T-79.514 Special Course on Cryptology / Privacy-Preserving Data Mining: Database randomization via RRT

Isto Niemi

November 18, 2003

1 Introduction

This survey is a review of randomized response techniques for privacy-preserving data mining as described in the papers "Using Randomized Response Techniques for Privacy-Preserving Data Mining" by W. Du, and Z. Zhan [1] and "Privacy Preserving Mining of Association Rule" by A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke.s [2].

There are two different approaches presented in the papers. One for the association rule mining and one for the data classification.

It is really easy to build up a survey utilizing Internet. Because of the privacy issues the accuracy of the collected data are not as good as it could be. People try to protect themselves from email spam by lying to the surveys.

The basic idea of the randomized response is to randomize the answers so that the true value cannot be estimated with sufficient precision but the aggregate information can still be utilized.

Randomized Response Technique is shortly presented in Section 2. Association rule mining based on [2] is reviewed in Section 3. In Section 4 the decision tree building is described based on [2]. In Section 5 short summary is presented.

2 Randomized Responses

Both papers are based on the Randomized response technique developed by Warner in 1965 [5]. The technique tries to solve the problem where responder has attribute A but he dares not share it out to the interviewer.

The technique consists of two models called Related-Question Model and Unrelated-Question Model. In both models two questions are asked from the responder instead of one. In Related-Question Model the questions are related so that the answers are opposite to each other. Questions can be:

1. Have you ever got the sensitive attribute A?
2. Have you never got the sensitive attribute A?

The Responder uses a randomizing device to decide which question to answer without letting the interviewer to know which question is answered. Let's say that the probability of choosing the first question is θ and the second one is $1-\theta$.

To estimate the proportion of people who have the attribute A, we can set the following equations

$$P^*(A = yes) = P(A = yes) \cdot \theta + P(A = no) \cdot (1 - \theta) \quad (1)$$
$$P^*(A = no) = P(A = no) \cdot \theta + P(A = yes) \cdot (1 - \theta) \quad (2)$$

where $P^*(A = yes)$ is the proportion of the 'yes' obtained from the disguised data. $P(A = yes)$ is the estimated proportion of the 'yes' in actually undisguised data. Similarly $P^*(A = no)$ and $P(A = no)$ are the proportions of the 'no'.

By solving the above equations we can get $P(A = yes)$ and $P(A = no)$ in the cases where $\theta \neq 1/2$. If $\theta = 1/2$ Unrelated-Question Model can be used. In this model two unrelated questions are asked with one probability for one of the questions is known. [1]

[1] Questions for Unrelated-Question Model can be like:

In these methods only *the privacy* of the responder is considered. It is also assumed that the responder follows the procedure.

3 Association Rule Mining

3.1 Uniform Randomization

Uniform Randomization is presented by Evfimievski et al. [2] as a generalization of Warner's idea. In this generalized form the responder is not answering a question but sending a transaction. Before sending a transaction to the server, the responder takes each item and with probability p replaces it by a new one not originally present in this transaction.

For large values of p, most of the items in most randomized transactions will be 'false' meaning that they are not from the original transaction. However, if there are enough clients and transactions, then frequent itemsets will still be 'visible'.

Uniform Randomization has its problems. For instance for a 3-itemset it is more likely that one or two items are changed than all three. If we have a 3-itemset that occurred seldom, after randomization it is even more unlike to occur even once. Then every time we see it in a randomized transaction we are quite sure that at least one item from this item set is 'true'. In this case we can say that a privacy breach has occurred.

Example 3.1 With $p = 80\%$ a 3-itemset that originally occurred in 1% transactions will occur in about $1\% \cdot (0.2)^3 = 0.008\%$ transactions. If there is 10 000 possible items, the probability that 10 randomly inserted items contain a given 3-itemset is less than $10^{-7}\%$.

3.2 Privacy Breaches

A privacy breach is a situation when, for some clients, the disclosure of its randomized private information to the server reveals that a certain property of unrandomized private information holds with high probability.

1. Do you ever have the sensitive attribute A?
2. Flip a coin. Did you get a head?

Let's have randomization operator R that randomly transforms a sequence of N transactions (T) into a (usually) different sequence of N transactions (T'). t_i is the i-th transaction in T and t'_i is the i-th transaction in randomized sequence T'.

Definition 3.2 We say that item set A *causes a privacy breach of level b* if some item $a \in A$ and for some $i \in 1, ..., N$ we have $P[a \in t_i | A \subset t'_i] \geq p$

Evfimievski et al. focus on the controlling privacy breaches given by above definition 3.2. Other information obtained from the randomization is ignored like the missing items and the size of the randomized transactions. Also all extra information the server might know about the client are skipped.

The problem with Definition 3.2 is that we have to randomize the data before we can calculate the privacy breach. If we select 'over safe' randomization parameters we might not have enough data to reach sufficient accuracy.

One way to find out a compromise between privacy and accuracy is to construct a situation that is pessimistic enough. Evfimievski et al. propose the situation when the item set and its subsets are frequent.

3.3 Cut-and-paste Randomization

Cut-and-paste randomization is an implementation of Uniform Randomization mentioned in Section 3.1. It is used to scramble the data set for the experiments presented at the section 3.7. The method is defined below:

Definition 3.3 A *cut-and-paste randomization* operator is a special case of select-a-size operator (which is presented on the paper [2]). For each input transaction size m, it has two parameters: $\rho_m \in (0, 1)$ (randomization level) and an integer $K_m > 0$ (the *cut-off*). The operator takes each input transaction t_i independently and proceeds as follows to obtain transaction t'_i(here $m = |t_i|$):

1. The operator chooses an integer j uniformly at random between 0 and K_m; if $j > m$ it sets $j = m$
2. The operator selects j items out of t_i uniformly at random (without replacement). These items are placed into t'_i

3. Each other item (including the rest of t_i) is placed into t'_i with probability ρ_m, independently.

3.4 Support Recovery

Suppose we have a set $\mathcal{I}$ of n items: $\mathcal{I} = \{a_1, a_2, ..., a_n\}$. Let T be a sequence of N transactions $T = (t_1, t_2, ..., t_N)$ where each transaction t_i is a subset of $\mathcal{I}$. Let A be some subset of items (that is, $A \subseteq \mathcal{I}$).

Definition 3.4 The fraction of the transactions in T that have intersection with A of size l among all transactions in T is called *partial support* of A for intersection size l:

$$supp_l^T(A) := \frac{\#\{t \in T | \#(A \cap t) = l\}}{N} \quad (3)$$

where support of A is $supp^T(A) = supp_k^T(A)$ for $k = |A|$.

Definition 3.5 Suppose that our randomization operator is both per-transaction[2] and item-invariant[3]. Consider a transaction t of size m and an item set $A \subset \mathcal{I}$ of size k. After randomization, transaction t becomes t'. We define

$$p_k^m[l \to l'] = p[l \to l'] := P[\#(t' \cap A) = l' | \#(t \cap A) = l]. \quad (4)$$

where both l and l' must be integers in $0, 1, ..., k$.

Let all transactions in T have the same size m_i. (If this is not so, we have to handle each transaction size separately.) Let define following arrays for partial supports:

$$\vec{s} := (s_0, s_1, ..., s_k)^T,$$
$$\vec{s'} := (s'_0, s'_1, ..., s'_k)^T$$

According the proof on the paper [2] the expected value of the partial support of the scrambled data is

$$\mathbf{E}\vec{s'} = P \cdot \vec{s} \quad (5)$$

[2]The operator uses only the knowledge of transaction t when randomizing it.

[3]The order of items belonging to transaction t does not effect to the randomization.

where P is the $(k+1) \times (k+1)$ matrix with elements $P_{(l'l)} = p[l \to l']$.

Denote $Q = P^{-1}$ (assume that it exists) and solve the vector $vecs$ from Equation 5.

$$\vec{s} = Q \cdot \mathbf{E}\vec{s'} \quad (6)$$

We have thus obtained an unbiased estimator for the original partial supports given randomized partial supports:

$$\vec{s}_{est} := Q \cdot \vec{s'} \quad (7)$$

The k-th coordinate $(\tilde{s})$of the $\vec{s}_{est}$ is in the special interest because it can be use as an estimate of the support s of the item set A in T. Denote $q[l \to l'] := Q_{(ll')}$.

$$\tilde{s} = \sum_{l'=0}^{k} s'_{l'} \cdot q[k \leftarrow l'] \quad (8)$$

There are also formulas for variance and the unbiased estimator of variance. Detailed proof can be found from the paper [2].

$$Var\ \tilde{s} = \frac{1}{N} \sum_{l=0}^{k} s_l \left(\sum_{l'=0}^{k} p[l \to l'] q[k \leftarrow l']^2 - \delta_{l=k}\right) \quad (9)$$

$$(Var\ \tilde{s})_{est} = \frac{1}{N} \sum_{l'=0}^{k} s'_{l'} (q[k \to l']^2 - q[k \leftarrow l']) \quad (10)$$

3.5 Limiting Privacy Breaches

The problem with Definition 3.2 is that we have to randomize the data before we can calculate the privacy breach. If we select 'over safe' randomization parameters we might not have enough data to reach sufficient accuracy.

One way to find out a compromise between privacy and accuracy is to construct a situation that is pessimistic enough. Evfimievski et al. propose the situation when the item set and its subsets are frequent. The procedure to limit privacy breaches can be found from the paper [2]. The general form is:

1. Estimate maximum possible support of a k-itemset in the transactions of given size m, for different k and m_i;

2. Given the maximum supports, find a situation most likely to cause a privacy breach.

3. Make randomization just strong enough to prevent such a privacy breach.

3.6 Discovering Associations

The associations are discovered by using the value of the estimated support. The basic idea of the discovery is to go through all item sets and select them whose support is above s_{min}. The algorithm used is modified Apriori algorithm [3]:

1. Let $k = 1$, let 'candidate sets' be all single-item sets. Repeat the following until k is too large for support recovery (or until no candidate sets are left):

   (a) Read the randomized data file and compute the supports of all candidate sets, separately for each unrandomized transaction size.

   (b) Recover the predicted supports and standard deviations[4] for the candidate sets from the equations 8,10.

   (c) Discard every candidate set whose support is below its *candidate limits*. A good value for the candidate limit is $s_{min} - \sigma$.

   (d) Save for the output only those candidate sets whose predicted support is at least s_{min}.

   (e) Form all possible $(k + 1)$-item sets such that all their k-subsets are among the remaining candidates. Let these item sets be the new candidate sets.

   (f) Let $k = k + 1$.

2. Output all the saved item sets.

3.7 Experiments of Association Discovery

The association discovery is experimented with two 'real life' data sets: the soccer data set and the mail-order data set. Cut-and-paste randomization were used with the cut-off of 7. The privacy breach level is 50%.

[4]Standard deviation = $\sigma = \sqrt{variance}$

The soccer data set is generated from the clickstream log of the 1998 world cup web site. There were 6 525 879 transactions. The mail-order data set is from the mail order company and consists of items ordered by a customer in a single mail order. There were 2.9 million transactions.

The experiment shows that the estimation of the support works quite well. There were only comparatively few false positives(item sets wrongly included into the output) and even fewer false drops (item sets wrongly omitted).

4 Classification of Disguise Data

Classification is one of the forms of data analysis that can be used to extract models describing important data classes or predict future data.

Classification is a two-step process:

1. A model is built from the input of training data set which is composed of data tuples described by attributes. Each tuple is assumed to belong to a predefined class described by one of the attributes, called the class label attribute.

2. The predictive accuracy of the model (or classifier) is estimated. The estimation is usually done by classifying a test data set (that is not used for training) and comparing the results to the class labels.

Du et al. provide a classification method for disguised binary data. The randomization technique is described in Section 4.1. The step one of the classification method is described in Section 4.2. And the step two in Section 4.3.

4.1 Multivariate Randomized Response

Multivariate Randomized Response presented by W. Du et al. [1] extends the randomized response technique so that instead of one question a set of questions is presented to the responder. The responder is supposed to either answer all the questions truthfully (with probability θ) or lie to all of them (with probability $1 - \theta$).

There is also other solution for sets that contain multiple questions [6]. According Du et al. the solution presented is not efficient enough to be extended to data mining. That's why the new method is proposed.

Suppose that there are N attributes and the data mining is based on these. Attributes are A_1,A_2,..A_N. Let E represent any logical expression based on those attributes. Let $P^*(E)$ be the proportion of the records in the whole disguised data set that satisfy $E = true$. Let $P(E)$ be the proportion of the records in the whole undisguised data set fulfilling the expression.

By using the randomized response technique with Related-Question Model, we can get the following equations.

$$P^*(E) = P(E) \cdot \theta + P(\overline{E}) \cdot (1 - \theta) \quad (11)$$
$$\overline{P^*(E)} = P(E) \cdot \theta + P(\overline{E}) \cdot (1 - \theta) \quad (12)$$

where $\overline{E}$ is the complement of E, where each attribute separately is given a complement value. An example of E and $\overline{E}$:

$$E = (A_1 = 1) \cap (A_2 = 1) \cap (A_3 = 0)$$
$$\overline{E} = (A_1 = 0) \cap (A_2 = 0) \cap (A_3 = 1)$$

4.2 Modified ID3 algorithm

In the paper [1] modified ID3 algorithm is presented to build decision tree based on disguised data. The difference to the original ID3 algorithm [4] is how the information gain is calculated and especially how $P(E)$ is calculated.

The attribute with the highest gain is selected to partition the training sample S. The training sets are recursively partitioned until each partition consists of samples from one class, the algorithm is stopped. The information gain for any candidate attribute A_k if it used to partition S is:

$$Gain(S, A_k) = Entropy(S) - \sum_{v \in A_k} (\frac{|S_v|}{|S|} Entropy(S_v)) \quad (13)$$

where v represents any possible value of attribute A_k, S_v is the subset of S for which attribute A_k has value v, $|S_v|$ is the number of elements in S_v and $|S|$ is the number of elements in S.

The entropy of S is

$$Entropy(S) = -\sum_{j=i}^{m} Q_j \log Q_j \quad (14)$$

where Q_j is the relative frequency of class j in S and m is the number of classes in the training set.

Because we are using disguised data we have to use estimates for $|S|$, $|S_v|$, $Entropy(S)$ and $Entropy(S_v)$. Du et al. present an example how to calculate the estimates in practice.

Example 4.1 We want to know the information gain for a node V that satisfies $A_i = 0$ and $A_j = 1$. Let S be the training data set consisting of the samples that belong to node V.

To compute $|S|$, the number of elements in S, let

$$E = (A_i = 1) \cap (A_j = 0)$$
$$\overline{E} = (A_i = 0) \cap (A_j = 1)$$

We can compute $P(E)$ as mentioned previously. Hence $|S| = P(E) * n$, where n is the number of records in the whole training set.

It is assumed that the class label is binary. Class label can be randomized also. Then the complementary class is used for the randomized data set. Let

$$E_c = (A_i = 1) \cap (A_j = 1) \cap (Class = 0)$$
$$\overline{E_c} = (A_i = 0) \cap (A_j = 0) \cap (Class = 1)$$

We can compute $P(E_c)$ directly from $P^*(E_c)$ and $P^*(\overline{E_c})$. Therefore, $Q_0 = \frac{P(E_c)*n}{|S|}$, $Q_1 = 1 - Q_0$ and $Entropy(S)$ can be computed. Note that results for $P(E)$ and $P(E_c)$ are different.

We still need values for $|S_{A_k=1}|$, $|S_{A_k=0}|$, $Entropy(S_{A_k=0})$ and $Entropy(S_{A_k=1})$. These can be similarly computed. For example, $|S_{A_k=1}|$ can be computed by letting

$$E = (A_i = 1) \cap (A_j = 1) \cap (A_k = 1)$$
$$\overline{E} = (A_i = 0) \cap (A_j = 0) \cap (A_k = 0)$$

Then we solve $P(E)$ using $P^*(E)$ and $P^*(\overline{E})$, and get $|S_{A_k=1}| = P(E) * n$.

4.3 Accuracy score

To avoid over-fitting in decision tree building, data set for testing is needed. This data set is used to determine how accurate the decision tree is. For disguised data this is not a trivial task.

Du et al. offer an example how the accuracy score can be calculated in their situation.

Example 4.2 Assume the number of attributes is 5 and the probability $\theta = 0.7$. One of the test records is 01101. It and its' complement 10010 are fed to the decision tree. If both prediction results are correct (incorrect) we can make an accurate conclusion about the testing results. On other situation we can make a conclusion with 0.7 certainties.

If number of testing data is large we can calculate $P(correct)$ by solving following equations.

$$P^*(correct) = P(correct) \cdot \theta + \overline{P}(correct) \cdot (1 - \theta) \quad (15)$$
$$\overline{P}^*(correct) = \overline{P}(correct) \cdot \theta + P(correct) \cdot (1 - \theta) \quad (16)$$

where $P^*(correct)$ is the proportion of correct predictions from testing data set S. $\overline{P}^*(correct)$ is the proportion of correct predictions from testing data set $\overline{S}$. $P(correct)$ is the proportion of correct predictions from testing data set U. $\overline{P}(correct)$ is the proportion of correct predictions from testing data set $\overline{U}$. S is the disguised data set and $\overline{S}$ is calculated from it by reversing the values from 0 to 1 and 1 to 0. U is the non-existing undisguised data set and $\overline{U}$ is non-existing undisguised reversed data set.

4.4 Classification Experiments

The data set used for experiment was from UCI Machine Learning Repository. It contains 48842 instances with 14 attributes (6 continues and 8 nominal) and a label describing the salary level. Prediction task is to determine whether a person's income exceeds \$50k/year.

First the data was binarized so that the values of each attribute are split from the median point. Data set was randomized using $\theta = 0.1, 0.2, 0.3, 0.4, 0.45, 0.51, 0.55, 0.6, 0.7, 0.8, 0.9, 1.0$. Randomization was done 50 times for every value. For every randomization the decision tree was build up and accuracy score calculated. For every θ the mean and the variance of the accuracy score was calculated.

The experiment shows that with θ ranges $[0, 0.4]$ and $[0.6, 1]$ the method can still achieve very high accuracy. When θ is close to 0.5 the data for a single attribute become uniformly distributed. On the other hand, when θ=0, all true information about the original data set is revealed. When θ is moving toward 0.5 the *privacy level* is enhancing.

5 Conclusions

Both papers are based on randomized response. Evfimievski et al. [2] propose an approach to do privacy preserving association rule mining so that each attribute is independently disguised. They define 10 as a maximum size for the attribute vector. Longer vectors, when disguised properly, cannot be utilized for data mining. Attributes contain categorical data such as books customer has bought.

Du et al. [1] propose an approach to do privacy preserving classification. Their randomization function isn't item-invariant and they do not offer any maximum size for the attribute vector. The attributes are assumed to be binary valued and the size of the attribute vector is not randomized.

In both approaches the randomization is done before the data are sent to the server and only the privacy of the responder is considered. It is also assumed that the responder follows the procedure although the transaction protocol isn't defined.

The randomization method of Evfimievski et al. is suitable for short transactions of categorical data such as the books bought together.

The randomization method of Du et al. is suitable for fixed size attribute vectors such as demographic profiles. They offer methods for binary data but in the future they are going to extend the solution to non binary data also.

The proofing of the equations and the detailed description of the experiments are scoped out of this survey. The support recovery (3.4) and the calculation of the information gain (4.2) are presented more detailed as an example.

References

[1] W. Du, and Z. Zhan. Using Randomized Response Techniques for Privacy-Preserving Data Mining. In *Proceedings of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2003.

[2] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 2002.

[3] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast Discovery of Association Rules. In U.M Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, chapter 12, pages 307-328. AAAI/MIT Press, 1996

[4] JK. Han and M. Kamber. *Data Mining Concepts and techniques*. Morgan Kaufmann Publishers, 2001.

[5] S.L. Warner. Randomized Response: a Survey Technique for Eliminating Evasive Answer Bias. *The American Statistical Association*, 60(309), pp. 6369, March 1965.

[6] A. C. Tamhane. Randomized response techniques for multiple sensitive attributes. *The American Statistical Association*, 76(376):916-923, December 1981.