A Novel Approach for hiding Sensitive Items in Data mining using correlation

Dr. K. Sreenivasa Rao¹
¹Associate Professor,
Department of Computer Science and Engineering,
Vignana Bharathi Institute of Technology,
Hyderabad, Telangana, India
ksrao517@gmail.com
February 1, 2018

Abstract

Data mining is a process of extracting hidden information from huge databases. Alongside the upside of extracting useful information, it additionally postures dangers of uncovering client’s sensitive data. Privacy preservation data mining (PPDM) does the duty of hiding sensitive information of the client. Various methods exist in literature for preserving sensitive information. Those hiding techniques are based on support and confidence framework, which suffers with limitations such as choosing a suitable value of these support and confidence measures setting up which cause losing of useful information of the user, and hence lack of tradeoff between privacy & accuracy of the database. There exist some other approaches which consider measures other than support and confidence such as utility of the data items. We proposed an approach based on correlation among items in sensitive item sets to hide the sensitive frequent item-sets. We proved that our approach maintained the tradeoff between privacy and accuracy of the database which is the challenge of PPDM techniques.
1 Introduction

Correlation is an area of Probability and Statistics that gives answers for measuring the life of reliance between numeric factors (Wikipedia-Correlation, 2013 [7]). In an extraordinary case, if two numeric, irregular factors, \((X, Y)\), are free, they are verbally communicated to be uncorrelated. Coefficient of relationship, \(\rho\), is a numeric measure of the force of reliance between the factors (Feller, 1961 [8], p. 211), (Grinstead, 2013, p.291 [9]).

Correlation is often used as a primary technique to discover or specify relationships between variables. More precisely, the correlation is considered as a measure of the linear relationship between two variables. 

\[
r = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2 \sum_{i=1}^{n}(Y_i-\overline{Y})^2}}
\]

(1.1)

2 Motivation

Inspiration of presented approach depends on the way that generally the procedures that are inspected in the writing depend on support and certainty. Different measures are additionally talked about in writing like conviction, all-certainty, use and lift. (Naeem et. al.2010 [12]). Another motivation for the research is to device a mechanism, which can full fill objective of privacy preservation of data mining in association rule at the cost of minimum side effects and maintaining database accuracy.

3 Related Work

Privacy Preserving Data Mining (PPDM) (2004) has turned into an outstanding exploration theme in information digging for as far
back as couple of years. Clifton et al. [5] dissected that information mining also brings danger against our databases and he tended to some conceivable answers for securing protection in information mining. Rao, K. Srinivasa, et al. 2014 [6] proposed Distortion Technique for Hiding Sensitive Association Rules that alters positions of sensitive items while maintaining the support. The technique used the idea of representative rules to prune the rules first and then hides the sensitive rules. Typically the methods examined in writing and have an issue of symptoms and absence of tradeoff between protection and precision. In light of this thought, Jieh-Shan Yeh et al. in 2010 [1] developed PPUM, Privacy Preserving Utility Mining. Jieh-Shan Yeh et al. (2010) demonstrated two novel calculations. Execution of the two calculations is broke down with the assistance of three parameters miss cost, hiding Failure and Data Difference Ratio (DDR) amongst initial and altered database.

4 Proposed Approach

In this work, we utilized Pearson’s Correlation coefficient measuring system proposed by Rao, K. Srinivasa et al. [11] to conceal sensitive information items in the original database. Creators in [1] took the necessary steps on T10I4D100K dataset where as in this work we took the necessary steps on three datasets to be specific T10I4D100K, T40I10D100K and retail datasets. So as opposed to concealing delicate things specifically as on account of regular PPDM methods, this approach chooses things having least relationship limit for concealing procedure. Proposed solution is given underneath.

![Proposed Solution for Hiding Sensitive Items](image-url)
Pseudocode of the Proposed Approach

a. Input:
   1. Source database D
   2. Min support
   3. min Per_Corr_thres.
   4. Sensitive items set U

b. Output:
   A transformed database D’, where Frequent item sets containing sensitive items from U are hidden Algorithm

Step 1: Generate Frequent Item Sets
Step 2: Select the Sensitive Frequent Item Sets having items from U and obtain SFI database.
Step 3: Pearson’s Correlation Weighing Mechanism
   3.1. Generate Correlation Matrix for the frequent item Sets
   3.2. Find the Person’s Correlation value for each of the pairs of the columns in the Matrix
   3.3. Select Columns satisfying minimum Per_Corr_thres
   3.4. Hiding process based on sensitive item set U
   3.5. Copy the Modified Database to New database
   Repeat
   Apply Steps 1 thru 3 on new database.
   Until
   No sensitive items occurred in columns having Pearson’s correlation above Per_Corr_thres
Step 4: Finally write the new database to form Sanitized database.

In step 3.4, unique columns (represents items) satisfying min_Corr threshold are selected and in step 3.5, sensitive items among those unique items from step 3.4 are deleted

*Effectiveness Measurement*
Sufficiency of our proposed figuring is measured by accepting the evaluation measures used by Oliveira and Zaiane [13] and K. Srinivasa Rao et al. [11].

(a) **Hiding Failure**: It is the measure of sensitive items that are even mined from sanitized database.

\[ HF = \frac{|U(DB')|}{|U(DB)|} \]

Where \( U(DB) \) AND \( U(DB') \) denote the sensitive item sets discovered from the original database \( DB \) and the sanitized database \( DB' \), respectively. The cardinality of a set \( S \) is denoted as \( |S| \). HF is calculated by mining frequent itemsets from original and sanitized database say initial frequent itemsets \( (F) \) and new frequent itemsets \( (F') \). Now let \( S \) be the number of sensitive items found in \( F \) and \( S' \) be the number of sensitive items found in \( F' \). Then \( HF = \frac{(S'/S) \times 100}{1} \). If its value is less means there are less no of sensitive items in sanitized database i.e. privacy becomes more. So Hiding Failure is inversely proportional to Privacy.

(b) **Miss Cost**: It is the measure of non-sensitive items that are hidden by accident during hiding process.

\[ MC = \frac{|\sim U(DB) - \sim U(DB')|}{|\sim U(DB)|} \]

Where \( \sim U(DB) \) and \( \sim U(DB') \) denote the non-sensitive item sets discovered from the original database \( DB \) and the sanitized database \( DB' \), respectively. Now let \( NS \) be the number of non sensitive items found in \( F \) and \( NS' \) be the number of non sensitive items found in \( F' \). Then \( MC = \frac{(NS'/NS) \times 100}{1} \). Less value of miss cost indicates that side effects are less.

(c) **DDR**: It is called Data Difference Ratio between modified database and original database. It is the measure of database accuracy.

\[ DDR(DB, DB') = \frac{|DB - DB'|}{|DB|} \]

Let \( N \) be the total number of items in original database \( DB \) and \( N' \) be the total number of items in sanitized database \( DB' \).
Now DDR = (|N − N'|/|N|)∗100. If it is more means accuracy will be less. so it’s value must be less. Hence the low values of Hiding Failure & DDR maintain the tradeoff between privacy & Accuracy. Our approach achieves this results which ensures tradeoff between privacy & accuracy.

5 EXPERIMENTAL RESULTS

The experiment was performed on two synthetic datasets namely T10I4D100K [K. Srinivasa Rao et al. [11]] and T40I10D100K [http://fini.ua.ac.be/data/] [10] which are also used by Jieh-Shan et al. (2010) [1, 16]. Experiment was also conducted on retail data set given by Tom Brijs.

Table I: Characteristics of the dataset Considered for the work

<table>
<thead>
<tr>
<th>Data Set Name</th>
<th>No. of Transactions</th>
<th>Avg. Length of Transactions</th>
<th>Distinct Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>T10I4D100K</td>
<td>100000</td>
<td>673</td>
<td>870</td>
</tr>
<tr>
<td>T40I10D100K</td>
<td>100000</td>
<td>2640</td>
<td>942</td>
</tr>
<tr>
<td>Retail</td>
<td>88162</td>
<td>606</td>
<td>16470</td>
</tr>
<tr>
<td>T10I4D100K</td>
<td>1000</td>
<td>7</td>
<td>795</td>
</tr>
<tr>
<td>T10I4D100K</td>
<td>2000</td>
<td>14</td>
<td>836</td>
</tr>
<tr>
<td>T10I4D100K</td>
<td>3000</td>
<td>20</td>
<td>845</td>
</tr>
<tr>
<td>T40I10D100K</td>
<td>1500</td>
<td>40</td>
<td>928</td>
</tr>
<tr>
<td>T40I10D100K</td>
<td>3000</td>
<td>79</td>
<td>936</td>
</tr>
<tr>
<td>Retail 2000</td>
<td>2000</td>
<td>13</td>
<td>4775</td>
</tr>
<tr>
<td>Retail 3000</td>
<td>3000</td>
<td>20</td>
<td>5897</td>
</tr>
</tbody>
</table>

5.1 Experimental Results Discussion

Performance of the proposed approach on T10I4D100K and T40I1D100K datasets gives optimal results with negative correlation thresholds as the items from our sensitive itemsets are negatively correlated and if we move on to positive correlation values Hiding failure increases to 30 % and 70 % and so on.DDR value follows reverse
direction of Hiding Failure i.e. it’s value is more in the negative values of Correlation threshold, and if we move on to positive correlation values DDR value decreases. we achieved the tradeoff between privacy and accuracy. But on Retail dataset negative correlation threshold values are giving hiding failure of 0% to 23% and DDR of 0.07% to 2.3% on various sensitive itemsets with various correlation thresholds and DDR of 0.07% to 0.5%. Positive correlation threshold values are giving hiding failure of 0% to 23% DDR of 0.1% to 2.3% which indicates that sensitive items are equally distributed at negative as well as positive correlation threshold values. There is no Miss cost. Detailed tables indicating the said results were not included in this paper but a consolidated comparison table is given in next section.

5.2 Experimental Results Comparison

Performance of the proposed approach is compared with that of Jieh-Shan et al. (2010) [1]. Three parameters were considered i.e. Hiding Failure, Missing Cost & Data Difference Ratio (DDR). Below mentioned Table II shows optimal results obtained by both the proposed approach and its counterpart. From the table, we know that even though we are getting some higher values of hiding failure & DDR for some sets of sensitive items, we are ensuring the lower DDR value and the other two measures reach the minimum value. Former approach used utility among the items as threshold for hiding sensitive items where as the proposed approach used the correlation among the items as threshold.

Table II: Results Comparison of Proposed work with existing work (Jieh-Shan et al. (2010) [1]

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Data set, Length</th>
<th>Sensitive Item set U &amp; its Size</th>
<th>Threshold</th>
<th>HF (%)</th>
<th>MC (%)</th>
<th>DDR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUIUF</td>
<td>T104D100K,1000</td>
<td>5</td>
<td>5,000 &amp; 5,000 (Utility)</td>
<td>0</td>
<td>0</td>
<td>2.95</td>
</tr>
<tr>
<td>MSIF</td>
<td>T104D100K,1000</td>
<td>5</td>
<td>5,000 &amp; 5,000 (Utility)</td>
<td>0</td>
<td>0</td>
<td>2.94</td>
</tr>
<tr>
<td></td>
<td>T104D100K,1000</td>
<td>U1=`(274,346,753,742,809)</td>
<td>-0.5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>T104D100K,1000</td>
<td>U2=`(392,461,569,801,862)</td>
<td>-0.5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>T104D100K,2000</td>
<td>U1=`(274,346,753,742,809)</td>
<td>-0.5</td>
<td>0</td>
<td>0</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Table II shows optimal results obtained by both the proposed approach and its counterpart. From the table, we know that even though we are getting some higher values of hiding failure & DDR for some sets of sensitive items, we are ensuring the lower DDR value and the other two measures reach the minimum value.

6 CONCLUSION

A Novel approach of privacy preserving data mining is proposed which improvises the values of hiding failure, missing cost and DDR. Here privacy preserving data mining is done by considering correlation among items in transaction in database. Pearson’s Correlation Coefficient based weighing mechanism is applied on the two IBM synthetic data sets namely T10I4D100K, T40I1D100K and retail datasets to hide the sensitive item sets. Performance of the approach is compared with Jieh-Shan et al. (2010) [1,16] and also with C.Saravanabhavan et al. (2013) [3].Our approach results in better values of hiding failure, miss cost and DDR values provided. In spite of having advantages, this framework has a drawback of resulting high hiding failure if the sensitive items having above correlation threshold is less in number (If the correlation values of the sensitive items are at longer intervals), so performance of our approach depends on how many sensitive items are satisfying correlation threshold, but in our work there is a provision to tune correlation threshold to different values so that we can have a control of increasing or decreasing the hiding failure. In future this framework can be extended by including a percent threshold for hiding specific percentage of sensitive items, repeat the same framework.
using replacement of hidden items by other items and to implement the same framework considering range wise selection of correlation threshold and also we want to extend this framework for a database having categorical attributes.

7 Acknowledgments

The authors would like to thank the authors of the papers cited in this paper which created interest to explore their ideas. The authors would like to thank the anonymous reviewers for their careful reading and insightful comments that have helped in improving of this paper.

References


