Privacy Preserving Association Rule Mining Using Perturbation Technique

Tobi Deborah Omoyele
Department of Computer Science
University of Ibadan, Ibadan, Nigeria
Omoyeletobi@yahoo.com

Solomon Olalekan Akinola
Department of Computer Science
University of Ibadan, Ibadan, Nigeria
solom202@yahoo.co.uk

ABSTRACT
The information age has enabled organizations to gather large volumes of data. However, the usefulness of this data is negligible if "meaningful information" cannot be extracted from it. Data mining answers this need. The problem of privacy-preserving data mining has become important recently because of the increasing ability to store personal data and the sophistication of data mining algorithms to leverage information. Many researches have been done in this field but few with quantitative data had drawbacks of high number of rules generated and few number of item hidden. This study, proposes a perturbation association rules hiding algorithm for privacy of quantitative data to provide a better algorithm for preserving quantitative data. In hiding of rules, the noise associated with each item was calculated. The noise was used to calculate the support and confidence of rules which were then compared with minimum support and confidence. Item whose support/confidence is less than or equal to minimum support or confidence would be hidden. Experimental result shows that the algorithm hides more rules than the existing works.

CCS Concepts
• Information systems → Information systems applications → Data mining → Association rules • Security and privacy → Human and societal aspects of security and privacy →Privacy protection

Keywords: Data Perturbation; Data Mining; Data Privacy; Randomization; Association rule mining; Summarization; Suppression.

1. INTRODUCTION
The information age has enabled many organizations to gather large volumes of data. However, the usefulness of this data is negligible if "meaningful information" or "knowledge" cannot be extracted from it. Data mining attempts to answer this need. Data mining is an iterative process within which progress is defined by discovery, prediction or classification of data through either automatic or manual methods. It is most useful in scenarios in which there are no predetermined notions about

what will constitute an interesting outcome and it also involve the search for new, valuable and nontrivial information in large volumes of data but the potential danger is privacy concerns of sensitive data [1]. The objective of privacy preserving data mining is to hide sensitive information so that they cannot be discovered through data mining techniques. Privacy preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited disclosure [2]. Most traditional data mining techniques analyze and model the dataset statistically, while privacy preservation is primarily concerned with protecting against disclosure of individual data records.

The term "privacy preserving data mining" was introduced in papers Rakesh and Ramakrishna [3] and Lindell and Pinkas [4]. These papers considered two fundamental problems of privacy preservation in data mining, privacy preserving in data collection and mining a dataset partitioned across several private enterprises. Rakesh and Ramakrishna [3] devised a randomization algorithm that allows a large number of users to contribute their private records for efficient centralized data mining while limiting the disclosure of their values, Lindell and Pinkas [4] invented a cryptographic protocol for decision tree construction over a dataset horizontally partitioned between two parties.

A number of techniques such as randomization, suppression, summarization, association rule, perturbation, cryptography and k-anonymity have been suggested in recent years according to survey carried out by Alexandre [5] in order to perform privacy-preserving data mining.

Ila [6] proposed an algorithm for privacy preservation of data mining in which he assumed that only sensitive data items can be found in the database. The proposed algorithm modified data in the database such that sensitive item can either be at the left hand side or right hand side of the rule and cannot be inferred through association rule mining algorithms. For the algorithm to hide sensitive item either the support or confidence is decreased to be smaller than pre-specified minimum support and minimum confidence.

Most of the studies carried out in this research are concentrated on hiding binary database which are only concerned with the presence or absence of item but in reality most real applications consists of quantitative values. For instance, many people have the problem of sugar, but this does not mean that one is sick or not, the only criterion that can be use to for determine the illness is the surplus/deficiency in sugar's quantity.

The problem of mining quantitative association rule was first introduced by Evfimievski *et al* [7]. Two works have been done in the field of hiding fuzzy association rule in quantitative data according to Manoj and Joshi [8] and Berberoglu [9]. However, little or no work has been done in the aspect of perturbation of association rules for quantitative data using the concept of noise.
Manoj and Joshi [8] proposed a hiding algorithm that integrates the fuzzy set concepts and Apriori mining algorithm to find

CoRI'16, Sept 7–9, 2016, Ibadan, Nigeria.

useful fuzzy association rules from a quantitative database and then hide them using privacy preserving technique. Unlike previous approaches which mainly deals with association rules in binary database, the approach deals with hiding the association rules in quantitative database. Numerical experiments were performed to measure the performance of the algorithm according to three criteria: the number of rules hidden, side effects and database effects of the algorithm. However, the algorithm only deals with association rule in small dataset in which the algorithm generated much rules but the algorithm is not efficient since a large part of the rules generated resulted into lost rules.

In this paper, we attempt to present a perturbation association rules hiding algorithm for privacy preservation of quantitative data with few rules and hides higher percentage of the rules by decreasing the support of the rule. The support of rule $A \rightarrow B$ is decreased by decreasing the support of A. This is achieved by decreasing the support value of either A or B of the rule until either the support or confidence is below the minimum support or minimum confidence value respectively.

The rest of this paper is organized as follows. Privacy preserving quantitative association rule hiding was defined in Section II, our approach to hide useful association rules in quantitative data was presented in Section III, the perturbation association rule hiding process was presented in Section IV, and the proposed algorithm related example was presented in Section V. Section VI includes the conclusion and future works.

2. PROBLEM STATEMENT

The objective of privacy preserving data mining is to hide certain information so that they cannot be discovered through data mining techniques. There have been two broad approaches for privacy preserving data mining. The first approach, called output privacy, is to alter the data before delivery to data miner so that real data is obscured and mining result will not disclose certain privacy. For example, perturbation, blocking, merging, swapping and sampling are some methods that have been proposed for this type of output privacy. The second approach, called input privacy, is to manipulate the data using data distribution methods in which the privacy of the data is protected before releasing to the user. In this approach, mining result is not affected or minimally affected. Almost all of studies that have been proposed in this research area concentrated on hiding Boolean association rules which are concerned only with whether an item is present in a transaction or not, without considering its quantity.

However, transactions with quantitative values are commonly found in real world application. For example, in a patient's blood test, many attributes could be found. In addition, attribute's quantity instead of just presence/absence of the attribute in blood is more important for determination of the illness. Furthermore, many people have the problem of sugar but this does not mean that one is sick or not, the only criterion to used to determine the illness is the surplus/deficiency in sugar 's quantity.

Some works have been done to discover association rules from quantitative data which generate set of rules but hides less than 30% of the rules generated. This paper proposes a privacy preserving data mining algorithm (that uses the output privacy approach in which the data is preserved before it will be released to the miner) that improves on the association rule hiding algorithms for quantitative data by combining the perturbation method for privacy and association rule mining technique for privacy. Furthermore, the paper addresses the challenges of applying privacy algorithm to only Boolean data by applying the algorithm to quantitative data and not just the presence or absence of data.

3. PROPOSED ALGORITHM

In order to hide an association rule, A→B we can either decrease its support to be smaller than minimum support value or its confidence to be smaller than its minimum confidence value. To decrease the confidence of a rule, two strategies can be used. The first one is to increase the support count of A i.e. LHS of the rule, but not support count of A→B. The second one is to decrease the support count of A→B. For the second case, if we only decrease the support of B, the right hand side of the rule, it would reduce the confidence faster than simply reducing the support of A∪B. Based on these two strategies, we propose a privacy preserving data mining algorithm for hiding sensitive quantitative data items called perturbation of association rule for quantitative data using the concept of noise. This algorithm first calculates the value of noise for each data items and the column with the highest noise value form the rule with the sensitive column. Secondly, the algorithm find useful association rule that consist of only one item on both sides of the rule and then hide them using privacy preserving technique. For hiding purpose, the algorithm tries to decrease the support of rule A→B by decreasing the support count of itemset AB until either support or confidence value of the rule goes below minimum support or minimum confidence value respectively.

Input:
(1)A source database D,
(2) Minimum support threshold
(3) Minimum confidence threshold
Output: A transformed database D where rules containing A on LHS (Left Hand Side) or B at the RHS (Right Hand Side) will be hidden.

Algorithm

1. Transform the rough dataset(Rd) into original database(D) as Rd→D
2. For each itemset XεD, for $x_1, \ldots\ldots\ldots\ldots\ldots, x_m \varepsilon X$. Calculate the value of the noisy attribute V such that $V = \frac{X+X^i}{2\sqrt{N}}$
3. Calculate the sumCount of each noisy attribute V_j for $j = 1\ to\ m$
4. If sumCount(V_j) = Maximum Value such that k = Maximum Value
   THEN
   1-itemset L_1= { $k_1, k_2, \ldots, k_j$}
5. EXIT sumCount
6. Sensitive column = {$S_1, \ldots\ldots, S_m$ } such that mεN
   1-itemset $L_2 = \{S_1, \ldots\ldots, S_m\}$
7. Find 2-itemset from the union of L_1 and L_2
8. Find c ={ rules from itemset X}
   //for X= {k_i, S_i} the possible rules are $k_1 \rightarrow S_1$
   $k_2 \rightarrow S_2$
   $k_3 \rightarrow S_3$
   $k_4 \rightarrow S_4$
   $\ldots \rightarrow \cdots$
   $k_j \rightarrow S_m$
9. Compute the support of rule R_x
   $Sup\big((LHS, RHS)\big) = \frac{\min(LHS,RHS)}{n} = \frac{\min(L_1, L_2)}{n}$

10. Compute the confidence of the rule R_x
$Conf\big((LHS, RHS)\big) = \frac{\text{Supp}(LHS,RHS)}{Supp(LHS)} = \frac{\min(LHS,RHS)}{Supp(LHS)} = \frac{\min(L_1, L_2)}{Supp(L_1)}$
11. if Sup(LHS,RHS) ≤ MST or conf(LHS,RHS) ≤ MCT
X_{bj}= X+V_{bj}
Else
$X_{bj} = X_{bj}$
12. go to line 8
13. select the next rule
14. repeat // line 9,10 and 11 on each new rule
15. if R_x is empty
THEN
16. transform the updated database D and output the updated D

3.1 Explanation to Abbreviations in the Algorithm

LHS (left hand side): this is the left hand side of the rule
RHS (right hand side): this is the right hand side of the rule generated
MST (minimum support threshold): this is the support specified by the user of the algorithm
MCT (minimum confidence threshold): this is the confidence threshold specified by the user
min = minimum
Sup/Supp = support

4. STEPS TO THE PROPOSED ALGORITHM

n = total number of transaction data
m= total number of attributes (items)
D= i^{th} attribute 1 ≤ i ≤ n
I_j= j^{th} attribute 1 ≤ j ≤ m
V= noise
X= original data
X^i= inverse of original data
MCT= minimum confidence threshold
MST= minimum support threshold
V_j= each noisy attribute 1≤ k ≤ i

STEP 1: for each transaction data D, i =1 to n, and for each attribute (item) j= 1 to m, transform the quantitative value into a noisy quantitative attribute value using the randomly generated formula $V = \frac{X+X^i}{2\sqrt{N}}$
For X = $x_1, \dots\dots\dots\dots\dots\dots, x_m$
X^i= $x_1^i, \dots\dots\dots\dots\dots\dots, x_m^i$
STEP 2: calculate the sumcount of each noisy attribute V_j on the transaction data as
$count_{jk} = \sum_{i=1}^{n} V_j$

STEP 3: for each noisy attribute V_j 1≤j≤m and 1≤k≤m check for the $count_j$ that has the maximum value. If $count_j$ satisfies the above condition then the column whose $count_j$ has the maximum value is put in the set of 1- itemset which form the left hand side (LHS) of the rule
i.e L_1={ V_j: $count_{jk}$ *has the maximum count value*}

STEP 4: join the 1-itemset L_1 to the sensitive column C_i in a way similar to that of apriori algorithm to form a 2- itemset. The 2-itemset is use to find the useful association rule by placing L_1 at the LHS and C_i at the RHS similar to that of apriori algorithm
STEP 5 (a): in order to hide sensitive rule, calculate the support and confidence of each rule

$$Sup\big((LHS, RHS)\big) = \frac{\min(LHS, RHS)}{n}$$

and

$$Conf\big((LHS, RHS)\big) = \frac{\text{Supp}(LHS, RHS)}{Supp(LHS)} = \frac{\min(LHS, RHS)}{Supp(LHS)}$$

STEP 5(b): A rule is hidden if
$Sup(LHS, RHS) \leq MST$
Or
$Conf(LHS, RHS) \leq MCT$
if Sup(LHS,RHS) ≤ MST or conf(LHS,RHS) ≤ MCT
X_{bj}= X+V_{bj}
Else
$X_{bj} = X_{bj}$

5. ILLUSTRATION OF THE ALGORITHM

In this section, we give an example to illustrate how the proposed algorithm can be used.
Let the original data be depicted below:

	A	B	C	D
T1	10	5	8	3
T2	3	11	6	14
T3	6	3	9	13
T4	7	5	8	12
T5	11	4	7	10

T1, T2, T3, T4, T5 are transactions and A, B, C, D are the attributes of the data in the relational table
Given:
Assuming the sensitive column to hide is column B
Minimum support threshold (MST) = 44% = 0.44
Minimum confidence threshold (MCT) = 75% = 0.75

STEP 1: for each transaction data D, i =1 to n, and for each attribute (item) j= 1 to m, transform the quantitative value into a noisy quantitative attribute value using the randomly generated formula $V = \frac{X+X^i}{2\sqrt{N}}$
For X = $x_1 \dots\dots\dots\dots\dots\dots x_m$
X^i= $x_1^i \dots\dots\dots\dots\dots\dots x_m^i$
For T1 x_{T1A}= 10
For T1 x_{T1A}^i=1/10
N= 5 i.e number of transactions in the database.

	A	A_I	B	B_I	C	C_I	D	D_I
T1	10	2.26	5	1.16	8	1.82	3	0.75
T2	3	0.75	11	2.48	6	1.38	14	3.15
T3	6	1.38	3	0.75	9	2.04	13	2.93
T4	7	1.59	5	1.16	8	1.82	12	2.70
T5	11	2.48	4	0.95	7	1.59	10	2.26

STEP 2: calculate the count of each noisy attribute Q_k on the transaction data as

$$count_{jk} = \sum_{i=1}^{n} V_j$$

	A	A_J	B	B_J	C	C_J	D	D_J
T1	10	2.26	5	1.16	8	1.82	3	0.75
T2	3	0.75	11	2.48	6	1.38	14	3.15
T3	6	1.38	3	0.75	9	2.04	13	2.93
T4	7	1.59	5	1.16	8	1.82	12	2.70
T5	11	2.48	4	0.95	7	1.59	10	2.26
count		8.46		6.5		8.65		11.79

STEP 3: for each noisy attribute V_j 1≤j≤m and 1≤k≤m check for the $count_j$ that has the maximum value. if $count_j$ satisfies the above condition then the column whose $count_j$ has the maximum count is put in the set of 1- itemset which form the left hand side (LHS) of the rule

i.e $L_1 = \{ V_j : count_{jk}\ has\ the\ maximum\ count\ value\}$

In this case

$L_1 = \{D_1, D_2, D_3, D_4, D_5\}$

STEP 4: join the 1-itemset L_1 to the sensitive column $C_i$in a way similar to that of apriori algorithm to form a 2- itemset. The 2-itemmset is use to find the useful association rule by L_1 at the LHS and C_i at the RHS similar to that of apriori algorithm

$D_1 \rightarrow B_1$

$D_2 \rightarrow B_2$

$D_3 \rightarrow B_3$

$D_4 \rightarrow B_4$

$D_5 \rightarrow B_5$

STEP 5 (a): in order to hide sensitive rule, calculate the support and confidence of each rule

$$Sup\big((LHS, RHS)\big) = \frac{\min(LHS, RHS)}{n}$$

and

$$Conf\big((LHS, RHS)\big) = \frac{\text{Supp}(LHS, RHS)}{Supp(LHS)} = \frac{\min(LHS, RHS)}{Supp(LHS)}$$

STEP 5(b): A rule is hidden if

$Sup(LHS, RHS) \leq MST$

Or

$Conf(LHS, RHS) \leq MCT$

if Sup(LHS,RHS) ≤ MST or conf(LHS,RHS) ≤ MCT

$X_{bj} = X_{bj} + V_{bj}$

Else

$X_{bj} = X_{bj}$

- $Sup(D_1 \rightarrow B_1) = \frac{\min(D_1, B_1)}{n} = \frac{\min(0.75, 1.16)}{5} = \frac{0.75}{5} = 0.15$

$$Conf(D_1 \rightarrow B_1) = \frac{\text{Supp}(D_1 \rightarrow B_1)}{Supp(D_1)} = \frac{\min(D_1, B_1)}{Supp(D_1)} = \frac{\min(0.75, 1.16)}{0.75} = \frac{0.75}{0.75} = 1$$

Since Sup($D_1 \rightarrow B_1$) < MST but conf($A_1 \rightarrow B_1$) > MCT ,hence b_1 is hidden

i.e

$X_{b1} = X_{b1} + V_{b1}$

$X_{b1} = 5 + 1.16 = 6.16 \approx 6$ to nearest whole number

- $Sup(D_2 \rightarrow B_2) = \frac{\min(D_2, B_2)}{n} = \frac{\min(3.15, 2.48)}{5} =$ 2.48/5=0.49

$$Conf(D_2 \rightarrow B_2) = \frac{\text{Supp}(D_2 \rightarrow B_2)}{Supp(D_2)} = \frac{\min(D_2, B_2)}{Supp(D_2)} = \frac{\min(3.15, 2.48)}{3.15} = \frac{2.48}{3.15} = 0.79$$

Since Sup($D_2 \rightarrow B_2$) > MST and conf($D_2 \rightarrow B_2$) > MCT, hence b_2 is not hidden

i.e

$X_{b2} = X_{b2} = 11$

- $Sup(D_3 \rightarrow B_3) = \frac{\min(D_3, B_3)}{n} = \frac{\min(2.93, 0.75)}{5} = \frac{0.75}{5} = 0.15$

$$Conf(D_3 \rightarrow B_3) = \frac{\text{Supp}(D_3 \rightarrow B_3)}{Supp(D_3)} = \frac{\min(D_3, B_3)}{Supp(D_3)} = \frac{\min(2.93, 0.75)}{2.93} = \frac{0.75}{2.93} = 0.256$$

Since Sup($D_3 \rightarrow B_3$) > MST but conf($D_3 \rightarrow B_3$) < MCT, hence b_3 is hidden

i.e $X_{b3} = X_{b3} + V_{b3}$

$X_{b3} = 3 + 0.75 = 3.75 \approx 4$ to nearest whole number

$$Sup(D_4 \rightarrow B_4) = \frac{\min(D_4, B_4)}{n} = \frac{\min(2.70, 1.16)}{5} = \frac{1.16}{5} = 0.232$$

$$Conf(D_4, B_4) = \frac{\text{Supp}(D_4, B_4)}{Supp(D_4)} = \frac{\min(D_4, B_4)}{Supp(D_4)} = \frac{\min(2.70, 1.16)}{2.70} = \frac{1.16}{2.70} = 2.43$$

Since Sup($D_4 \rightarrow B_4$) < MST and conf($D_4 \rightarrow B_4$) > MCT, hence b_4 is hidden

i.e

$X_{b4} = X_{b4} + V_{b4}$

$X_{b4} = 5 + 1.16 = 6.16 \approx 6$ to nearest whole number

- $Sup(D_5 \rightarrow B_5) = \frac{\min(D_5, B_5)}{n} = \frac{\min(2.26, 0.95)}{5} = \frac{0.95}{5} = 0.19$

$$Conf(D_5 \rightarrow B_5) = \frac{\text{Supp}(D_5 \rightarrow B_5)}{Supp(D_5)} = \frac{\min(D_5, B_5)}{Supp(D_5)} = \frac{\min(2.26, 0.95)}{2.26} = \frac{0.95}{2.26} = 0.42$$

Since Sup($D_5 \rightarrow B_5$) < MST and conf($D_5 \rightarrow B_5$) < MCT, hence b_5 is hidden i.e

$X_{b5} = X_{b5} + V_{b5}$

$X_{b5} = 4 + 0.95 = 4.95 \approx 5$ to nearest whole number

Transformed Database

	A	B	C	D
T1	10	**6**	8	3
T2	3	11	6	14
T3	6	**4**	9	13
T4	7	**6**	8	12
T5	11	**5**	7	10

6. DISCUSSION OF RESULT

From the result shown in the transformed database, the algorithm was able to hide 4 out of the 5 items on the sensitive column resulting into 80% privacy protection of the sensitive column B. Comparing with the previous works of Manoj and Joshi [8] and Berberoglu and Kaya [9] on the same dataset, which hide only 1 out of the 5 items, resulting into 20% privacy protection of the sensitive column B of the data while the proposed algorithm hides 80% of the sensitive column B.

7. CONCLUSION AND FUTURE WORK

In this work, we proposed an improved algorithm for hiding sensitive quantitative data that combines the concept of noise with association rule mining to generate the useful association rules in quantitative data and then hide them using privacy concept. The algorithm works on association rules in quantitative data unlike previous work done in this field which works on binary data. In the future we plan to apply the algorithm to real life dataset; also

we propose that in the proposed algorithm the association rule should contain more than one item on each side of the rule and that the algorithm should be used for text data after they might have been normalized into quantitative data.

8. REFERENCES

[1] Doug Alexander *(2016)* - data mining *http://www.laits.utexas.edu/anorman/BUS.FOR/course.mat/Alex/. Accessed on January 2016*

[2] Alexandre Evfimirvski, Ramakrishnan Srikant, Rakesh Agrawal and Johannes Gerhke Privacy preserving mining of association rules. *In proceedings of the 8th ACM SIGKDD international Conference on Knowledge discovery in databases and data mining*, Edmonton, Akberta, Canada July 23-26, 2002, pages 217-228

[3] Rakesh Agrawal and Ramakrishna Srikant (2000) SIGMOD *'oo proceedings of the 2000 ACM SIGMOD international conference on Management of data* pages 439-450

[4] Lindell Y and Pinkas B (2000) Advances in cryptology *'crypt '000 proceedings, LNCS 1880, Springer-Vertag,pp.20-24,August 2000.*

[5] Alexandre Evfimievski (2002)- randomization in a privacy preserving data mining SIGKDD explorations. Volume 4-issue-2, 2002- *international journal of computer science and information technologies*

[6] Ila Chandrakar (2016). *Hybrid algorithm for privacy preserving association rule mining. Department of Information Technology VNR Vignana Jyothi Institute of Engineering and Technology Hyderabad, India http://www.ieee.org/documents/hybrid-algorithm. Accessed May 2016.*

[7] Evfimievski A., Srikant R., Agrawal R., and Gehrk J., (2002). Privacy Preserving Mining of Association Rules, in *Proceedings of the 8th Conference on Knowledge Discovery and Data Mining*, New York, USA, pp. 1-12, 2002.

[8] Manoj Gupta and Joshi R. C. (2009). Privacy Preserving Fuzzy Association Rules Hiding in Quantitative Data. *International Journal of Computer Theory and Engineering*, Vol. 1, No. 4, October, 2009, 1793-8201.

[9] Berberoglu T. and Kaya, M. (2008). Hiding Fuzzy Association Rules in Quantitative Data, *The 3rd International Conference on Grid and Pervasive Computing Workshops*, May 2008, pp. 387-392.