On Addressing Accuracy Concerns in Privacy Preserving Association Rule Mining

Ling Guo, Songtao Guo, and Xintao Wu

University of North Carolina at Charlotte
{lguo2,sguo,xwu}@uncc.edu

Abstract. Randomized Response techniques have been empirically investigated in privacy preserving association rule mining. In this paper, we investigate the accuracy (in terms of bias and variance of estimates) of both support and confidence estimates of association rules derived from the randomized data. We demonstrate that providing confidence on data mining results from randomized data is significant to data miners. We propose the novel idea of using interquantile range to bound those estimates derived from the randomized market basket data. The performance is evaluated using both representative real and synthetic data sets.

1 Introduction

Privacy is becoming an increasingly important issue in many data mining applications. A considerable amount of work on privacy preserving data mining [2,1,11,10] has been investigated recently. Among them, randomization has been a primary tool to hide sensitive private data for privacy preserving data mining. The issue of maintaining privacy in association rule mining has attracted considerable attention in recent years [7,8,4,13]. Most of techniques are based on a data perturbation or Randomized Response (RR) approach [5], wherein the 0 or 1 (0 denotes absence of an item while 1 denotes presence of an item) in the original user transaction vector is distorted in a probabilistic manner that is disclosed to data miners.

In [13,4,3], the authors proposed the MASK technique to preserve privacy for frequent itemset mining and addressed the issue of providing efficiency in calculating the estimated support values. Their results empirically showed a high degree of privacy to users and a high level of accuracy in the mining results can be simultaneously achieved. To evaluate the privacy, they defined a privacy metric and presented an analytical formula for evaluating the privacy obtained under the metric. However, accuracy metric on data mining results was only defined in an aggregate manner as support error and identity error computed over all discovered frequent itemsets.

Our paper moves one step further to address the issue of providing accuracy in privacy preserving mining of association rules. We investigate the issue of how the accuracy (i.e., support and confidence) of each association rule mined from randomized data is affected when the randomized response technique is applied.

Specifically, we present an analytical formula for evaluating the accuracy (in terms of bias and variance of estimates) of both support and confidence measures of association rules derived from the randomized data. From the derived bias and variance of estimates, we further derive approximate interquantile ranges. Data miners are ensured

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 124–135, 2008.

that their estimates lie within these ranges with a high confidence, say 95%. We would emphasize that providing confidence on estimated data mining results is significant to data miners since they can learn how accurate their reconstructed results are. We illustrate the importance of those estimated interquantile ranges using an example.

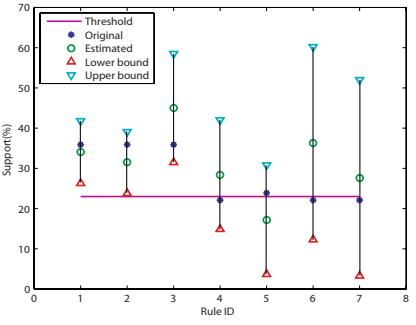


Fig. 1. Accuracy of the estimated support values of association rules derived from randomized data with p=0.65

Figure 1 shows the original support values, the estimated support values from the randomized data, and their corresponding 95% interquantile ranges of 7 association rules, which were derived from COIL data sets[1]. A distortion parameter $p = 0.65$ and support threshold sup_{min} = 23% were used in the experiment. The interquantile range of each rule can give data miners confidence about their estimate derived from randomized data. For example, the estimated support of rule 2 is 31.5% and its 95% interquantile range is [23.8%,39.1%], which suggests the original support value lies in this range with 95% probability. Furthermore, we can observe the 95% interquantile ranges for rules 1-3 are above the support threshold, which guarantees those are true frequent itemsets (with at least 95% confidence).

We emphasize providing accuracy of data mining results is important for data miners during data exploration. When the support threshold is set as 23%, we may not only take rule 2 and 6 as frequent sets from the estimated support values, but also conclude rule 6 (35.9%) is more frequent than rule 2 (31.5%). However, rule 2 has the original support as 36.3% while rule 6 has the original support as 22.1%, we mistakenly assign the infrequent itemset 6 as frequent. By using the derived interquantile ranges, we can determine that rule 2 is frequent with high confidence (since its lower bound 23.8% is above the support threshold) and rule 6 may be infrequent (since its lower bound 12.3% is below the support threshold).

The remainder of this paper is organized as follows. In Section 2, we present the distortion framework and discuss how the Randomized Response techniques are applied to privacy preserving market association rule mining. We conduct the theoretical analysis on how distortion process affects the accuracy of both support and confidence values derived from the randomized data in Section 3. In Section 4, empirical evaluations on various datasets are given. We conclude our work in Section 5.

[1] http://kdd.ics.uci.edu/databases/tic/tic.html

2 Distortion Framework

2.1 Association Rule Revisited

Denoting the set of transactions in the database D by $\mathcal{T} = \{T_1, \cdots, T_n\}$ and the set of items in the database by $\mathcal{I} = \{A_1, \cdots, A_m\}$. An association rule $\mathcal{X} \Rightarrow \mathcal{Y}$, where $\mathcal{X}, \mathcal{Y} \subset \mathcal{I}$ and $\mathcal{X} \cap \mathcal{Y} = \phi$, has two measures: the support s defined as the $s(100\%)$ of the transactions in $\mathcal{T}$ contain $\mathcal{X} \cup \mathcal{Y}$, and the confidence c is defined as $c(100\%)$ of the transactions in $\mathcal{T}$ that contain $\mathcal{X}$ also contain $\mathcal{Y}$.

2.2 Randomization Procedure

Let there be m sensitive items $A_1, A_2, \cdots, A_m$, each being considered as one dichotomous variable with 2 mutually exclusive and exhaustive categories (0 = absence, 1 = presence). One transaction can be logically translated as a fixed-length sequence of 0's and 1's. For each transaction, we apply the Warner RR model [15] independently on each item using different settings of distortion. If the original value is in the *absence*(*presence*) category, it will be kept in such category with a probability θ_0 (θ_1) and changed to *presence*(*absence*) category with a probability $1 - \theta_0$ ($1 - \theta_1$). For item A_j, the distortion probability matrix P_j generally takes the form $P_j = \begin{pmatrix} \theta_0 & 1-\theta_1 \\ 1-\theta_0 & \theta_1 \end{pmatrix}$. In this paper, we follow the original Warner RR model by setting $\theta_0 = \theta_1 = p_j$. This setting indicates users have the same level of privacy for both 1's and 0's. In general customers may expect more privacy for their 1's than for their 0's, since the 1's denote specific actions whereas the 0's are the default options.

Denote $\pi^{(j)} = (\pi_0^{(j)}, \pi_1^{(j)})'$ ($\lambda^{(j)} = (\lambda_0^{(j)}, \lambda_1^{(j)})'$) as the vector of marginal proportions corresponding to item A_j in the original (randomized) data set, where $j = 1, \cdots, m$. We have

$$\lambda^{(j)} = P_j \pi^{(j)} \tag{1}$$

Note that each vector $\pi^{(j)}$ has two values $\pi_0^{(j)}, \pi_1^{(j)}$ and the latter corresponds to the support value of item A_j. For a market data set with n transactions, let $\hat{\lambda}^{(j)}$ be the vector of sample proportions corresponding to $\lambda^{(j)}$. Then an unbiased estimate of $\pi^{(j)}$ is $\hat{\pi}^{(j)} = P_j^{-1} \hat{\lambda}^{(j)}$.

2.3 Estimating k-Itemset Supports

We can easily extend Equation 1, which is applicable to one individual item, to compute the support of an arbitrary k-itemset. For simplicity, let us assume that we would compute the support of an itemset which contains the first k items $\{A_1, \cdots, A_k\}$ (The general case with any k items is quite straightforward but algebraically messy).

Let $\pi_{i_1, \cdots, i_k}$ denote the true proportion corresponding to the categorical combination $(A_{1i_1}, \cdots, A_{ki_k})$, where $i_1, \cdots, i_k \in \{0, 1\}$. Let π be vectors with elements $\pi_{i_1, \cdots, i_k}$ arranged in a fixed order. The combination vector corresponds to a fixed order of cell entries in the contingency table formed by the k-itemset. When we have k items, the number of cells in the k-dimensional contingency table is 2^k. Table 1(a) shows one

Table 1. 2×2 contingency tables for two variables A,B

(a) Original

	$\bar{B}$	B	
$\bar{A}$	π_{00}	π_{01}	π_{0+}
A	π_{10}	π_{11}	π_{1+}
	π_{+0}	π_{+1}	π_{++}

(b) After randomization

	$\bar{B}$	B	
$\bar{A}$	λ_{00}	λ_{01}	λ_{0+}
A	λ_{10}	λ_{11}	λ_{1+}
	λ_{+0}	λ_{+1}	λ_{++}

contingency table for a pair of two variables. We use the notation $\bar{A}$ ($\bar{B}$) to indicate that A (B) is absent from a transaction. The vector $\pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})'$ corresponds to a fixed order of cell entries π_{ij} in the 2×2 contingency table. π_{11} denotes the proportion of transactions which contain both A and B while π_{10} denotes the proportion of transactions which contain A but not B. The row sum π_{1+} represents the support frequency of item A while the column sum π_{+1} represents the support frequency of item B.

The original database D is changed to D_{ran} after randomization. Assume $\lambda_{\mu_1,\cdots,\mu_k}$ is the probability of getting a response $(\mu_1, \cdots, \mu_k)$ and λ the vector with elements $\lambda_{\mu_1,\cdots,\mu_k}$ arranged in a fixed order (e.g., the vector $\lambda = (\lambda_{00}, \lambda_{01}, \lambda_{10}, \lambda_{11})'$ corresponds to cell entries λ_{ij} in the randomized contingency table as shown in Table 1(b)), we can obtain

$$\lambda = (P_1 \times \cdots \times P_k)\pi$$

where $\times$ stands for the Kronecker product.

Let $P = P_1 \times \cdots \times P_k$, an unbiased estimate of π follows as

$$\hat{\pi} = P^{-1}\hat{\lambda} = (P_1^{-1} \times \cdots \times P_k^{-1})\hat{\lambda} \tag{2}$$

where $\hat{\lambda}$ is the vector of sample proportions corresponding to λ and P_j^{-1} denotes the inverse of the matrix P_j. Note that although the distortion matrices $P_1, \cdots, P_k$ are known, they can only be utilized to estimate the proportions of itemsets of the original data, rather than precisely reconstruct the original 0-1 data.

In this paper we follow the Moment Estimation method as shown in Equation 2 to get the unbiased estimate of the distribution for original data. This method has been broadly adopted in the scenarios where RR is used to perturb data for preserving privacy. Although it has good properties as computational simplicity and unbiasedness, some awkward property exists due to random errors [5,6]. That is, the estimate may fall out of the parameter space, which makes the estimate meaningless. This is one reason that Maximum Likelihood Estimation (MLE) is adopted to estimate the distribution in literature [6].

It has been proved in [6] that a good relation holds between these two methods in the scenarios of RR: The moment estimate is equal to the MLE estimate within parameter space. Based on that, we can know that moment estimate from Equation 2 achieves the Cramér-Rao bound as MLE does. Therefore, moment estimate is the minimum variance unbiased (MVU) estimator in RR contexts. Our later analysis on accuracy of association rule is based on such unbiased estimate under the assumption that the estimate is within parameter space.

3 Theoretical Analysis on Accuracy of Association Rule

In this section, we theoretically analyze the variance of the estimates of both s and c for any individual association rule $\mathcal{X} \Rightarrow \mathcal{Y}$. To derive their interquantile ranges, we also analyze the distributions of those estimates derived from the randomized data.

3.1 Accuracy on Support s

From Equation 2, we know how to derive the estimate of support values of any itemset from the observed randomized data. Now we address the question how accurate the estimated support value is.

The whole contingency table is usually modeled as a multinomial distribution in statistics. When we have k items, the number of cells in the contingency table is 2^k. For each cell d, where $d = 1, 2, \cdots, 2^k$, it has a separate binomial distribution with parameters n and η_i. The binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent 0/1 experiments, each of which yields success with probability η_i. When n is large enough (one rule of thumb is that both $n\eta_i$ and $n(1-\eta_i)$ must be greater than 5), an approximation to $B(n, \eta_i)$ is given by the normal distribution $N(n\eta_i, n\eta_i(1-\eta_i))$.

Result 1. *Since each cell $\pi_{i_1,\cdots,i_k}$ approximately follows normal distribution, its $(1-\alpha)100\%$ interquantile range can be approximated as*

$$[\hat{\pi}_{i_1\cdots i_k} - z_{\alpha/2} * \sqrt{\hat{v}ar(\hat{\pi}_{i_1\cdots i_k})}, \hat{\pi}_{i_1\cdots i_k} + z_{\alpha/2} * \sqrt{\hat{v}ar(\hat{\pi}_{i_1\cdots i_k})}]$$

$z_{\alpha/2}$ is the upper $\alpha/2$ critical value for the standard normal distribution.

$\hat{v}ar(\hat{\pi}_{i_1\cdots i_k})$ can be derived from the covariance matrix [5]:

$$\begin{aligned} \hat{cov}(\hat{\pi}) &= \Sigma_1 + \Sigma_2 \\ &= (n-1)^{-1}(\hat{\pi}^{\delta} - \hat{\pi}\hat{\pi}^{'}) + (n-1)^{-1}P^{-1}(\hat{\lambda}^{\delta} - P\hat{\pi}^{\delta}P^{'})P^{'-1} \end{aligned}$$

Note that Σ_1 is the dispersion matrix of the direct estimator of π, which is only related to the data size for estimation. While the data size is usually large in most market basket analysis scenarios, it can be neglected. Σ_2 represents the component of dispersion associated with RR distortion.

We can simply use the derived $\hat{\pi}_{i_1\cdots i_m}$ (from Equation 2) as an estimate of μ and the derived $\sqrt{\hat{v}ar(\hat{\pi}_{i_1\cdots i_m})}$ as an estimate of σ, where μ and σ are unknown parameters of the normal distribution of each cell. An $(1-\alpha)100\%$ interquantile range, say $\alpha = 0.05$, shows the interval contains the original $\pi_{i_1,\cdots,i_m}$ with 95% probability.

To illustrate this result, we use a simple example $G \Rightarrow H$ (rule 2 in Figure 1). The proportion of itemsets of the original data is given as

$$\pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})' = (0.415, 0.043, 0.183, 0.359)'$$

Using the RR scheme presented in the previous section, with the distortion parameters $p_1 = p_2 = 0.9$, we get the randomized responses

$$\hat{\lambda} = (0.368, 0.097, 0.218, 0.316)'$$

By applying Equation 2, we derive the unbiased estimate of π as

$$\hat{\pi} = (0.427, 0.031, 0.181, 0.362)'$$

The covariance matrix of $\hat{\pi}$ is unbiasedly estimated as

$$\hat{cov}(\hat{\pi}) = \begin{bmatrix} 7.113 & -1.668 & -3.134 & -2.311 \\ -1.668 & 2.902 & 0.244 & -1.478 \\ -3.134 & 0.244 & 5.667 & -2.777 \\ -2.311 & -1.478 & -2.777 & 6.566 \end{bmatrix} \times 10^{-5}$$

The diagonal elements of the above matrix represent the variances of the estimated $\hat{\pi}$, e.g., $\hat{var}(\hat{\pi}_{00}) = 7.113 \times 10^{-5}$ and $\hat{var}(\hat{\pi}_{11}) = 6.566 \times 10^{-5}$. Those off-diagonal elements indicate the estimated covariances, e.g., $\hat{cov}(\hat{\pi}_{11}, \hat{\pi}_{10}) = -2.777 \times 10^{-5}$.

From Result 1, we can derive 95% interquantile range of s_{GH} as

$$[\hat{\pi}_{11} - z_{0.025}\sqrt{\hat{var}(\hat{\pi}_{11})}, \hat{\pi}_{11} + z_{0.025}\sqrt{\hat{var}(\hat{\pi}_{11})}] = [0.346, 0.378]$$

We can also see this derived interquantile range [0.346, 0.378] for rule 2 with $p_1 = p_2 = 0.9$ is shorter than [0.238, 0.391] with $p_1 = p_2 = 0.65$ as shown in Figure 1.

3.2 Accuracy on Confidence c

We first analyze the accuracy on confidence of a simple association rule $A \Rightarrow B$ where A and B are two single items which have 2 mutually exclusive and exhaustive categories. We denote s_A, s_B, and s_{AB} as the support values of A, B, and AB respectively. Accordingly, we denote $\hat{s}_A$, $\hat{s}_B$, and $\hat{s}_{AB}$ as the estimated support values from randomized data of A, B, and AB respectively.

Result 2. *The confidence (c) of a simple association rule $A \Rightarrow B$ has estimated value as*

$$\hat{c} = \frac{\hat{s}_{AB}}{\hat{s}_A} = \frac{\hat{\pi}_{11}}{\hat{\pi}_{1+}}$$

with the expectation of $\hat{c}$ approximated as

$$\hat{E}(\hat{c}) \approx \frac{\hat{\pi}_{11}}{\hat{\pi}_{1+}} + \frac{\hat{\pi}_{11}}{\hat{\pi}_{1+}^3}\hat{var}(\hat{\pi}_{10}) - \frac{\hat{\pi}_{10}}{\hat{\pi}_{1+}^3}\hat{var}(\hat{\pi}_{11}) + \frac{\hat{\pi}_{11} - \hat{\pi}_{10}}{\hat{\pi}_{1+}^3}\hat{cov}(\hat{\pi}_{11}, \hat{\pi}_{10}) \quad (3)$$

and the variance of $\hat{c}$ approximated as

$$\hat{var}(\hat{c}) \approx \frac{\hat{\pi}_{10}^2}{\hat{\pi}_{1+}^4}\hat{var}(\hat{\pi}_{11}) + \frac{\hat{\pi}_{11}^2}{\hat{\pi}_{1+}^4}\hat{var}(\hat{\pi}_{10}) - 2\frac{\hat{\pi}_{10}\hat{\pi}_{11}}{\hat{\pi}_{1+}^4}\hat{cov}(\hat{\pi}_{11}, \hat{\pi}_{10}) \quad (4)$$

according to the delta method [12].

Confidence can be regarded as a ratio (W) of two correlated normal random variables (X, Y), $W = X/Y$. However, it is hard to derive the critical value for the distribution of W from its cumulative density function $F(w)$ [14], we provide an approximate interquantile range of confidence based on Chebyshev's Inequality.

Theorem 1. *(Chebyshev's Inequality) For any random variable X with mean μ and variance σ^2*

$$Pr(|X - \mu| \geq k\sigma) \leq 1/k^2 \qquad k > 0$$

Chebyshev's Inequality gives a conservative estimate. It provides a lower bound to the proportion of measurements that are within a certain number of standard deviations from the mean.

Result 3. *The loose $(1-\alpha)100\%$ interquantile range of confidence (c) of $A \Rightarrow B$ can be approximated as*

$$[\hat{E}(\hat{c}) - \frac{1}{\sqrt{\alpha}}\sqrt{\hat{var}(\hat{c})}, \hat{E}(\hat{c}) + \frac{1}{\sqrt{\alpha}}\sqrt{\hat{var}(\hat{c})}]$$

From Chebyshev's Inequality, we know for any sample, at least $(1 - 1/k^2)$ of the observations in the data set fall within k standard deviations of the mean. When we set $\alpha = \frac{1}{k^2}$, we have $Pr(|X-\mu| \geq \frac{1}{\sqrt{\alpha}}\sigma) \leq \alpha$. Hence, $Pr(|X-\mu| \leq \frac{1}{\sqrt{\alpha}}\sigma) \geq 1-\alpha$. We can simply use the derived $\hat{E}(\hat{c})$ (from Equation 3) as an estimate of μ and the derived $\sqrt{\hat{var}(\hat{c})}$ (from Equation 4) as an estimate of σ, where μ and σ are unknown parameters of the distribution of confidence. An approximate $(1-\alpha)100\%$ interquantile range of confidence c is then derived.

All the above results can be straightforwardly extended to the general association rule $\mathcal{X} \Rightarrow \mathcal{Y}$ and further details can be found in [9].

4 Empirical Evaluation

In our experiments, we use the COIL Challenge 2000 which provides data from a real insurance business. Information about customers consists of 86 attributes and includes product usage data and socio-demographic data derived from zip area codes. The training set consists of 5822 descriptions of customers, including the information of whether or not they have a Caravan insurance policy. Our binary data is formed by collapsing non-binary categorical attributes into binary form (the data can be found at www.cs.uncc.edu/∼xwu/classify/b86.dat), with $n = 5822$ baskets and $m = 86$ binary items.

4.1 Accuracy of Individual Rule vs. Varying p

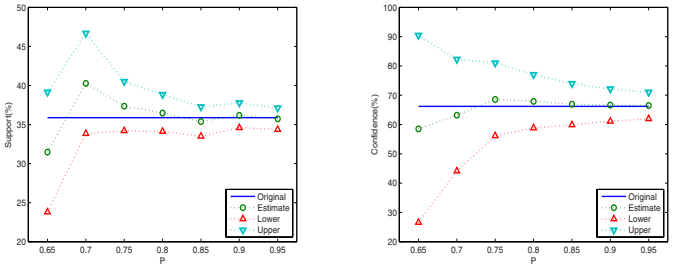
Table 2[2] shows the 7 randomly chosen association rules derived from the randomized COIL data with distortion parameter $p = 0.65$. In this table, s ($\hat{s}$) indicates the original (estimated) support value. $\hat{s}_l$ ($\hat{s}_u$) denotes the lower bound (upper bound) of the 95% interquantile range of the estimated support value. Similarly, c ($\hat{c}$) indicates the original (estimated) confidence value. $\hat{c}_l$ ($\hat{c}_u$) denotes the lower bound (upper bound) of the 95% estimated confidence value. We have shown how the accuracy of the estimated support values varies in Figure 1 (Section 1). One observation is that interquantile ranges of

[2] The meaning of these items can be found in Table 2 of [16].

Table 2. Accuracy of the estimated support and confidence for 7 representative rules of COIL

ID	$\mathcal{X}$	$\mathcal{Y}$	s	$\hat{s}$	$\hat{s}_l$	$\hat{s}_u$	c	$\hat{c}$	$\hat{c}_l$	$\hat{c}_u$
1	G	E	35.9	34.1	26.3	41.8	66.2	64.7	31.3	95.3
2	G	H	35.9	31.5	23.8	39.1	66.2	62.2	26.6	90.4
3	EH	G	35.8	45.0	31.5	58.5	89.3	77.5	33.5	100
4	EG	I	22.1	28.4	14.9	42.0	61.7	75.2	0	100
5	HF	I	23.9	17.2	3.7	30.8	100	91.0	0	100
6	EGH	F	22.1	36.3	12.3	60.2	61.7	99.4	0	100
7	FGI	E	22.1	27.6	3.32	52.0	77.9	86.3	0	100

confidence estimates are usually wider than that of support estimates. For example, the 95% interquantile range of the estimated confidence for rule 2 is $[26.6\%, 90.4\%]$, which is much wider than that of the estimated support $[23.8\%, 39.1\%]$. This is due to three reasons. First, we set the distortion parameter $p = 0.65$ which implies a relatively large noise (the perturbed data will be completely random when $p = 0.5$). Second, the variance of the ratio of two variables is usually larger than the variance of either single variable. Third, the estimated support can be modeled as one approximate normal distribution so we can use the tight interquantile range. On the contrary, we derive the loose interquantile range of confidence using the general Chebyshev's Theorem. We expect that the explicit form of the $F(w)$ distribution can significantly reduce this width. We will investigate the explicit form of the distribution of confidence and all other measures, e.g. correlation, lift, etc. to derive tight bounds in our future work.



(a) support (b) confidence

Fig. 2. Accuracy vs. varying p for rule $G \Rightarrow H$

Our next experiment shows how the derived estimates (support, confidence, and their corresponding interquantile ranges) of one individual rule vary with the distortion parameter p. We vary the distortion parameter p from 0.65 to 0.95. Figure 2(a) (2(b)) shows the accuracy of the estimated support (confidence) values with varied distortion p values for a particular rule $G \Rightarrow H$. As expected, the larger the p, the more accurate the estimate and the tighter the interquantile range is. It was empirically shown in

[13] that a distortion probability of $p = 0.9$ (equivalently $p = 0.1$) is ideally suited to provide both privacy and good data mining results for the sparse market basket data. We can observe from Figure 2(b) that the 95% interquantile range of the confidence estimate with $p \geq 0.9$ is tight.

4.2 Accuracy of All Rules vs. Varying p

The above study of the accuracy of the estimate in terms of each individual rule is based on the variance as criterion. In the case of all rules together, we can evaluate the overall accuracy of data mining results using the average support error, the average confidence error, percentage of false positives, percentage of false negatives etc. as defined in [4].

The metric $\rho = \frac{1}{|R|} \sum_{r \in R} \frac{|\hat{s_r} - s_r|}{s_r} \times 100$ represents the average relative error in the reconstructed support values for those rules that are correctly identified. The identity error σ reflects the percentage error in identifying association rules. $\sigma^+ = \frac{|R-F|}{|F|} \times 100$ indicates the percentage of false positives and $\sigma^- = \frac{|F-R|}{|F|} \times 100$ indicates the percentage of false negatives where R (F) denotes the reconstructed (actual) set of association rules. In addition to the support error (ρ) and the identity error (σ^+, σ^-), we define the following three measures.

- γ: the confidence error $\gamma = \frac{1}{|R|} \sum_{r \in R} \frac{|\hat{c}_r - c_r|}{c_r} \times 100$ represents the average relative error in the reconstructed confidence values for those rules that are correctly identified.
- s-p: the number of pairs of conflict support estimates. We consider $\hat{s}_1, \hat{s}_2$ as a pair of conflict estimates if $\hat{s}_1 < \hat{s}_2$ but $s_1 > \hat{s}_{1l} > s_{min} > s_2$ where $\hat{s}_{1l}$ denotes the lower bound of interqunatile range for s_1.
- c-p: the number of pairs of conflict confidence estimates (similarly defined as the above s-p).

Errors in support estimation due to the distortion procedure can result in falsely identified frequent itemsets. This becomes especially an issue when the support threshold setting is such that the support of a number of frequent itemsets lie very close to this threshold value (s_{min}). Such border-line itemsets can cause many *false positives* and *false negatives*. Even worse, an error in identifying a frequent itemset correctly in early passes has a ripple effect in terms of causing errors in later passes.

Table 3(a) shows how the above measures are varied by changing distortion parameter p from 0.65 to 0.95. We can observe all measures (the support error ρ, the confidence error γ, the false positives σ^+, the false negatives σ^-) decrease when p increases. The number of conflict support pairs (s-p) and conflict confidence pairs (c-p) also have the same trend. Our experiment shows that when $p \geq 0.85$, there are no or very few conflict support (confidence) pairs, which implies the reconstructed set of association rules is close to the original set. However, when $p \leq 0.80$, there are significant number of conflict pairs, which implies the reconstructed set may be quite different from the original one. By incorporating the derived interquantile range for each estimate, we can decrease the error caused by conflict pairs. In Section 1, we have shown one conflict support pair: rule 2 and rule 6. We can see that $\hat{s}_2 < \hat{s}_6$ (but $s_2 > s_6$). As $\hat{s}_{2l} > s_{min}$ and

Table 3. $sup_{min} = 25\%$ $conf_{min} = 65\%$ for COIL

(a)

p	ρ	σ^-	σ^+	s-p	γ	c-p
0.65	25.6	34.0	53.8	27817	9.90	737
0.70	12.3	21.2	38.1	4803	6.39	393
0.75	7.35	11.8	30.8	729	4.44	85
0.80	3.64	6.82	16.9	0	2.47	28
0.85	2.64	6.67	7.76	0	1.76	0
0.90	1.91	5.18	4.24	0	1.10	0
0.95	0.84	4.63	1.02	0	0.51	0

(b)

p	σ^-	σ_l^-	σ_u^-	σ^+	σ_l^+	σ_u^+
0.65	34.0	98.8	1.25	53.8	0.00	110.7
0.70	21.2	90.9	0.08	38.1	0.08	105.7
0.75	11.8	66.3	0.00	30.8	1.18	96.5
0.80	6.82	50.7	0.31	16.9	0.24	80.9
0.85	6.67	37.7	0.00	7.76	0.55	53.0
0.90	5.18	31.8	0.00	4.24	0.00	35.0
0.95	4.63	26.8	0.00	1.02	0.00	25.7

$\hat{s}_{6l} < s_{min}$, data miners can safely determine rule 2 is frequent but rule 6 may be infrequent. We would emphasize again that providing estimates together with their interquantile ranges (especially for those conflict pairs) through some visualization is very useful for data exploration tasks conducted on the randomized data.

Table 3(b) shows the comparison between the identity errors derived using lower bound and upper bound respectively. We define $\sigma_l^+ = \frac{|R_l - F|}{|F|} \times 100$ ($\sigma_u^+ = \frac{|R_u - F|}{|F|} \times 100$) as the false positives calculated from R_l (R_u) where R_l (R_u) denotes the reconstructed set of association rules using lower (upper) bound of interquantile range respectively. Similarly we define σ_l^- and σ_u^-. We can observe from Table 3(b) that σ_u^- is significantly lower than σ_- while σ_l^+ is significantly lower than σ_+. In other words, using the upper bound of the derived interquantile range can decrease the false negatives while using the lower bound can decrease the false positives. In some scenario, we may emphasize more on decreasing the false positive error. Hence, we can use the lower bound of the derived interquantile range, rather than the estimated value, to determine whether the set is frequent or not (i.e., frequent only if $\hat{s}_l \geq s_{min}$, infrequent otherwise).

4.3 Other Datasets

Since the COIL Challenge data is very sparse (5822 tuples with 86 attributes), we also conducted evaluations on the following representative databases used for association rule mining.

1. BMS-WebView-1[3]. Each transaction in the data set is a web session consisting of all the product detail pages viewed in that session. There are about 60,000 transactions with close 500 items.
2. A synthetic database generated from the IBM Almaden market basket data generator with parameters T10.I4.D0.1M.N0.1K., resulting in 10k customer tuples with each customer purchasing about ten items on average.

Tables 4 and 5 show our results on these two data sets respectively. We can observe similar patterns as shown in COIL data set.

[3] http://www.ecn.purdue.edu/KDDCUP

Table 4. $sup_{min} = 0.20\%$ $conf_{min} = 20\%$ for BMS-WebView-1

(a)

p	ρ	σ^-	σ^+	s-p	γ	c-p
0.65	362.4	64.1	80.6	632	114.7	11
0.75	72.9	39.9	68.7	418	57.9	2
0.85	19.5	27.9	54.0	67	24.5	0
0.95	5.47	9.66	16.5	56	7.23	0

(b)

p	σ^-	σ_l^-	σ_u^-	σ^+	σ_l^+	σ_u^+
0.65	63.9	100.0	1.34	81.8	0.0	187.6
0.75	40.1	100.0	1.07	69.8	0.0	155.3
0.85	27.9	99.1	0.40	54.0	0.0	152.8
0.95	9.66	70.6	0.00	16.5	0.0	123.8

Table 5. $sup_{min} = 0.20\%$ $conf_{min} = 60\%$ for IBM data with T10.I4.D0.1M.N0.1K

(a)

p	ρ	σ^-	σ^+	s-p	γ	c-p
0.65	1234.9	73.4	171.9	971	47.8	7
0.75	99.7	57.8	168.0	11	38.3	0
0.85	19.9	49.7	165.6	3	18.6	0
0.95	5.14	21.3	50.3	0	4.61	0

(b)

p	σ^-	σ_l^-	σ_u^-	σ^+	σ_l^+	σ_u^+
0.65	73.7	100.0	2.99	172.8	0.0	722.5
0.75	57.8	100.0	1.20	167.9	0.0	674.3
0.85	49.7	100.0	0.90	165.6	0.0	673.4
0.95	21.3	99.7	0.00	50.3	0.0	460.8

5 Conclusion and Future Work

In this paper, we have considered the issue of providing confidence ranges of support and confidence in privacy preserving association rule mining. Providing the accuracy of discovered patterns from randomized data is important for data miners. To the best of our knowledge, this has not been previously explored in the context of privacy preserving data mining.

Randomization still runs certain risk of disclosures. It was observed as a general phenomenon that maintenance of item privacy and precise estimation were in conflict. We will investigate how to determine distortion parameters optimally to satisfy both privacy and accuracy constraints. We will explore some scenario where some sensitive items are randomized while the remaining are released directly or where some transactions are randomized while the remaining are unperturbed. We also plan to investigate the extension of our results to generalized and quantitative association rules.

Acknowledgments

This work was supported in part by U.S. National Science Foundation IIS-0546027.

References

1. Agrawal, D., Agrawal, C.: On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of the 20th Symposium on Principles of Database Systems (2001)
2. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Dallas, Texas, May 2000, pp. 439–450 (2000)

3. Agrawal, S., Haritsa, J.: A framework for high-accuracy privacy-preserving mining. In: Proceedings of the 21st IEEE International Conference on Data Engineering, pp. 193–204 (2005)
4. Agrawal, S., Krishnan, V., Haritsa, J.: On addressing efficiency concerns in privacy-preserving mining. In: Lee, Y., Li, J., Whang, K.-Y., Lee, D. (eds.) DASFAA 2004. LNCS, vol. 2973, pp. 113–124. Springer, Heidelberg (2004)
5. Chaudhuri, A., Mukerjee, R.: Randomized Response Theory and Techniques. Marcel Dekker, New York (1988)
6. den Hout, A.V., der Heijden, P.G.M.V.: Randomized response, statistical disclosure control and misclassification: A review. International Statistical Review 70(2), 269–288 (2002)
7. Evfimievski, A., Gehrke, J., Srikant, R.: Limiting privacy breaches in privacy preserving data mining. In: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 211–222 (2003)
8. Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J.: Privacy preserving mining of association rules. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 217–228 (2002)
9. Guo, L., Guo, S., Wu, X.: On addressing accuracy concerns in privacy preserving association rule mining. Technical Report, CS Dept., UNC Charlotte (March 2007)
10. Huang, Z., Du, W., Chen, B.: Deriving private information from randomized data. In: Proceedings of the ACM SIGMOD Conference on Management of Data, Baltimore, MA (2005)
11. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On the privacy preserving properties of random data perturbation techniques. In: Proceedings of the 3rd International Conference on Data Mining, pp. 99–106 (2003)
12. Kendall, M.G., Stuart, A.: The advanced theory of statistics, vol. 1. Hafner Pub. Co., New York (1969)
13. Rizvi, S., Haritsa, J.: Maintaining data privacy in association rule mining. In: Proceedings of the 28th International Conference on Very Large Data Bases (2002)
14. Springer, M.D.: The Algebra of Random Variables. John Wiley and Sons, New York (1979)
15. Warner, S.: Randomized response: a survey technique for eliminating evasive answer bias. J. Amer. Statist. Assoc. 60, 63–69 (1965)
16. Wu, X., Barbará, D., Ye, Y.: Screening and interpreting multi-item associations based on log-linear modeling. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washinton, August 2003, pp. 276–285 (2003)