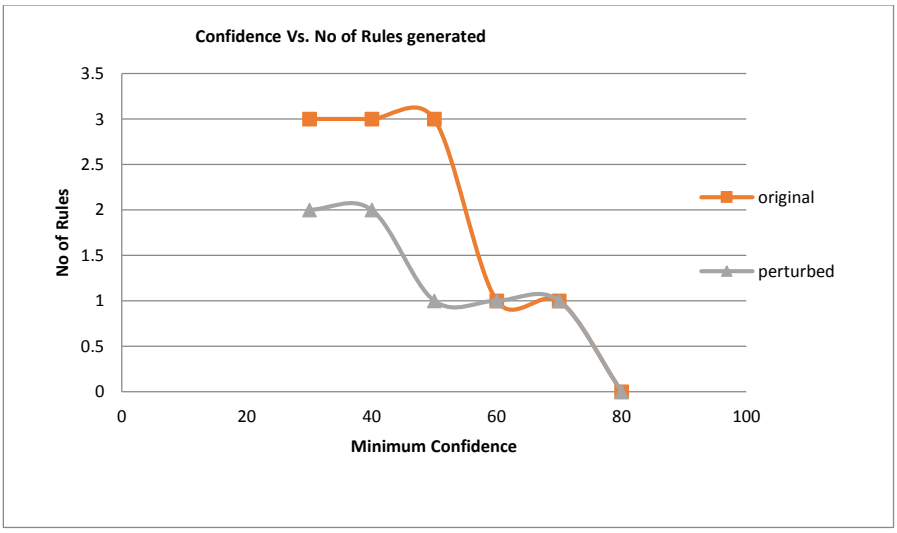

Figure 3. Minimum Confidence Vs. No of rules (random perturbation)

6. CONCLUSIONS AND FUTURE DIRECTIONS

We have proposed techniques for generating synthetic perturbed data for privacy preservation in Association Rule Mining. We used different synthetic perturbation techniques for generating perturbed datasets and these perturbed data were given as input to apriori algorithm and the association rules were generated. Rules generated in synthetic multiplicative perturbation, logarithmic transformations and produces most similar no of rules both in original and perturbed data values and thus concludes the effectiveness of the privacy preservation in association rule mining. But the Random

perturbation results in wide deviation in no of rules generated 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

[1] Agrawal D. Aggarwal C. C. On the Design and Quantification of Privacy-Preserving Data Mining Algorithms. ACM PODS Conference, 2002.

[2] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. "Mining frequent item sets for boolean association rules". In Proceedings of International Conference on Management of Data, pp. 216.

[3] C. Clifton and D. Marks" 1996, Security and privacy implications of data mining", In Workshop on Data Mining and Knowledge Discovery, pages 15–19.

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

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

[6] J.J. Kim 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.

[7] Kim,J., 1986,"A method for limiting disclosure in microdata based on random noise and transformation", American Statistical Associationn 1986 Proceedings of the Section on Survey Reaearch Methods, 370-374.

[8] H. Kargupta, 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.

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

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

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

[12] A. Cavoukian and J. Jonas, "Privacy by Design in the Age of Big Data," Office of the Information and Privacy Commissioner, 2012.

[13] Cheng Hongbing; Rong Chunming; Hwang Kai; Wang Weihong; Li Yanyan, "Secure big data storage and sharing scheme for cloud tenants," in *Communications, China* , vol.12, no.6, pp.106-115, June 2015 doi: 10.1109/CC.2015.7122469

[14] M. Li et al., "Toward Privacy-Assured and Searchable Cloud Data Storage Services," IEEE Network, vol. 27, no. 4, 2013, pp. 1–10.

[15] S. Liu, "Exploring the Future of Computing," IT Professional, vol. 15, no. 1, 2013, pp. 2–3.

[16] Matturdi, B.; Zhou Xianwei; Li Shuai; Lin Fuhong, "Big Data security and privacy: A review," in *Communications, China*, vol.11, no.14, pp.135-145, Supplement 2014 doi:10.1109/CC.2014.7085614