



JOURNAL OF APPLIED SCIENCES RESEARCH

ISSN: 1819-544X EISSN: 1816-157X

JOURNAL home page: <http://www.aensiweb.com/JASR>

2015 June; 11(10); pages 55-59.

Published Online 10 June 2015.

Research Article

Synthetic Data Perturbation Techniques for Privacy Preservation in Association Rule Mining

¹Dr. T. Ravi, ²R. Prasanna Kumar, ³Komal Kumar. N., ⁴G. Ragu

¹Principal, Srinivasa Institute of Engg. and Tech, Chennai.

²Research Scholar, Anna University, Chennai.

³Assistant Professor, T.J.S. Engineering College, Chennai.

⁴Assistant professor, T.J.S. Engineering College, Chennai.

Received: 6 March 2015; Revised: 25 April 2015; Accepted: 29 May 2015

© 2015 AENSI PUBLISHER All rights reserved

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.

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

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 [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 patient'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

Corresponding Author: Dr. T. Ravi, Principal, Srinivasa Institute of Engg. and Tech, Chennai.

E-mail: travi675@gmail.com

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 $I = \{i_1, i_2, \dots, i_m\}$ is a set of items, $T = \{t_1, t_2, \dots, t_n\}$ 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 hand side" 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

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

2.1 Apriori Algorithm:

Apriori is an 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($I, \text{Min_sup}, \text{Min_con}$)

Initialize: $k := 1$, $C_1 =$ all the 1- item sets;
read the database to count the support of C_1 to determine L_1 .

$L_1 :=$ {frequent 1- item sets};

$k:=2$; //k represents the pass number//

while ($L_{k-1} \neq \emptyset$) do

begin

$C_k :=$ gen_candidate_itemsets with the given L_{k-1}
prune(C_k)

for all transactions $t \in T$ do

increment the count of all candidates in C_k that are contained in t ;

$L_k :=$ All candidates in C_k 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):

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.

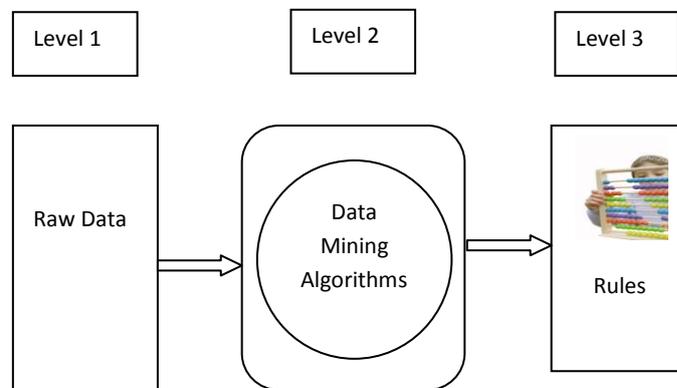


Fig. 3.1: Levels of PPARM.

4. Synthetic Data Perturbation Techniques:

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

Algorithm synthetic(x,e)

begin

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

for all x_i

compute e by normal distribution with 0 mean and some variance

$y_{ij} = x_{ij} e_{ij}$

return y_{ij}

end

4.1.2 Synthetic logarithmic transformation perturbation:

We define x_{ij} , $V(Y) = \Sigma$,

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

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

where Σ is the variance/covariance matrix of variables x_1, x_2, \dots, x_p . We generate the random numbers following a multivariate normal distribution, where c is a positive number $N(0, c\Sigma)$ between zero and one. We denote the noise variables e_1, e_2, \dots, e_p

Algorithm logarithmic(x,e)

begin

let $x_i = \{x_1, x_2, x_3, \dots, x_n\}$ 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 $0 < \epsilon < 1$, a set X of m points in \mathbf{R}^N , and a number $n > 8 \ln(m) / \epsilon^2$, there is a linear map $f : \mathbf{R}^N \rightarrow \mathbf{R}^n$ such that $(1 - \epsilon) \|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon) \|u - v\|^2$ for all $u, v \in X$.

4.2.1 Synthetic Random perturbation:

Kargupta[8] Let $X \in \mathbf{R}^{n \times m}$ where X is the dataset, m is the datapoints, n -dimensional space and R be $k \times n$ ($k < n$) 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 $k \times n$ ($k < n$) of elements with mean=0 and some variance

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 analyse 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 most similar rules for the case of 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

Random perturbation 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.

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.



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

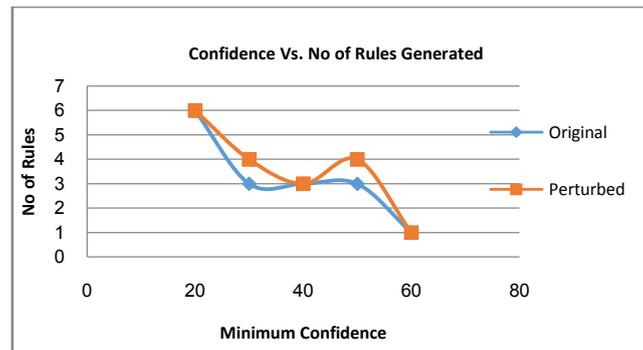


Fig. 2: Minimum Confidence Vs. No of rules (logarithmic transformations).



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

References

1. Agrawal, D., C.C. Aggarwal, 2002. On the Design and Quantification of Privacy-Preserving Data Mining Algorithms. ACM PODS Conference.
2. Agrawal, R., T. Imielinski and A.N. Swami, 1993. "Mining frequent item sets for boolean association rules". In Proceedings of International Conference on Management of Data, pp: 216.
3. Clifton C. and D. Marks, 1996. Security and privacy implications of data mining", In Workshop on Data Mining and Knowledge Discovery, pages 15-19.
4. Agrawal, R. and R. Srikant, 2000. "Privacy-Preserving Data Mining", SIGMOD, pp: 161-172.
5. Dalenius, T. and S.P. Resiss, 1982. "Data Swapping: A technique for Disclosure Control", Journal of Statistical Planning and Inference, 6: 73-85.
6. 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.
7. Kim, J., 1986. "A method for limiting disclosure in microdata based on random noise and transformation", American Statistical Association 1986 Proceedings of the Section on Survey Research Methods, 370-374.
8. 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.
9. Evfimievski, R. Srikant, R. Agrawal and J. Gehrke, 2002. "Privacy Preserving Mining of Association Rules," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD'02).
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, 2010. "Association Rule Hiding For Data Mining" Springer, DOI 10.1007/978-1-4419-6569-1, Springer Science + Business Media, LLC.