The Nucleus **50**, No. 4 (2013) 359-367

The Nucleus
A Quarterly Scientific Journal of Pakistan Atomic Energy Commission
NCLEAM, ISSN 0029-5698

SANITIZING SENSITIVE ASSOCIATION RULES USING FUZZY CORRELATION SCHEME

SONIA HAMEED, F. SHAHZAD and *S. ASGHAR[1]

Department of Computer Science, Mohammad Ali Jinnah University, Islamabad, Pakistan

[1]Institute of Information Technology, PMAS University of Arid Agriculture, Rawalpindi, Pakistan

(Received June 16, 2013 and accepted in revised form November 26, 2013)

Data mining is used to extract useful information hidden in the data. Sometimes this extraction of information leads to revealing sensitive information. Privacy preservation in Data Mining is a process of sanitizing sensitive information. This research focuses on sanitizing sensitive rules discovered in quantitative data. The proposed scheme, Privacy Preserving in Fuzzy Association Rules (PPFAR) is based on fuzzy correlation analysis. In this work, fuzzy set concept is integrated with fuzzy correlation analysis and Apriori algorithm to mark interesting fuzzy association rules. The identified rules are called sensitive. For sanitization, we use modification technique where we substitute maximum value of fuzzy items with zero, which occurs most frequently. Experiments demonstrate that PPFAR method hides sensitive rules with minimum modifications. The technique also maintains the modified data's quality. The PPFAR scheme has applications in various domains e.g. temperature control, medical analysis, travel time prediction, genetic behavior prediction etc. We have validated the results on medical dataset.

Keywords: PPDM, Fuzzy association rule hiding, PPFAR**.**

1. Introduction

Recent advances in digital data storing capacity have made possible to collect and analyze millions of transactions within the database [1]. The data includes bank records, medical records, criminal records and phone call records. This tremendous growth in data has led people and organizations to face the challenge of transformation of data into useful information or knowledge.

Data Mining is an ideal approach to fulfill this challenging task. In ref. [2] the authors state, "data mining can be referred to as extracting or mining knowledge from data. It is an important step for knowledge discovery".

The authors in ref. [3] introduced association rule mining in binary datasets, which is a pivoted tool for finding the useful pattern. Association rule mining is a process to find out the item sets which occur frequently together. It is widely used in promoting businesses.

Although classical association rule mining is very valuable but it has limitations. It cannot work on every kind of data. Rules generated from binary data only focuses on existence and non-existence of any item in the data set. In reality, it is not always the case. Data can be of quantitative nature. Consider an example discussed by [5], numerous attributes could be present in blood test of a patient. However, attribute's quantity is more important than its presence or absence for determination of illness. The concept of quantitative association rules arose from above mentioned fact.

In these types of rules, the quantitative attributes are divided into intervals and single elements are either members or non-members of those intervals. This approach has limitation of rigid boundaries i.e. an item can exist or cannot exist in the intervals. For surmounting this limitation, an approach was developed called fuzzy association rules. This approach has the advantage of interval overlapping which makes fuzzy sets [6].

The basis of the data mining is to generalize data across people, rather than disclose information about individuals. Besides advantages of finding valuable information, it also originates threats of illuminating sensitive information. The concept of Privacy Preservation was introduced by [7]. Privacy preservation is a technique used for sanitizing sensitive knowledge. The sensitive knowledge is represented by a particular group of association rules called sensitive association rules. For decision making, such rules are useful and must remain hidden.

* Corresponding author : sohail.asg@gmail.com

The literature has proposed numerous techniques e.g. [7-10] to cope with the problem of privacy preservation. Most of the proposed literature concentrates on sanitizing Boolean sensitive association rules. In our work, we have devised a scheme for sanitizing sensitive fuzzy association rules. We achieved this sanitization using quantitative data. The proposed scheme has the potential to accomplish comprehensive sanitization of these sensitive rules. Our method does not produce ghost rules.

The remaining paper is divided into following sections. Section 2 defines problem statement. Section 3 discusses related work. Section 4 describes the proposed scheme, PPFAR. The example of PPFAR scheme is included in section 5. Section 6 describes experimental evaluation. Discussions on results are given in section 7. In section 8, we discuss conclusions and future work.

2. Problem Statement

This section discusses the problem statement. First we define the problem of finding association rules. Secondly we define the problem of sanitizing sensitive fuzzy association rules.

Let I = {i_1, i_2, i_3, ….., i_n} be the of itemset. Let D be the dataset where each T $\in$ D. Given an item set X $\in$ I, X is contained in T if X $\subseteq$ T. An association rule is an implication of the form $S_a \rightarrow S_b$. Both S_a and S_b are item sets. A rule has support *s* if and only if s% transactions in D contain S_a U S_b. A rule has confidence *c* if and only if c% transactions in D contain S_a also contain S_a U S_b. An item set is frequent if and only if its support exceeds certain support threshold *misupp*.

Sanitization of sensitive fuzzy association rules can be stated as follows. Let D be the dataset. Let FAR be a fuzzy association rules set $\{(F_X \rightarrow Fy,),(F_x \rightarrow F_z),.....(F_i \rightarrow f_N)\}$. SFAR represents sensitive fuzzy association rules [19] like $\{(SF_X \rightarrow SF_Y),(SF_X \rightarrow SF_Z)\}$. The goal is to change D in D' in such a way where all sensitive fuzzy association rules are sanitized. In this case, D' is the released database.

3. Related Work

In this section we have compared different methodologies in context of sanitizing sensitive association rules. In ref. [8], the authors highlight notion of limiting disclosure for sensitive rules and put the pains for sanitizing some frequent item sets. Sanitization can be achieved in such a way that non sensitive data can be minimally affected. For this purpose, authors modify database in such a way that support of sensitive rules is decreased. Later on, this work was extended by [13], where the authors investigated confidentiality issues of association rules. They have based their solution on modifying original data values. Following the same path [13] proposed a technique in which they removed individual values from data to thwart sensitive rules. To achieve hiding, authors proposed algorithms namely CH (Cyclic Hide), GIH method (Generating Itemset Support Reduction Algorithm) and CR (Confidence Reduction Algorithm). In all the above algorithms the concept of replacing known values with unknown values is used. They performed experiments on web log data containing 32711 instances. They have evaluated their algorithm on the basis of the side effects of ghost rules and lost rules. According to analysis, GIR has least number of side effects.

The specific algorithms proposed [14] only remove information from database, rather than modifying the existing information. Parameters used for evaluating proposed algorithm were hiding failure, miss cost and artifactual patterns. Though algorithm does not produce artifactual patterns but factors of miss cost and hiding failure was present in algorithm, even the more legitimate patterns were missed. Four algorithms Naive, MaxFIA, MinFIA and IGA are proposed [14]. The experiments were conducted using the IBM Synthetic dataset generator with 500 different items and 100k transactions, which shows that IGA is better than other approaches because IGA's impact on the database was lesser and the miss cost of IGA was minimum as compared to other approaches. In ref. [14] the authors used three diverse methods to measure dissimilarity between the source and sanitized database. The algorithm's efficiency was measured with respect to CPU, by keeping both the dataset size and the set of restrictive patterns constant. For checking scalability the size of input was increased.

Authors in ref. [7] proposed a technique based on reconstruction. This technique was used in estimation of probability distribution of original numeric data values for building a decision classifier from perturbed training data. They evaluate the algorithms for uniform and Gaussian perturbation. This method was simple but it lacked a formal framework for proving quantification privacy. Similarly, algorithm proposed [15] was based on Expectation Maximization (EM) for distribution reconstruction, which converges to the maximum likelihood estimate of original

distribution. The method produces quantification and dimension of confidentiality and data loss. Another reconstruction based technique, proposed [16], in which they give an idea of distortion to pre-process data before applying mining process and their key aim was to guarantee isolation altitude of individual entries in each tuple.

In ref. [12], authors analyze the impact of fuzzy set concept. In refs. [5, 12] authors used the same strategy for sanitizing fuzzy association rules in quantitative data, but their techniques are different from each other in the following manner [12] applied decreasing confidence strategy, while [5] used strategy to decrease the support. Both used same data set of Breast Cancer (UCI) for experimentations and after performing analysis on experimentation of both, it can be concluded that technique proposed [5] perform better hiding than [12]. Later on [17] used a border based technique for sanitizing fuzzy weighted sensitive rules; in which they put their efforts to discover extremely projecting non-sensitive data having fuzzy weights and dig out recurrent data with fuzzy weights. After that they sanitize them in combination with sensitive items. Results of this technique are comparatively better than previous one.

We have analyzed the literature in context of PPDM and concluded that two areas are neglected. First, existing techniques are working on Boolean data. In real world scenarios data is not always in the form of 0's and 1's. There are only few techniques, working on quantitative data. Therefore, the need arises for a technique, which caters for the quantitative data. Second, existing techniques are using support and confidence framework for selection of sensitive rules; which do not provide the strong basis for selection of sensitive rules.

4. PPFAR Scheme

In the proposed scheme, first we fuzzify the quantitative data with the help of the membership function given below in the Figure 1.

After performing the fuzzification we have fuzzy itemsets. In ref. [7] authors applied an algorithm for getting fuzzy association rules. This algorithm is called fuzzy correlation rule mining. In Fuzzy Correlation every rule shows the related Association Rules set and this set is called sensitive. We then sanitize these sensitive rules by using PPFAR method, keeping in view the above mentioned relationship. The PPFAR method has below mentioned steps. The PPFAR method's flow is given in Figure 2.

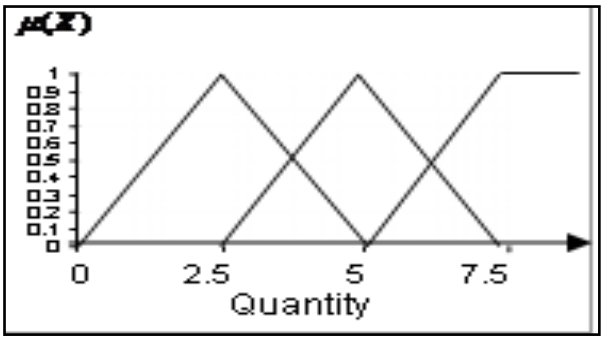

Figure 1. Membership function.

4.1. First Step Quantitative Data's Fuzzification

Fuzzy sets were introduced by Zadeh in 1965. They are an expansion of the classical crisp sets. Fuzzy sets assign values which plunge between 0 and 1. *X* represents the crisp set and μ_A represents membership function that defines set A. This is formally shown as $\mu_A : X \rightarrow [0,1]$ as defined in [19]. Figure 1 represents membership function used in PPFAR method. This function assigns a membership degree to every element present in *X*. When fuzzifying, one thing must be taken care of: fuzzy set A's support be determined from crisp set having all those items whose membership grade is not 0 in *A*.

$$Sup(A) = \{x \in X \mid \mu_A(x) > 0\} \quad (1)$$

4.2. Second Step: Mining Fuzzy Association Rule

In this step, Apriori algorithm [3] is used for mining fuzzy association rules. The method is defined formally as: let $F = \{f_1, f_2, \ldots\ldots f_m\}$ denote fuzzy itemset, $T = \{t_1, t_2, \ldots\ldots t_n\}$ denote fuzzy records set, while every record t_i is shown in the form of vector having m values, $(f_1(t_1), f_2(t_2), \ldots\ldots f_m(t_i))$ and $f_i(t_i)$ represents membership grade that t_i belongs to fuzzy item $f_j(t_i) \in [0,1]$. FAR has an implication form like $F_x \rightarrow F_y$. Both $F_x, F_y \subset F$ are fuzzy item sets. A fuzzy association rule of the form $F_x \rightarrow F_y$, is present in fuzzy dataset *T* with fuzzy support $f\sup(\{F_x, F_y\})$. The calculation is, as shown in Equation 2 discussed in [11].

$$f\sup(F_x, F_y\}) = \frac{\sum_{i=1}^{n} Min(f_j(t_i) \mid f_j \in \{F_x, F_y\})}{n} \quad (2)$$

If $f\sup(\{F_x, F_y\})$ for a rule is greater than or equal to the user-defined fuzzy support (S_f), that rule $(F_x \rightarrow F_y)$ is called a useful fuzzy association rule. This shows that F_x and F_y are found recurrently together. Fuzzy correlation is calculated for discovering useful fuzzy association rules.

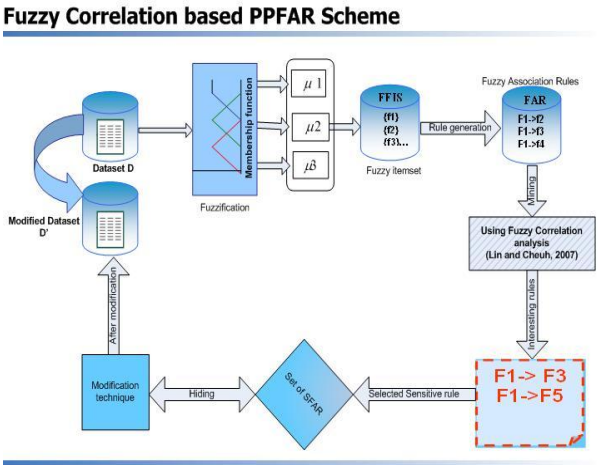

Figure 2. Correlation based scheme for PPFAR.

4.3. *Fuzzy Correlation Analysis*

The procedure of calculating fuzzy correlation is; Let A and B be two fuzzy item sets where $A,B \subset F$. Here F represents fuzzy collection. *A, B* are defined over set *X* with μ_A and μ_B. μ_A and μ_B are membership functions. Fuzzy item sets *A* and *B* are shown formally in Equations 3 and 4 as defined in [19].

$$A = (x, \mu_A(x)) \mid x \in X) \quad (3)$$

$$B = (x, \mu_B(x)) \mid x \in X) \quad (4)$$

In above equations $\mu_A, \mu_B \in X$, shows paired data sequence, and are represented as $\{(x_i, \mu_A(x_i), \mu_B(x_i)) \mid i = 1....n\}$. This represents membership degree for fuzzy item sets A, B defined over X. Equation 5 discussed in [11] represents fuzzy correlation coefficient between the two sets.

$$Cor_{A,B} = \frac{S_{A,B}}{\sqrt{S_A^2, S_B^2}} \quad (5)$$

The values obtained after applying Equation 5, lie in range of -1 and 1. The linguistic variables of PPFAR method are as follows:

- If $|Cor_{A,B}|$ is $\geq$ 0.5 and $\leq$ 1, then they have high correlation.
- If $|Cor_{A,B}|$ is $\geq$ 0 and < 0.5; it means low correlation.
- If $|Cor_{A,B}|$ is 0, then it means no correlation.
- If $|Cor_{A,B}|$ < 0, then they are negatively correlated.

4.4. *Fuzzy Correlation Rules Mining Algorithm*

In [19] authors define fuzzy items as, "Let $F = \{f_1, f_2, ... f_m\}$ represent Fuzzy items, $T = \{t_1, t_2, ... t_n\}$ be a fuzzy data records, and each record t_i represents a vector with m values, $(f_1(t_i), (f_2(t_i), ... f_m(ti))$ where $f_j(t_i)$ is membership degree that t_i belongs to fuzzy item f_j, means that $f_j(t_i) = \mu_{fj}(ti)), \ f_j(t_i) \in [0,1]$; S_f is the user-defined fuzzy support; α: A user Specified fuzzy Correlation threshold".

4.5. *Third Step: Sensitive Fuzzy Association Rules Identification*

Once we have performed both coefficient correlation analysis and fuzzy association rules creation, we would now find all those fuzzy association rules having value greater than α and $f\sup$. These rules would be considered sensitive fuzzy association rules.

4.6. *Fourth Step: Prune Non-Sensitive Items*

After selecting sensitive fuzzy association rules, only those fuzzy items would be amended, which have high correlation with each other. We would replace higher values for fuzzy items with 0 in all transactions. It would decrease values of concerned sensitive rules. The method is repeated until there are no sensitive fuzzy association rules having values greater than specified threshold. This step results in modified dataset. Table 1 in [19], shows the notations used in algorithm shown in Figure 3. The algorithm in figure 3 has been adopted from [19] and changed according to our modification technique.

5. Putting Scheme into Work

This section describes demonstration of proposed scheme with the help of an example for achieving sanitization of sensitive rules.

Table 1. Notations used in PPFAR scheme.

D	Source Data
D'	Released data
TC	Transaction ID
N	Total number of transaction
Ant/ FX	Tail
Con/FY	Head
FAR	Fuzzy Association rules set
FSAR	Sensitive Association Rules set
Cor	Correlation
F_i	Fuzzy Item set
F	Item set $\{f_1, f_2, f_3 \ldots, f_m\}$
S	candidate combinations set
$f\sup(f_i)$	Fuzzy Support
S_f	Minimum Fuzzy Support

Input

Dataset D (Fuzzified dataset)

α : User defined threshold

S_f

Output

Dataset D' (Modified dataset)

FAR' (Fuzzy Association rule set)

Step 1: Data Fuzzification using membership function given in Figure 3.1.

Step 2: Finding $(Cor_{A,B})$ and FAR support S_f

Step 2.1: For every $f_i \in F$, $f\sup(f_i)$ is calculated.

Step 2.2: Let $S_1 = \{ F_i \mid F_i \in f, f\sup(F_i) \geq S_f \}$ represent the set of recurrent fuzzy items.

Step 2.3: let $CA_2 = \{(F_A, F_B)\}$ be two fuzzy item sets of S_1 , where $F_A, F_B \in S_1$ and $F_A \neq F_B$.(CA_2 comes from S_1 cross join with S_1 , since F_A and F_B are members of S_1)

Step 2.4: For every member of CA_2 , (F_A, F_B) we compute the fuzzy support $(f\sup(\{F_A, F_B\}))$ and the fuzzy Correlation F_A and F_B $(Cor_{A,B})$.

Step 2.4.1: If $(f\sup(\{F_A, F_B\})) \geq S_f$ and $(Cor_{A,B}) \geq \alpha$, than grouping of (F_A, F_B) is a member of S_2

Step 2.5: Then CA_k , if $k \geq 3$ then we find CA_k by joining S_{k-1} with S_{k-1} .

Step 3: Apply alteration Method (substitute 1 with 0) till we do not reach minimum threshold.

Step 4: Repeat Step 2.

End

Figure 3. Algorithm for proposed PPFAR.

In this example, every attribute contains 3 fuzzy regions and 3 membership values are created for every item by using the predefined membership function given in Figure 1.

Table 2. Sample fuzzy data with fuzzy support.

T	fi_1	fi_2	fi_3	fi_4	fi_5
T1	0.1	0.9	0.3	0.5	0.3
T2	0.2	0.8	0.4	0.8	0.4
T3	0.2	0.7	0.1	0.3	0.5
T4	0.4	0.5	0.6	0.1	0.1
T5	0.7	0.3	0.4	0.8	0.2
T6	0.5	0.6	0.7	0.9	0.8
T7	0.8	0.1	0.9	0.4	0.9
T8	0.7	0.1	0.5	0.2	0.7
T9	0.8	0.4	0.3	0.5	0.6
T10	0.2	0.7	0.2	0.4	0.5
FSUPPORT	0.46	0.51	0.44	0.49	0.50

A sample fuzzy dataset is shown in Table 2.We assume the value of specified fuzzy support (S_f) as 0.30 and fuzzy correlation as 0.40. Firstly, the fuzzy support of each fuzzy item is calculated. As in given example, fuzzy support is greater than defined threshold. So, we obtain the frequent item sets $S_1 = \{fi_1, fi_2, fi_3, fi_4, fi_5\}$.

In the next step, the set of combinations of two fuzzy items is calculated. We call it CA_2 and it is obtained by joining S_1 with itself. CA_2 is defined as below and shown in Figure 4:

$$CA_2 = \{(fi_1, fi_2),(fi_1, fi_3),(fi_1, fi_4),(fi_1, fi_5),(fi_2, fi_3),$$
$$(fi_2, fi_4),(fi_2, fi_5),(fi_3, fi_4),\ (fi_3, fi_5),(fi_4, fi_5)\}.$$

For each element of CA_2 fuzzy support and fuzzy correlation coefficient are calculated by using the Equation 5.

CA_2	$M(fi_1,fi_2)$	$M(fi_1,fi_3)$	$M(fi_1,fi_4)$	$M(fi_1,fi_5)$	$M(fi_2,fi_3)$	$M(fi_2,fi_4)$	$M(fi_2,fi_5)$	$M(fi_3,fi_4)$	$M(fi_3,fi_5)$	$M(fi_4,fi_5)$
1.	0.1	0.1	0.1	0.1	0.3	0.5	0.3	0.3	0.3	0.3
2.	0.2	0.2	0.2	0.2	0.4	0.8	0.4	0.4	0.4	0.4
3.	0.2	0.1	0.2	0.2	0.1	0.3	0.5	0.1	0.1	0.3
4.	0.4	0.4	0.1	0.1	0.5	0.1	0.1	0.1	0.1	0.1
5.	0.3	0.4	0.7	0.2	0.3	0.3	0.2	0.4	0.2	0.2
6.	0.5	0.5	0.5	0.5	0.6	0.6	0.6	0.7	0.7	0.8
7.	0.1	0.8	0.4	0.8	0.1	0.1	0.1	0.4	0.9	0.4
8.	0.1	0.5	0.2	0.7	0.1	0.1	0.1	0.2	0.5	0.2
9.	0.4	0.3	0.5	0.6	0.3	0.4	0.4	0.3	0.3	0.5
10.	0.2	0.2	0.2	0.2	0.2	0.4	0.5	0.2	0.2	0.4
fsupp	0.25	0.35	0.31	0.36	0.29	0.36	0.32	0.31	0.37	0.36
Cor A,B	-0.91	0.54	0.01	0.44	-0.56	0.25	-0.41	0.08	0.43	0.10

Figure 4. Dataset with fuzzy support and fuzzy correlation coefficient of CA_2.

Table 3. Fuzzy correlation computation method.

Column number	Col. 1	Col. 2	Col. 3	Col. 4	Col. 5
T	$fi_1\mu_A$	$fi_2\mu_B$	$S_{A,B}$ (each item)	S_A^2	S_B^2
T1	0.1	0.9	-0.1404	0.1296	0.1521
T2	0.2	0.8	-0.0754	0.0676	0.0841
T3	0.2	0.7	-0.0494	0.0676	0.0361
T4	0.4	0.5	0.0006	0.0036	0.0001
T5	0.7	0.3	-0.0504	0.0576	0.0441
T6	0.5	0.6	0.0036	0.0016	0.0081
T7	0.8	0.1	-0.1394	0.1156	0.1681
T8	0.7	0.1	-0.0984	0.0576	0.1681
T9	0.8	0.4	-0.0374	0.1156	0.0121
T10	0.2	0.7	-0.0494	0.0676	0.0361
F Support	0.46	0.51			
$S_{A,B}$			-0.070666667		
Sum of col 4/9 and Col 5/9				0.076	0.078778
$Cor_{A,B}$					-0.91328

As shown in Table 3, for all elements fuzzy correlation coefficient and fuzzy support are calculated.

After performing these steps, an element whose fuzzy support is greater or equal to s_f (specified threshold i.e. 0.30) and α (Specified Fuzzy Correlation Coefficient i.e. 0.40) will be an element of S_2, so $S_2 = \{(\{fi_1\},\{fi_3\}),(\{fi_1\},\{fi_5\}),(\{fi_3,\},\{fi_5\})\}$

Similarly again the fuzzy support and fuzzy correlation coefficient of each element of CA_3 is calculated. CA_3 is shown in Table 4.

Table 4. Fuzzy Support and Fuzzy Correlation Coefficient of CA3.

$(\{fi_1\},(\{fi_3, fi_5\})$	$(\{fi_3\}),(\{fi_1, fi_5\})$	$(\{fi_5\},\{fi_1, fi_3\})$
0.30	0.30	0.30
0.47	0.57	0.55

After performing all these steps we concluded that interesting item sets are *fi1, fi3* and *fi* and we mark them as sensitive fuzzy association rules. We modify the values in these sensitive fuzzy items. In modification step, we replace highest value with 0. In the given example we replace value of *fi3, fi5* i.e. 0.9 and with 0 in T7. Now the support of sensitive fuzzy association rules is 0.27, 0.27, 0.28, which is below the specified threshold, so the fuzzy rules that generate with the combination of *fi3 and fi5* automatically hide i.e. when we apply s_f, these rules will not appear in the results. In table 5 we have shown the modified dataset. The value of *fi3* in T7 is 0 now.

Table 5. Modified dataset.

T	fi_1	fi_2	fi_3	fi_4	fi_5
T1	0.1	0.9	0.3	0.5	0.3
T2	0.2	0.8	0.4	0.8	0.4
T3	0.2	0.7	0.1	0.3	0.5
T4	0.4	0.5	0.6	0.1	0.1
T5	0.7	0.3	0.4	0.8	0.2
T6	0.5	0.6	0.7	0.9	0.8
T7	0.8	0.1	0	0.4	0.9
T8	0.7	0.1	0.5	0.2	0.7
T9	0.8	0.4	0.3	0.5	0.6
T10	0.2	0.7	0.2	0.4	0.5
FSUPPORT	0.46	0.51	0.27	0.49	0.50

6. Experimental Results and Analysis

In this section, analysis of PPFAR scheme is presented. Three experiments were performed for analyzing performance of PPFAR method by using Wisconsin Breast Cancer dataset of UCI Machine Learning Repository. The dataset contains 699 transactions. Comparison of proposed scheme is done with datasets given by [5, 11, 12]. Following criteria was used to measure the performance of PPFAR; 1) Count of ghost rules created; 2) Count of modification in fuzzy transactions; 3) Lost rules' count in the modification process. Figure 5 presents relationship between count of ghost rules generated and number of transactions. The results are depicted in Figure 5. It is clear from the figure that PPFAR scheme outperforms existing techniques e.g. [5, 12] in generation of ghost rules (new rules generated after modification process) i.e. minimum ghost rules are generated. The number of ghost rules generated is based upon the whole dataset. These rules should be minimum after modifications made in transactions.

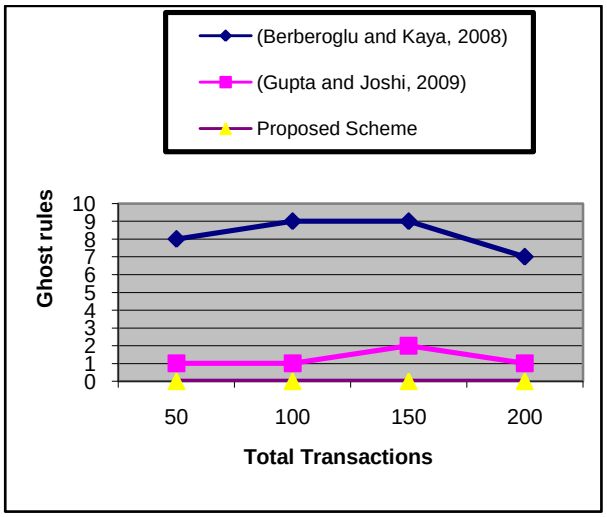

Figure 5. Comparison of new rules generated with existing techniques.

Figure 6 represents association between total modifications after sanitization and total number of transactions. Experiments were performed on three datasets. The first one contains 5 transactions having five fuzzy items. We applied fuzzy support and fuzzy correlation analysis and found two items that were sensitive. To hide sensitive items, we performed just one modification in the dataset and were able to accomplish complete sanitization. Second dataset consist of 10 transactions having 5 fuzzy items. We found three sensitive rules from this dataset. We analyzed fuzzy items, which had high correlation with each other and had most frequent occurrence. We directly targeted the fuzzy items for modification. We achieved complete sensitive fuzzy rule hiding with only two modifications in this experiment. In experiment number 3, the dataset had 682 transactions, containing 9 quantitative attributes, fuzzified into three regions. We applied fuzzy support and correlation analysis and found 11 rules as sensitive. We achieved complete hiding with 235 modifications in the dataset. The results are depicted in Figure 6.

The basic rationale for lesser modifications in PPFAR scheme is twofold. First is that it takes into consideration only those fuzzy rules that have higher correlation with other fuzzy rules. Secondly it also looks for most occurrences. By using PPFAR scheme, modification is performed only in the above mentioned fuzzy items. Therefore, higher data quality is maintained in modified dataset by applying PPFAR scheme as compared to existing techniques [5, 12]. Existing techniques apply typical support and confidence criterion for selecting transactions and alter all of them for a rule, generating large number of side effects. It also decreases data quality of released dataset.

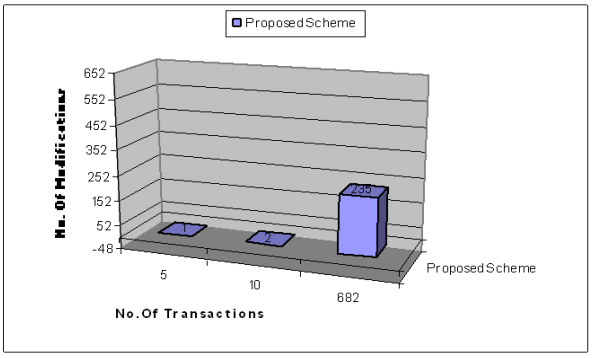

Figure 6. Comparison of No. of modifications with No. of Transactions.

Figure 7 shows the comparison between number of sensitive rules identified with total number of hidden fuzzy association rules. The results demonstrate that all the identified sensitive rules are hidden. Figure 8 shows factor of lost rules in proposed scheme. We observe that PPFAR Scheme results in less number of lost rules.

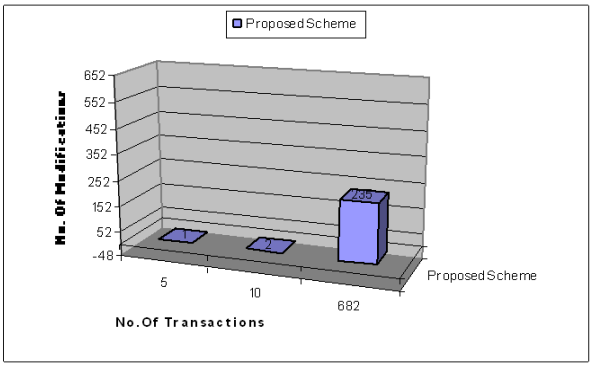

Figure 7. Comparison of sensitive rules identified with hidden rules.

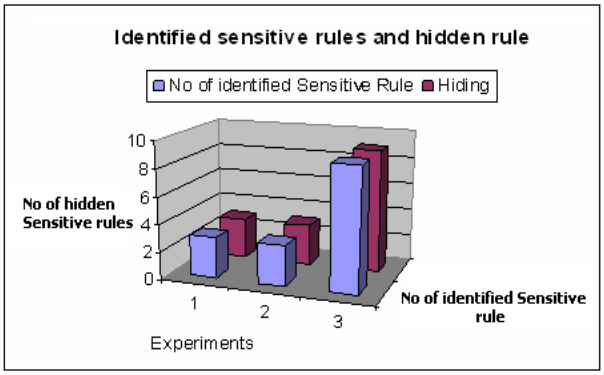

Figure 8. Comparison of Lost rules.

7. Conclusions and Future Work

We have highlighted a number of notable techniques proposed in the literature within the context of PPDM. Most of the techniques sanitize association rules in Boolean data. There are many scenarios in which data is not in Boolean form, where the quantity of data is more important. In this work, we have devised a new method for concealing sensitive fuzzy association rules named, Fuzzy Correlation based Scheme for PPDM with minimum side effects. We investigated that, by using fuzzy correlation coefficient analysis, we can get highly correlated fuzzy association rules. These rules can be considered sensitive. For hiding purpose, we apply modification technique in those fuzzy items, which frequently occur in different transactions. The results of fuzzy correlation based scheme (PPFAR) are comparatively better than the existing techniques. The primary objectives of PPFAR scheme are to completely hide sensitive fuzzy association rules with lesser modifications and generation of no ghost rules. Results show that PPFAR scheme has no side effects of ghost rules. In addition, PPFAR scheme requires minimum number of modifications for hiding than previous techniques. In future, we will investigate further similarity measures that can be used to find out the interesting association rules in quantitative data. We would also look at evolutionary approaches for sanitizing sensitive fuzzy association rules in the context of quantitative data.

References

[1] S. R. D. M. Oliveira, "Data Transformation for Privacy-Preserving Data Mining," Doctoral Dissertation, University of Alberta Edmonton, Alta (2005).

[2] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed.,The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd Ed., Vol. 2. Oxford: Clarendon (1892) pp.68–73.

[3] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules, Presented at the Proceedings of the 20th VLDB Conference, Santiago, Chile (1994).

[4] C. Clifton and D. Marks, Security and Privacy Implications of Data Mining, SIDMOD Workshop on Data Mining and Knowledge Discovery (1996).

[5] M. Gupta and R. C. Joshi, International Journal of Computer Theory and Engineering **1**, No. 4, (2009) 1793.

[6] Computational Intelligence: Principles, Techniques and Applications, Prof. Dr Amit Konar, Springer, The Netherlands, ISBN: 3-540-20898-4 (2005) pages 705.

[7] R. Agrawal and R. Srikant, "Privacy preserving data mining.," presented at the Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, New York, NY, USA (2000).

[8] M. Atallah et al., Disclosure Limitation of Sensitive Rules, Presented at the Knowledge and Data Engineering Exchange Chicago, IL, USA (1999).

[9] Y. Saygin et al., Using Unknowns to Prevent Discovery of Association Rules, Presented at the ACM SIGMOD Record, New York, USA (2001).

[10] M. Naeem and S. Asghar, A Novel Architecture for Hiding Sensitive Association Rules, Proceedings of the International Conference on Data Mining, Las Vegas, Nevada, USA (July 12-15, 2010) CSREA Press 2010. ISBN: 1-60132-138-4, Robert Stahlbock and Sven Crone (Eds.)

[11] N. P. Lin and H.-E. Chueh, Fuzzy Correlation Rules Mining, Presented at the 6th WSEAS International Conference on Applied Computer Science, Hangzhou, China (2007).

[12] T. Berberoglu and M. Kaya, Hiding Fuzzy Association Rules in Quantitative Data, Presented at the 3rd International Conference on Grid and Pervasive Computing Workshops, Kunming, China. (May 25-28, 2008)**.**

[13] E. Dasseni et al., Hiding Association Rules by Using Confidence and Support, Presented at the 4th Information Hiding Workshop, Pittsburg, PA, USA (2001).

[14] S. R. M. Oliveira and O. R. Zaiane, Privacy Preserving Frequent Itemset Mining, Presented at the Workshop on Privacy, Security and Data Mining, Maebashi City, Japan (December 2002).

[15] D. Agrawal and C. C. Aggarwal, On the Design and Quantification of Privacy Preserving Data Mining Algorithms, Presented at the Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, New York, NY, USA (2001).

[16] S. J. Rizvi and J. R. Haritsa, Maintaining Data Privacy in Association Rule Mining, Presented at the 28th International Conference on Very Large Databases, Hong Kong, China (2002).

[17] D. Toshniwal and M. Verma, A Border-Based Approach for Hiding Fuzzy Weighted Sensitve Itemsets, Presented at the International Conference on DMIN'10, Las Vegas, Nevada,USA, (2010).

[18] UCI Machine Learning Repository, Breast Cancer Wisconsin, http://archive.ics.uci.edu/ml/.