ISSN 1512-1232

AN IMPROVED ALGORITHM FOR PRIVACY PRESERVING QUANTITATIVE DATA USING ASSOCIATION RULE MINING AND PERTURBATION TECHNIQUE

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

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

Abstract

Mining of database is required in order to get useful information from it. One of the challenges of data mining is privacy of some sensitive data in the database for the mining task. One way to overcome the issue of privacy in data mining is to incorporate privacy techniques. Privacy Preserving Data Mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited disclosure. Existing studies in PPDM using quantitative data had drawbacks of high number of rules generated but few number of the sensitive item hidden. Also the scalability of the algorithms are not measured hence it is impossible to ascertain how the algorithm will perform as the data size increases. This study proposes a perturbation association rules hiding algorithm for privacy of quantitative data to provide an improved algorithm which performs efficiently as the data size increases. 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 value would be hidden. Experimental results show that the algorithm performs efficiently on large data sets.

Keywords: *Data perturbation, Data mining, Data privacy, Association rule mining*

1. Introduction

With the growing use of computers, there is a great amount of data being generated by such computer systems. Government agencies, scientific institutions and businesses have all dedicated large resources to collecting and storing data. In reality, only a small amount of these data will ever be used because, in many cases, the volumes are too large to manage. According to Kantardzic and Mehmed [1], in today's fiercely competitive business environments, companies need to rapidly turn their terabytes of raw data into significant insights into their customers and markets to guide their marketing, investment and management strategies. In order to get useful information from the database, mining of the database is required. Data mining is the process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. As asserted by Rakesh and Ramakrishnan [2], “the fruitful direction for the future of data mining research will be the development of techniques that incorporate privacy concerns into data mining”. The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users and sophistication of data mining algorithms to

ISSN 1512-1232

leverage this information. Privacy preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited disclosure.

A number of techniques such as randomization, suppression, summarization, association rule, perturbation, cryptography and k-anonymity have been suggested in recent years according to a survey carried out by Alexandre and Tyrone [3].

The present paper is an extension of our paper entitled Privacy Preserving Association Rule Mining Using Perturbation Technique [4], delivered at Ibadan ACM Chapter International Conference on Computing Research and Innovations, University of Ibadan, Nigeria, 7 – 9 September, 2016.

The rest of this paper is organized as follows. Related works are presented in Section 2 while our algorithm is presented in Section 3. In Section 4, the results obtained and discussion on the scalability, effect of increasing the minimum support and confidence values and classification accuracy of the algorithm are presented. Section 5 concludes the paper.

2. Related works

Nikhil *et al* [5] proposed an approach that modifies few transactions in a transaction database so that it gains the support of the sensitive rules and confidence of the sensitive rules and it also reduces the side effects. The techniques presented in the work increased the number of hidden sensitive rules and also reduce the number of modified entries.

A privacy preservation data mining algorithm in which it was assumed that only sensitive data items can be found in the database was proposed by Ila [6]. The 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 association rule, either the support or confidence is decreased to be smaller than pre-specified minimum support and minimum confidence.

Sathiyapriya *et al* [7] introduce the method for hiding sensitive quantitative data using genetic algorithm. The use of the genetic algorithm is to find useful association rule from the data. The approach contains two parts which are finding interval for rule and hiding the sensitive association rules. The algorithm is implemented with breast cancer and wine quality datasets. A fitness function is used for identifying transaction that will be perturbed in order to preserve non-sensitive rules. The approach minimizes lost rules but does not use the standard minimum support and minimum confidence value for a rule to be hidden. The main weakness of this approach is that many rules that are not needed for sensitive items are generated.

Manoj and Joshi [8] proposed a hiding algorithm that integrates the fuzzy set concepts and *a priori* mining algorithm to find useful fuzzy association rules from a quantitative database and then hide them using privacy preserving technique.. The algorithm considers fuzzy association rules which consist of only one item on both side of the rule. The algorithm is not efficient in that a large part of the rules generated resulted into lost rules. Berberoglu and Kaya [9] also worked on privacy of quantitative data using fuzzy logic but with low efficient result.

The objective of privacy preserving data mining is to hide certain information so that they cannot be inferred 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. The second approach, called input privacy, is to manipulate the data 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 studies that have been done 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.

ISSN 1512-1232

However, some works have been done to discover association rules from quantitative data which produce set of rules but hides less than 30% of the rules generated. In this present study, we propose a privacy preserving data mining algorithm, which uses the output privacy approach where data is preserved before it is released to the miner. The algorithm improves on the existing association rule hiding algorithms for quantitative data by combining perturbation and association rule mining techniques for privacy. Furthermore, the study addresses the challenges of applying privacy algorithm to only Boolean data by applying the algorithm to quantitative data. In addition, the scalability of the algorithm as data size increases was also carried out.

3. The Proposed Algorithm

As earlier mentioned in Section 1 of this paper, the present paper is an extension of the conference paper [4] we earlier presented on this work. For clarity sake, we re-present the algorithm with a simple illustrative example as follows:

In order to hide an association rule, $A \rightarrow B$, 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 $A \rightarrow B$. The second one is to decrease the support count of $A \rightarrow B$, 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 $A \cup B$. Based on these two strategies, we propose a privacy preserving data mining algorithm for hiding sensitive quantitative data using the concept of noise. The 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 consists 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 $A \rightarrow B$ 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.

3.1 Explanation of 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
Supp = support

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.

n = total number of transaction data
m= total number of attributes (items)
D= i^{th} attribute $1 \leq I \leq n$
I_j= j^{th} attribute $1 \leq j \leq m$
V= noise

ISSN 1512-1232

X= original data
X^i= inverse of original data
V_j= each noisy attribute 1≤ k ≤ I

Let the original data be depicted as shown in Table 1 containing five transactions T1, T2, T3, T4, T5 and A, B, C, D are the itemsets for each transaction.

Table 1: The Original Data Set

	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

Given that the sensitive column is column B

Set Minimum support threshold (MST) = 44% = 0.44
Set 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 = $x_1 \dots \dots \dots \dots \dots x_m$

$X^i = x_1^i \dots \dots \dots \dots \dots x_m^i$

For T1 x_{T1A}^i=1/10

N= 5, i. e., number of transactions in the database.

Table 2 gives the transformed noisy table obtained

Table 2: The Transformed Noisy Table

	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

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

ISSN 1512-1232

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

Table 3: The Count of Each Noisy Attribute

	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 $1 \leq j \leq m$ and $1 \leq k \leq m$, 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-itemset was used 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

ISSN 1512-1232

$$Conf(LHS,RHS) \leq MCT$$

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

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

ISSN 1512-1232

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

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

Table 4 shows the transformed database in which column B is said to be sensitive and it is secured using the algorithm. The hidden items are shown in bold text.

Table 4: The original data set

	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

The transformed database shows that 4 out of the 5 items on column B were hidden which implies that the algorithm hides 80% of the data.

4. Scalability Study of the Proposed Algorithm

The scalability of an algorithm is very important since it determines how efficient an algorithm is when applied to large dataset. An algorithm is scalable if it is suitably efficient and practical when applied to large data set [10]. The scalability of the algorithm (i.e., how the algorithm handles a growing data size / the capability of the algorithm to increase its total output as data size increases) was measured on dataset ranging from 10,000 to 100,000 dataset. The dataset was obtained from UCI repository [11]. It is a diabetic patient dataset. The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals. It includes over 50 features representing patient and hospital outcomes. The data contains attributes such as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc. The dataset is multivariate in nature. The attributes are integer data types and number of instances are 100, 000. In this study, the column used were admission type, discharge disposition, source of admission, time in hospital, number of lab test performed, number of lab procedures, number of medication, number of diagnoses, Also all the 100, 000 rows were used.

For the purpose of scalability study of the algorithm, the value of minimum support was set to 44% and minimum confidence was set to 75%. Table 5, Figures 1 and 2 show the results obtained.

Table 5: Result obtained from increasing Data Sizes with the Proposed Algorithm

Data Size	Time Taken (seconds)	Number of item hidden	% of item hidden
10000	14.63	9539	95.4
20000	26.88	19075	95.4
30000	41.41	28622	95.4

ISSN 1512-1232

40000	51.20	38275	95.7
50000	62.01	47983	96.0
60000	81.48	57670	96.1
70000	91.89	67333	96.2
80000	157.11	77038	96.3
90000	192.65	86746	96.4
100000	206.06	96483	96.5

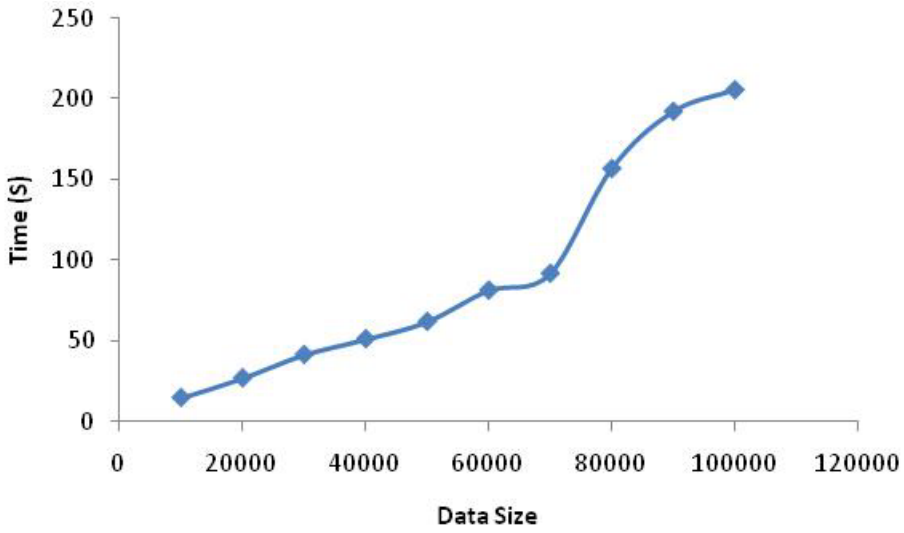

Figure 1: Time Taken (s) and Data Size

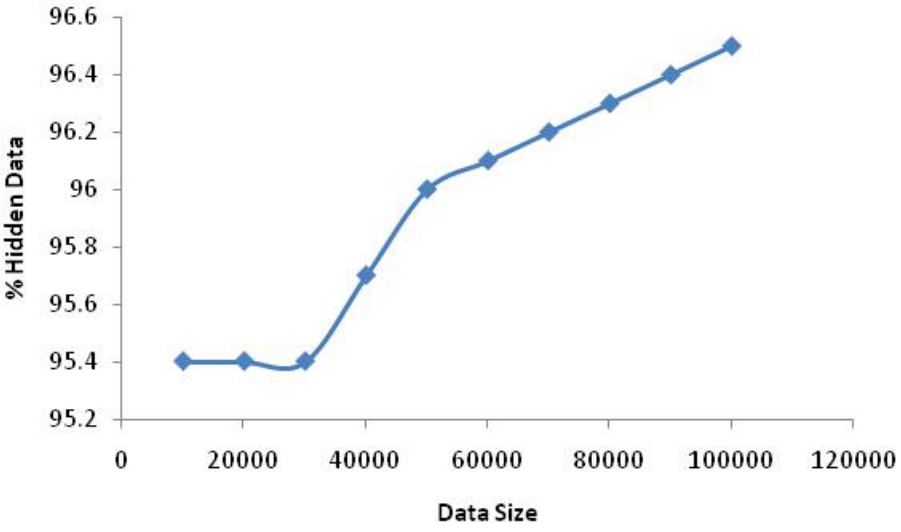

Figure 2: Percentage Hidden Data and Data Size

Table 5, Figures 1 and 2 show that as the database size increases, both the time taken to hide the sensitive column data and percentage rows of data hidden linearly increased as well. With high data size, the percentage data hidden is approaching 100%.

ISSN 1512-1232

4.1 Effect of Changing the Values of MST and MCT

Both the Minimum Support Threshold (MST) and Minimum Confidence Threshold (MCT) values were varied in order to study the effect these changes will have on the percentage of data hidden in a table.

(1) Table 6, Figures 3 and 4 show the result obtained when the minimum support and the minimum confidence were set to 0.34 and 0.65 respectively.

Table 6: Result with 0.34 Minimum Support and 0.65 Minimum Confidence

Data Size	Time Taken (Seconds)	Number of item hidden	% of item hidden
10000	12.94	9033	90.339
20000	23.77	18065	90.329
30000	40.57	27144	90.483
40000	54.29	36405	91.015
50000	64.94	45703	91.407

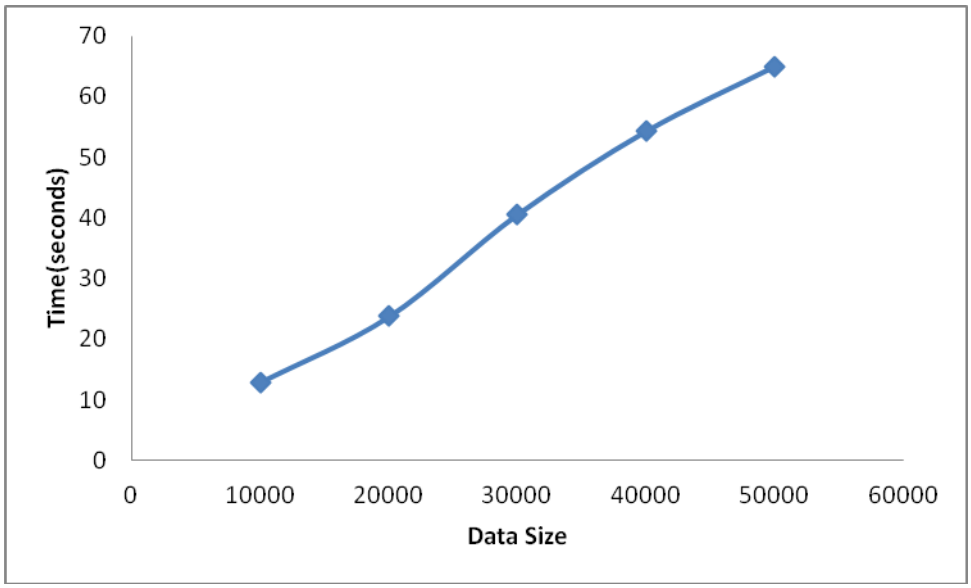

Figure 3: Time Taken(s) and Data Size when minimum support = 0.34 and minimum confidence = 0.65

ISSN 1512-1232

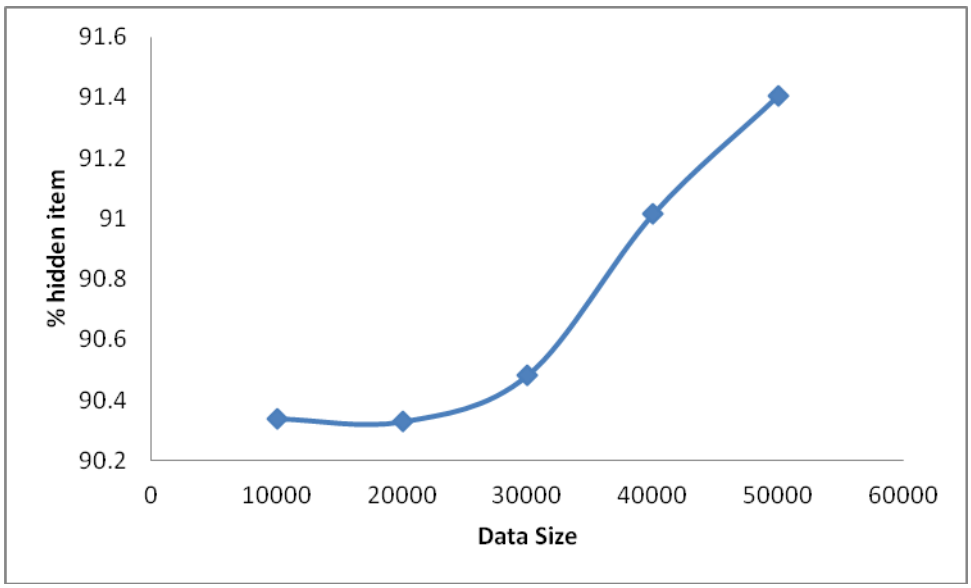

Figure 4: Percentage Hidden Item and Data Size when minimum support = 0.34 and minimum confidence = 0.65

(2) Table 7, Figures 5 and 6 show the result obtained when the minimum support is set to be 0.54 and the minimum confidence is set to 0.85.

Table 7: Result with 0.54 Minimum Support and 0.85 Minimum Confidence

Data Size	Time Taken	Number of item hidden	Number of item not hidden	% of item hidden
10000	13.14	9999	0	100
20000	30.77	19999	0	100
30000	41.92	29999	0	100
40000	55.38	39999	0	100
50000	71.97	49999	0	100

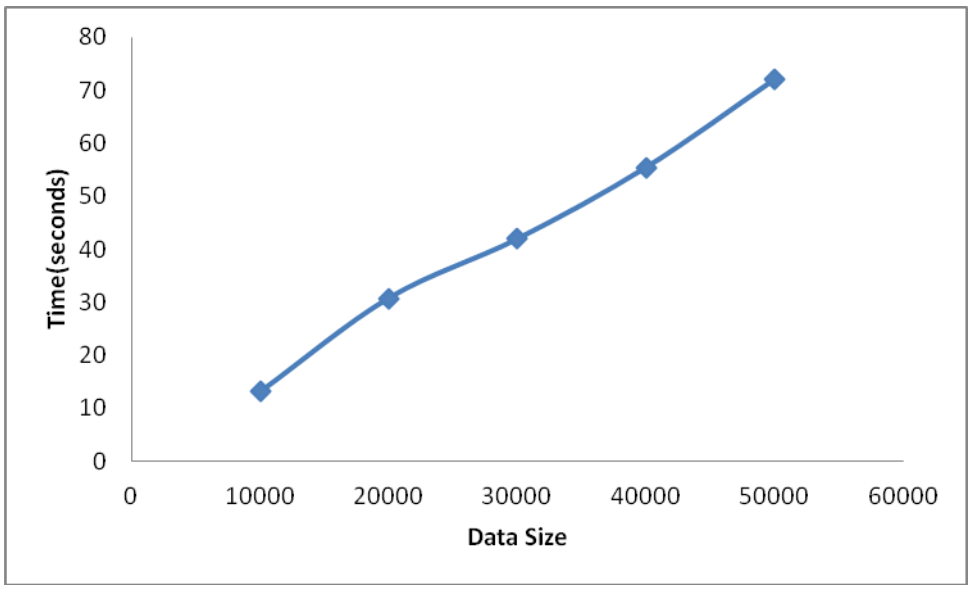

Figure 5: Time Taken(s) and Data Size when minimum support = 0.54 and minimum confidence = 0.85

ISSN 1512-1232

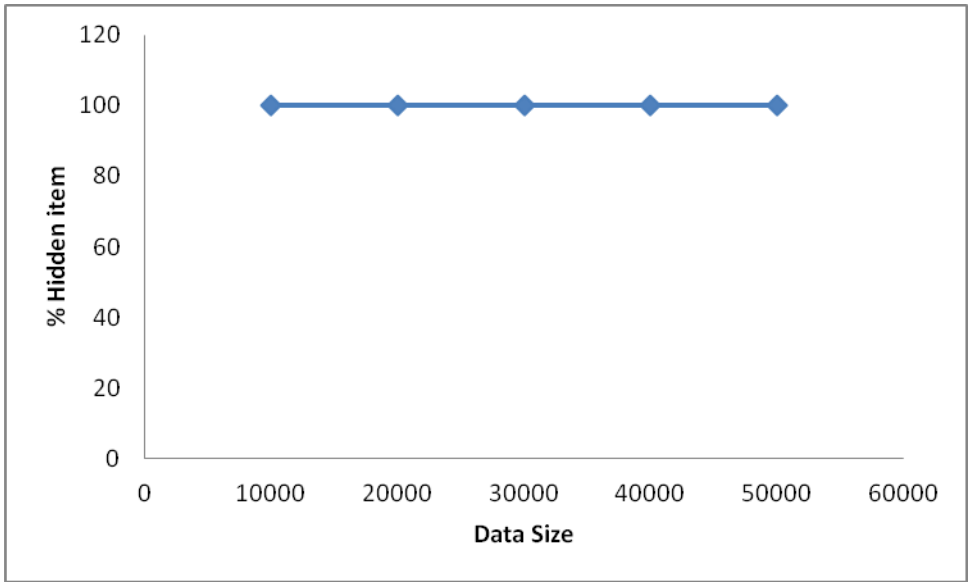

Figure 6: Percentage Hidden Item and Data Size when minimum support = 0.54 and minimum confidence = 0.85

Generally, it is observed that with the increase in minimum support and minimum confidence threshold values, all sensitive column data items were hidden.

4.2 Classification Accuracy Before and After the Privacy Algorithm was Applied to Data

When J48 (Decision Tree) classification algorithm was applied to the original dataset (1000 dataset), the correctly classified instances was 5, 597 (55.98%). Also after the proposed algorithm has been implemented on the dataset, the new dataset called transformed dataset has 5, 385 (53.86%) correctly classified instances. Hence, it shows that mining accuracy result after privacy preserving quantitative data algorithm was applied on the data will be minimally affected.

4.3 Brief Discussion of Results

There are various privacy preservation techniques such as summarization, perturbation, association rule hiding which are used to secure data items before releasing to the miner. Most of the research works carried out in this field are basically on hiding the presence/absence of a data. However, in this research work we joined the few researchers who worked on not just the presence/absence of data but on quantitative data.

With perturbation techniques and association rule hiding, a simple privacy preservation data mining algorithm was used to design a new algorithm which saves time and work on any size of data. It is established from this research work that the proposed algorithm worked efficiently than most of the other algorithms that worked on privacy of quantitative data, such as Manoj and Joshi [8]; and Berberoglu and Kaya [9].

ISSN 1512-1232

References

1. Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.
2. Rakesh Agrawal and Ramakrishnana Srikant (2000). Privacy-preserving data mining. In Proceedings of the ACM international Conference on Management of Data, pages 439-450, Dallas, Texas, United States.
3. Alexandre Evfimievski, Tyron Grandison (2013). Privacy Privacy Preserving Data Mining, IBM Almaden Research Center 650 Harry Road, San Jose, California 95120, USA.
4. Omoyele, Tobi Deborah and Akinola, Solomon Olalekan (2016). Privacy Preserving Association Rule Mining Using Perturbation Technique, Proceedings of the Ibadan ACM Chapter International Conference on Computing Research and Innovations (CORI), University of Ibadan, Nigeria, 7 – 9 September, 2016, pp. 91 - 95.
5. Nikhil N. Vaidya, Amit Pimpalkar, Ashwini Meshram (2015). Association of Data with Privacy Preserving of Sensitive Information, IJCAT - International Journal of Computing and Technology, Volume 2, Issue 4, April 2015 ISSN: 2348 – 6090.
6. Ila Chandrakar. Hybrid algorithm for privacy preserving association rule mining. Department of Information Technology VNR Vignana Jyothi Institute of Engineering and Technology Hyderabad, India.
7. Sathiyapriya K., Sudha Sadasivam G., Aarthi V. C, Divya K., Suganya C.J. P. (2012). Privacy preserving quantitative association rule mining. Trends in Innovative Computing 2012 - Intelligent Systems Design.
8. Manoj Gupta and Joshi R. C. 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, May2008, pp. 387-392
10. Wikipedia, Scalability, https://en.m.wikipedia.org/wiki/Scalability, accessed in November 2016.
11. http://UCI Machine Learning Repository_ Diabetes 130-US hospitals for years 1999-2008 Data Set_files

Article received: 2016-12-16