Privacy-Preserving Data Mining in Electronic Surveys

Justin Zhan[1] and Stan Matwin[2]
(Corresponding author: Justin Zhan)
Carnegie Mellon Cylab Japan[1]
Kobe Harborland Center Building 17F, 1-3-3 Higashikawasaki-cho, Chuo-ku, Kobe 650-0044, Japan
Email: justinzh@andrew.cmu.edu
School of Information Technology & Engineering, University of Ottawa[2]
800 King Edward Ave., P.O. Box 450 Stn A, Ottawa Ontario, K1N 6N5, Canada
Email: stan@site.uottawa.ca

(Received August 23, 2005; revised and accepted Oct. 11, 2005 & Jan. 4, 2006)

Abstract

Electronic surveys are an important resource in data mining. However, how to protect respondents' data privacy during the survey is a challenge to the security and privacy community. In this paper, we develop a scheme to solve the problem of privacy-preserving data mining in electronic surveys. We propose a randomized response technique to collect the data from the respondents. We then demonstrate how to perform data mining computations on randomized data. Specifically, we apply our scheme to build a Naive Bayesian classifier from randomized data. Our experimental results indicate that accuracy of classification in our scheme, when private data is protected by randomization, is close to the accuracy of a classifier build from the same data with the total disclosure of private information. Finally, we develop a measure to quantify privacy achieved by our proposed scheme.

Keywords: Data mining, privacy, randomization

1 Introduction

Data mining has emerged as a means for identifying patterns and trends from large amounts of data. To conduct data mining computations, we need to collect data first. However, because of privacy concerns, people might decide to selectively divulge information, or give false information, or simply refuse to disclose any information at all. There is research evidence [2] that providing privacy protection measures is a key to the success of data collection.

There are many ways to collect data. For instance, data may be collected using transaction records. This can often be done without people's knowledge, and individuals have no control over what information can be collected. The evolving legal developments will hopefully soon preclude this questionable practice. Another way to collect data is to solicit respondents' responses via surveys, for example, respondents might be asked to rate certain products, or they might be asked whether they have a certain medical condition, etc. The collected data is entered into a database. Although answering survey questions gives respondents control over whether they want to disclose their information or not, privacy concerns might hinder the respondents from telling the truth or responding at all (we will refer to this problem as *respondent privacy in electronic surveys*). How can we improve the chance to collect more truthful data that are useful for data mining while preserving respondents' privacy? How can respondents contribute their personal information without compromising their privacy?

One way to achieve privacy is to let each respondent randomize their data, such that data collector cannot derive the truthful information about a respondent's private information. The challenge is how to conduct data mining on randomized data. To address this challenge, we propose the following computation model depicted in Figure 1. The model consists of a data collection step and a computation step. In data collection step, each respondent utilizes certain techniques to randomize her data, then sends randomized data to data collector (solid line from the respondents to data collector) who *cannot* access the actual respondents' data (dashed line from data collector to respondents), and should not be able to find out any respondent's actual data with probabilities better than a pre-defined threshold. In computation step, data collector constructs a database using randomized data, and conducts data mining computations on this database. The goal of data collector is to derive useful information (or knowledge) out of this randomized database. In this paper, we focus on the naive Bayesian (e.g., NB) clas-

sification [7]. However, the proposed approach can be applied to other data mining algorithms as well.

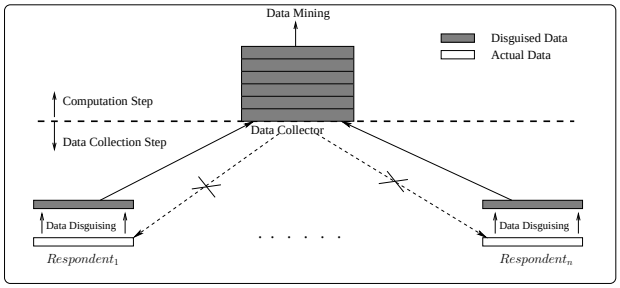

Figure 1: Privacy-oriented mining of survey data

We propose to use the *Randomized Response* techniques [11] to solve the problem of respondent privacy in electronic surveys. The basic idea of randomized response is to scramble the data in such a way that the data collector cannot tell with probabilities better than a pre-defined threshold whether the data from a respondent contain truthful information about the sensitive, private information. Although information from each individual respondent is scrambled, if the number of respondents is significantly large, the aggregate information of these respondents can be estimated with reasonable accuracy. Such property is useful for naive Bayesian classification since it is based on aggregate values of a data set, rather than individual data items.

The contributions of this paper are as follows: (1) We have modified naive Bayesian classification algorithm [7] to make it work with data disguised by randomized response techniques, and implemented the modified algorithm. (2) We then conducted a series of experiments to measure accuracy of our modified naive Bayesian algorithm on randomized data. Our results show that if we choose the appropriate randomization parameters, the accuracy we have achieved is very close to the accuracy achieved by the standard, unmodified naive Bayesian classifier on the undisguised data. (3) We develop a method to measure privacy achieved by proposed approach.

The rest of the paper is organized as follows: we discuss related work in Section 2. In Section 3, we describe how to utilize multi-variant randomized response technique to build a naive Bayesian classifier on randomized data. In Section 4, we describe our experimental results. Further discussion is provided in Section 5. We give our conclusion in Section 6.

2 Related Work

There are currently two approaches to achieve privacy-preserving data mining: one is to use Secure Multi-party Computation (SMC) techniques [12]. Several SMC-based privacy-preserving data mining schemes have been proposed [8, 10]. The other is the randomization approach. Agrawal and Srikant proposed a scheme for privacy-preserving data mining using random perturbation approach [1]. In their scheme, a random number is added to the value of a private attribute. For example, if x_i is the value of a private attribute, $x_i + r$, rather than x_i, will appear in the database, where r is a random value drawn from some distribution. The paper shows that if the random number is generated with some known distribution (e.g., uniform or Gaussian distribution), it is possible to recover the distribution of the values of that private attribute. Assuming independence of the attributes, the paper then shows that a decision tree classifier can be built with the knowledge of distribution of each attribute.

Rizvi and Haritsa presented a scheme called **MASK** to mine associations with secrecy constraints in [9]. Evfimievski et al. proposed an approach to conduct privacy-preserving association rule mining based on randomization techniques [4]. Du and Zhan [3] utilized randomized response technique for decision tree classification. The method presented here is also based on randomized response technique. The difference is that the randomized response technique in [3] is based on the related-question model, and our randomized response technique is based on the unrelated-question model. In Section 6, we discuss the advantages of our approach over [3].

3 Building Naive Bayesian Classifiers Using Multi-variant Randomized Response Techniques

Randomized Response techniques were first introduced by Warner [11] to solve the following survey problem: to estimate the percentage of respondents in a population that has attribute A, queries are sent to a group of respondents. Since the attribute A is related to some confidential aspects of human life, respondents may decide not to reply at all or to reply with incorrect answers.

For the purpose of this discussion, we will distinguish two types of questions in a survey: questions about the respondent's *private* information, and questions about the respondent's *personal* information. Both kinds of information refer to the attributes of the respondent. The private information is an attribute the respondent would rather not disclose, including its probability distribution (e.g., whether the respondent has a certain medical condition; or whether she takes a given medication); the personal information is also an attribute of the respondent, but unlike the private information the respondents do not normally mind that the data collector knows the probability distribution of the personal information (e.g., what is the probability that the color of the respondent's hair being black, or what is the probability that she lives near a lake). We also assume that the private and personal information are unrelated - e.g., taking a medication in unrelated to one's hair color.

To enhance the level of cooperation, instead of asking

each respondent whether she has the attribute A, the data collector asks each respondent two *unrelated* questions. One of them asks private information, i.e., the one that the data collector is interested in. The other refers to the personal information. The answers to the two questions are unrelated to each other [11]. For example, the survey questions can be designed as follows:

1) Do you have the *private* attribute A?

2) Do you have the *personal* attribute Y?

In practice, the first question could be "Are you taking medicine A?", and the second question could be "Do you live near a lake?". Respondents answer one of these two questions. They use a randomization device to decide which question to answer, without letting the data collector know which question is answered. Each randomization device tells the respondent which question she is to answer: the probability of choosing the first question is θ, and the probability of choosing the second question is $1-\theta$. Although the data collector learns a response (i.e., "yes" or "no"), he does not know which question was answered by the respondents. It is important to engineer the interaction between data collector and respondent in such a way that the respondent will trust the system, i.e., the respondent will clearly understand that data collector has no way of knowing which of the two questions is answered. Thus the respondent feels that her privacy is preserved. We further comment on this in Section 5.1. Note that data collector only knows the probability distribution of the respondent's attribute Y. This is consistent with the interpretation of a personal attribute - the data collector could know the distribution of the values (e.g., hair colors) of the personal attribute in the general population, without knowing the value of that attribute for a specific respondent.

The randomized response technique discussed above considers only one attribute. However, data sets usually consist of multiple attributes; finding the relationship among these attributes is one of the major goals for data mining. Therefore, we need techniques that can handle multiple attributes while supporting various data mining computations.

In this paper, we provide multi-variant randomized response technique (MRR) to address the problems of respondent privacy in electronic surveys.

3.1 Notations

In this work, we assume data are binary, but the techniques can be extended to categorical data. Suppose there are N *private* attributes (A_1, A_2, ..., A_N) in a data set A. We construct N *personal* attributes (Y_1, Y_2, ..., Y_N). We want one *private* attribute (question) to pair with one *personal* attribute (question), therefore we make the number of attributes of Y and the number of attributes of A be equal. Let A and Y represent any logical expression based on those attributes $A_i (i \in [1, N])$ and $Y_i (i \in [1, N])$. For example, A can be $(A_1 = 0) \wedge (A_2 = 1)$ and Y can be $(Y_1 = 0) \wedge (Y_2 = 1)$.

Let $P(Y)$ be the proportion of the records in the personal data that satisfy $Y =$ `true`. Let $P^*(A)$ be the proportion of the records in the whole *randomized* data set that satisfies $A =$ `true`. Let $P(A)$ be the proportion of the records in the whole *non-randomized* data set that satisfy $A =$ `true` (the potential non-randomized data set which in reality does not exist). $P^*(A)$ can be observed from the randomized data, but $P(A)$, the actual proportion that we are interested in, cannot be observed from the randomized data because the non-randomized data set is not available to anybody; we have to estimate $P(A)$. The goal of MRR is to find a way to estimate $P(A)$ from $P^*(A)$.

3.2 Multi-variant Randomized Response Scheme

In this scheme, all the attributes including the class label either keep the same values or obtain the values from personal data. In other words, when sending the private data to the data collector, respondents either tell their answers to the private questions or tell their answers to the personal questions. The probability for the first event is θ, and the probability for the second event is $1-\theta$. For example, assume a respondent's attribute values A_1 and A_2 are 11 for private data; and the respondent's attribute values Y_1 and Y_2 are 01. The respondent generates a random number between 0 and 1; if the number is less than θ, she sends 11 to the data collector; if the number is bigger than θ, she sends 01 to the data collector. Since the data collector only knows θ which is the same for all respondents and does not know the random number generated by each respondent, he cannot know whether the respondent tells the values from private data or personal data. To simplify our presentation, we use $P(A(11))$ to represent $P(A_1 = 1 \wedge A_2 = 1)$, $P(Y(11))$ to represent $P(Y_1 = 1 \wedge Y_2 = 1)$ where"$\wedge$" is the logical `and` operator. Because the contributions to $P^*(A(11))$ partially come from $P(A(11))$, and partially come from $P(Y(11))$, we can derive the following equation:

$$P^*(A(11)) \quad = \quad P(A(11)) \cdot \theta + P(Y(11)) \cdot (1-\theta).$$

Since $P(Y(11))$ is known as Y is personal data, θ is determined before collecting the data, and $P^*(A(11))$ can be directly computed on the disguised (randomized) data set. By solving the above equation, we can obtain $P(A(11))$, the information needed to build a naive Bayesian classifier. The general model is described in the following:

$$P^*(A) \quad = \quad P(A) \cdot \theta + P(Y) \cdot (1-\theta). \qquad (1)$$

3.3 Building Naive Bayesian Classifiers

The naive Bayesian classifier is one of the most successful algorithms in many classification domains. Despite of its simplicity, it is shown to be competitive with other

complex approaches, especially in text categorization and content based filtering. The naive Bayesian classifier applies to learning tasks where each instance x is described by a conjunction of attribute values and where the target function $f(x)$ can take on any value from some finite set V. A set of training examples of the target function is provided, and a new instance is presented, described by the tuple of attribute values $< a_1, a_2, \cdots, a_n >$. The learner is asked to predict the target value for this new instance. Under a conditional independence assumption, i.e., $P(a_1, a_2, \cdots, a_n|v_j) = \Pi_{i=1}^{n} P(a_i|v_j)$, a naive Bayesian classifier can be derived as follows:

$$\begin{aligned} v_{NB} &= argmax_{v_j \in V} P(v_j) \Pi_{i=1}^{n} P(a_i|v_j) \\ &= argmax_{v_j \in V} P(v_j) \Pi_{i=1}^{n} \frac{P(a_i \wedge v_j)}{P(v_j)} \end{aligned}$$

To build a NB classifier, we need to compute $P(v_j)$ and $P(a_j \wedge v_j)$. To compute $P(v_j)$, we can use the general model (Equation 1) with A being $(C = v_j)$ and Y being $(CY = v_j)$ where C is the class label for the private data A and CY is the class label of personal data Y. $P^*(A)$ can be computed directly from the (whole) randomized data set. $P(Y)$ is known since it is personal and θ is known as well. By knowing θ, data collector, who conducts the training, only knows the probability of the training data being private, but does not exactly know if each value is private data or not. By solving the above equation, we can get $P(A)$ which is $P(C = v_j)$ in this case. Similarly, we can compute $P(a_i \wedge v_j)$ using the general model (Equation 1) with A being $(A_i = a_i \wedge C = v_j)$ and Y being $(Y_i = a_i \wedge CY = v_j)$.

3.4 Testing

Conducting the testing is straightforward when data are not randomized, but it is a non-trivial task when the testing data set is randomized. When we choose a record from the testing data set, compute a predicted class label using the naive Bayesian classifier, and find out that the predicated label does not match the record's actual label, can we say this record fails the testing? If we knew whether the record represents the private or the personal data, and if we knew the true class for each data, we could easily answer this question. But how can we compute the accuracy score of a NB classifier when data are randomized? Our answer is to apply the multi-variant randomized response technique once again to compute the accuracy. Let us use an example to illustrate how to compute the accuracy. Assume the number of attributes is 2. To test a record $(A_1 = 1, A_2 = 0)$ (denoted by $A(10)$), we feed $A(10)$ and $Y(10)$, where $Y = (Y_1 = 1, Y_2 = 0)$ to the NB classifier built in Section 3.3. Let $P^*(A(cc))$ be the proportion of correct predictions using the disguised (randomized) testing data set, $P(Y(cc))$ be the proportion of correct predictions in the personal data, and let $P(A(cc))$ be the proportion of correct predictions in the private data. $P(A(cc))$ is what we want to estimate.

Because $P^*(A(cc))$ consist of contributions from $P(A(cc))$ and $P(Y(cc))$, we have the following equation:

$$P^*(A(cc)) = P(A(cc)) \cdot \theta + P(Y(cc)) \cdot (1 - \theta),$$

where $P^*(A(cc))$ can be obtained from disguised testing data set. θ is known and by knowing θ, data collector, who conducts the testing, only knows the probability of the testing data being private, but does not exactly know if each value is private data or not. How does the data collector know $P(Y(cc))$? One implementation is as follows: each respondent is given the same classifier by the data collector. The classifier is constructed during the training (Section 3.3). Each respondent applies this classifier on her personal data Y and communicates the number of correct predictions (0 or 1) to the data collector, who then computes $(Y(cc))$. Note that the data collector does not know the values of the Y attributes, only the result of the classifier. The data collector can solve the above equation and get $P(A(cc))$, the accuracy score of testing.

4 Measuring Privacy

Enlighten by [9] where a probability-based method is provided. We develop privacy measure for our proposed scheme as follows:

- First, we measure privacy for a single entry.
- Second, we select the minimal privacy value and treat it as the privacy level for the group. The reason why we choose the minimal value for the group is that, the entries are randomized together, by finding the original value for one entry will cause disclosing the original values for other entries in the group.

4.1 Measure Privacy for a Single Entry Before Mining

For a single entry, original value can be 1 or 0; randomized value can be 1 or 0 as well. Privacy comes from uncertainty of original value given a randomized value. In other words, if original value is 1, given randomized value 1 or 0, privacy will be the probability of data collector guess the original value being 0. There are four possible randomization results:

- Original value is 1, the value after randomization is still 1;
- Original value is 1 but the value after randomization is 0;
- Original value is 0 but the value after randomization is 1;
- Original value is 0, the value after randomization is still 0.

Consequentially, there are four components in the privacy measure:

- The probability that original value is 1, multiplies the probability that original value is 1 and the value after randomization is still 1 , then times the probability that guessing the original value is 0 given the randomized value is 1.

- The probability that original value is 1, multiplies the probability that original value is 1 but the value after randomization is still 0, then times the probability that guessing the original value is 0 given the randomized value is 0.

- The probability that original value is 0, multiplies the probability that original value is 0 but the value after randomization is 1 , then times the probability that guessing the original value is 1 given the randomized value is 1.

- The probability that original value is 0, multiplies the probability that original value is 0 and the value after randomization is still 0, then times the probability that guessing the original value is 1 given the randomized value is 0.

Let's use the following denotations:

- Let's O_m be the original value;

- Let's R_m be the value after randomization;

- Let's W_a be the probability that a value is 1 in data set A, and the probability that a value is 0 in data set A is $(1 - W_a)$;

- Let's W_y be the probability that a value is 1 in data set Y, and the probability that a value is 1 in data set Y is $(1 - W_y)$;

Privacy denoted by $PRE(PSE)$ for a single entry before mining can be derived as follows:

$$\begin{aligned} & PSE(PRE) \\ = \; & Pr(O_m = 1) * Pr(R_m = 1|O_m = 1) * Pr(O_m = 0|R_m = 1) + \\ & Pr(O_m = 1) * Pr(R_m = 0|O_m = 1) * Pr(O_m = 0|R_m = 0) + \\ & Pr(O_m = 0) * Pr(R_m = 1|O_m = 0) * Pr(O_m = 1|R_m = 1) + \\ & Pr(O_m = 0) * Pr(R_m = 0|O_m = 0) * Pr(O_m = 1|R_m = 0) \\ = \; & Component_1 + Component_2 + Component_3 + Component_4 \end{aligned}$$

The first component can be computed as follows:

$$\begin{aligned} & Component_1 \\ = \; & W_a * [\theta + (1-\theta) * W_y] * \frac{Pr(R_m = 1|O_m = 0) * Pr(O_m = 0)}{Pr(R_m = 1)} \\ = \; & \frac{W_a * [\theta + (1-\theta) * W_y] * (1-\theta) * (1-W_y) * (1-W_a)}{Pr(R_m = 1|O_m = 1) * Pr(O_m = 1) + Pr(R_m = 1|O_m = 0) * Pr(O_m = 0)} \\ = \; & \frac{W_a * [\theta + (1-\theta) * W_y] * (1-\theta) * (1-W_y) * (1-W_a)}{[\theta + (1-\theta) * W_y] * W_a + (1-\theta) * (1-W_y) * (1-W_a)} \end{aligned}$$

Similarly, we can obtain other components.

$$\begin{aligned} & Component_2 \\ = \; & \frac{W_a * (1-\theta) * (1-W_y) * [\theta + (1-\theta) * (1-W_y)] * (1-W_a)}{[\theta + (1-\theta) * (1-W_y)] * (1-W_a) + (1-\theta) * (1-W_y) * W_a} \end{aligned}$$

$$\begin{aligned} & Component_3 \\ = \; & \frac{(1-W_a) * (1-\theta) * W_y * [\theta + (1-\theta) * W_y] * W_a}{(\theta + (1-\theta) * W_y) * W_a + (1-\theta) * W_y * (1-W_a)}, \end{aligned}$$

$$\begin{aligned} & Component_4 \\ = \; & \frac{(1-W_a) * [\theta + (1-\theta) * (1-W_y)] * (1-\theta) * (1-W_y) * W_a}{[\theta + (1-\theta) * (1-W_y)] * (1-W_a) + (1-\theta) * (1-W_y) * W_a}. \end{aligned}$$

We then get

$$\begin{aligned} & PSE(PRE) \\ = \; & \frac{W_a * (\theta + (1-\theta) * W_y) * (1-\theta) * (1-W_y) * (1-W_a)}{(\theta + (1-\theta) * W_y) * W_a + (1-\theta) * (1-W_y) * (1-W_a)} + \\ & \frac{W_a * (1-\theta) * (1-W_y) * (\theta + (1-\theta) * (1-W_y)) * (1-W_a)}{(\theta + (1-\theta) * (1-W_y)) * (1-W_a) + (1-\theta) * (1-W_y) * W_a} + \\ & \frac{(1-W_a) * (1-\theta) * W_y * (\theta + (1-\theta) * W_y) * W_a}{(\theta + (1-\theta) * W_y) * W_a + (1-\theta) * W_y * (1-W_a)} + \\ & \frac{(1-W_a) * (\theta + (1-\theta) * (1-W_y)) * (1-\theta) * (1-W_y) * W_a}{(\theta + (1-\theta) * (1-W_y)) * (1-W_a) + (1-\theta) * (1-W_y) * W_a} \\ = \; & \frac{(1-W_a) * W_a * (1-\theta) * (\theta + (1\theta) * W_y)}{(\theta + (1-\theta) * W_y) * W_a + (1-\theta) * W_y * (1-W_a)} + \\ & \frac{2 * W_a * (1-W_a) * (1-\theta) * (1-W_y) * (\theta + (1-\theta) * (1-W_y)}{(\theta + (1-\theta) * (1-W_y)) * (1-W_a) + (1-\theta) * (1-W_y) * W_a} \end{aligned}$$

4.2 Measure Privacy for a Single Entry After Mining

There is another issue which may decrease the privacy level that we obtained from pre-mining. That is privacy leak because of inference from the mining middle steps and mining output results or other sources. We will compute the final data privacy as the difference between $PSE(PRE)$ and recoverability from the middle steps and outputs or other sources.

There are mainly two scenarios for current data mining applications:

1) Centralized data mining. In this scenario, there is a data collector who collects all the disguised data from data providers and forms a data set. she then conduct mining on the disguised data set.

2) Distributed data mining. There are several parties, with each of them having a data set, want to collaborate to conduct mining on their combined data set. According to data format,there are two cases:

   - Horizontally Partitioned Data. Each party's data set has the same set of attributes, however, the transactions for each party are different.
   - Vertically Partitioned Data. Each party's data set has different attributes, but the identities for each transaction are the same.

The denotation of recoverability for different scenarios is not the same.

Scenarios I: For inference where the original values for pre-condition of an inference rule being *known*, e.g., vertically partitioned distributed mining, the recoverability can be denoted as:

$$REC = Pr(precondition) * Pr(Confidence),$$

where precondition is the conditions for an inference rule, and $Pr(Confidence)$ is the confidence for an inference rule. The inference rules means that the rules obtained during mining, after mining and other sources. For example, assume we obtain a rule $A_1 \Rightarrow A_2$ with $Pr(A_1 \Rightarrow A_2) = 0.6$, then $Pr(Confidence) = 0.6$.

Let's use an example to show how to compute REC. Assume that data collector obtains the following rules:

$$Pr([A_1 = 1] => [A_2 = 1]) = 80\%;$$
$$Pr([A_1 = 1] => [A_2 = 0]) = 60\%;$$
$$Pr([A_1 = 0] => [A_2 = 1]) = 40\%;$$
$$Pr([A_1 = 0] => [A_2 = 0]) = 30\%.$$

Assume that A_1 belongs to Alice and A_2 belongs to Bob. Then REC will be

$$\begin{aligned} REC &= Pr(precondition) * Pr(Confidence) \\ &= Pr(A_1 = 1) * Pr(A_2 = 1|A_1 = 1) + \\ &\quad Pr(A_1 = 1) * Pr(A_2 = 0|A_1 = 1) + \\ &\quad Pr(A_1 = 0) * Pr(A_2 = 0|A_1 = 0) + \\ &\quad Pr(A_1 = 0) * Pr(A_2 = 1|A_1 = 0). \end{aligned}$$

Scenario II: For inference where the original values for pre-condition of an inference rule being *unknown*, e.g., horizontally partitioned distributed mining and centralized mining, the recoverability can be computed as:

$$REC = Pr(Confidence) * Pr(OriginalValues|RandomizedValues),$$

where $Pr(Confidence)$ is the confidence for an inference rule; $Pr(OriginalValues|RandomizedValues)$ is the probability to make a correct guess for original values given the randomized values.

Let's still use the above example, then REC will be

$$\begin{aligned} REC &= Pr(A_2(R) = 1|A_1(R) = 1) * Pr(A_2(O) = 1|A_2(R) = 1) + \\ &\quad Pr(A_2(R) = 0|A_1(R) = 1) * Pr(A_2(O) = 0|A_2(R) = 0) + \\ &\quad Pr(A_2(R) = 1|A_1(R) = 0) * Pr(A_2(O) = 1|A_2(R) = 1) + \\ &\quad Pr(A_2(R) = 0|A_1(R) = 0) * Pr(A_2(O) = 0|A_2(R) = 0), \end{aligned}$$

where $A_i(O)$ represents the original values and $A_i(R)$ represents the randomized values.

If the inference rules are obtained from other sources, the recoverability is also computed according to the second scenarios. In general, we compute REC for each entry, we then select the largest value among these values. We then compute the final privacy for each entry as follows:

$$PSE = PSE(PRE) - REC.$$

We compute PSE for each single entry. We then select the smallest value $PSE(Min)$ as the privacy value for the group.

5 Experimental Results

To evaluate the effectiveness of our proposed scheme, We conducted experiments on two real life data sets *Adult* and *Breast Cancer* which were obtained from the UCI Machine Learning Repository (ftp://ftp.ics.uci.edu/pub/machine-learning-databases).

5.1 Experimental Steps

We modified naive Bayesian classification algorithm to handle randomized data based on our proposed scheme. We applied our scheme to obtain a privacy-oriented classifier. We also ran the naive Bayesian classification algorithm on original data set, and obtained a base classifier. We then applied the same testing data to both classifiers. Our goal is to compare classification accuracy of these two classifiers. Obviously we want accuracy of privacy-oriented classifier to be close to accuracy of the base classifier. Our experiments consist of the following steps.

Preprocessing: Since we assume that the data set contains only binary data, we first discretize the original non-binary data to become binary. We split the value of each attribute from the median point of the range of the attribute. After preprocessing, we randomly divided the data sets into a training data set D (80%) and a testing data set B (20%). Note that B will be used for comparing our results with the benchmark results.

Benchmark: We use D and the original NB classification algorithm to build a classifier T_D; we use the data set B to test the classifier, and get an accuracy score. We call this score the original accuracy (or the benchmark score).

θ Selection: For $\theta = 0, 0.05, 0.1$, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0, we conduct the following 4 steps:

1) Randomization: We create a disguised data set G. For each record in the training data set D, we generate a random number r from 0 to 1 using uniform distribution. If $r < \theta$, we copy the record of D to G without any change; if $r \geq \theta$, we randomly generate the values for a record of Y according to the pre-defined probability and copy the record values to G. In this paper, each record of Y is randomly generated such that each logical expression (Y) appears with the probability of 0.5. That is $W_y = 0.5$. We perform this randomization step for all the records in the training data set D , then generate the new data set G.

2) Classifier Construction: We use the data set G and our modified NB classification algorithm to build a naive Bayesian classifier T_G.

3) Testing: We use the data set B to test T_G,and get an accuracy score S.

4) Repeating: We repeat Steps 1 - 3 for 1000 times, and get $S_1, \ldots, S_{1000}$. We then compute the mean and the variance of these 1000 accuracy scores.

5.2 Accuracy Analysis

5.2.1 The Analysis of Mean

Figures 2 and 3 show the mean values of the accuracy scores for *Adult* and *Breast-Cancer* data sets respectively. We can see from the figures that when $\theta = 1$, the results are exactly the same as the results when the standard, unmodified classification algorithm is applied. This is because when $\theta = 1$, the randomized data sets are all from the private data D. For θ approaching 1, the contribution of the private data is enhanced; with θ deviating from 1, the contribution of the private data is decreasing (when θ is 0, the collected data set is all from the personal data). Therefore, when θ moves from 1 towards 0, the mean of accuracy has the trend of decreasing.

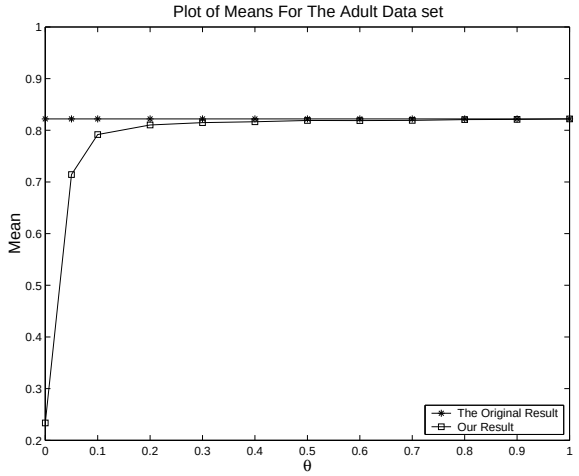

Figure 2: Plot of means for the adult data set

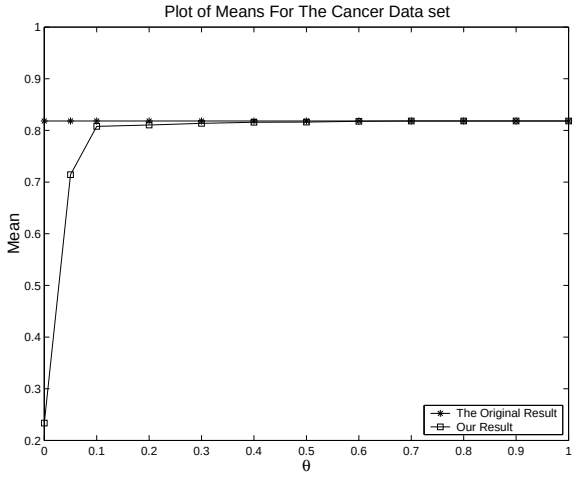

Figure 3: Plot of means for the cancer data set

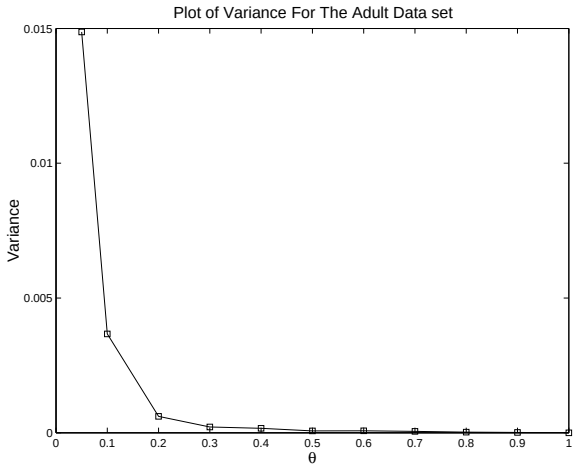

Figure 4: Plot of variance for the adult data set

5.2.2 The Analysis of Variance

Figures 4 and 5 shows the variances of the accuracy scores. When θ moves from 1 towards 0, the degree of randomness in the disguised data is increasing, the variance of the estimation used in our method becomes larger. The variance changes with different randomization levels (θ). When θ is near 0, the randomization level is much higher and the private data is better disguised. We do not show the variance when $\theta = 0$. In this case, since the collected data set is actually the personal data and the probability distribution for it is always the same for each iteration, the variance is 0.

5.3 Privacy Analysis

To get better sense of our proposed privacy measure, we conduct a set of experiments on the data sets with various distributions. Since we don't know inference rules after mining, we solely evaluate privacy before mining. Specially, we conduct experiments when $W_a = 0.1, 0.2, 0.3, 0.4, 0.5$. For each data distribution, we compute the privacy value for the cases where $\theta = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$. Notice that the privacy when $W_a = 0.9, 0.8, 0.7, 0.6$ is symmetric with the privacy when $W_a = 0.1, 0.2, 0.3, 0.4$. Therefore, I only evaluate half of them. As we see from results in Figure 6.

- When $\theta = 1$, the private data is fully disclosed. Privacy value is 0;
- When $\theta = 0$, the data collector gets no private data, and the data obtained are all the personal data. In this case, the privacy level of private data is the highest.
- When θ is away from 1 and approaches 0, the elements of the private data contribute less to the classification, and the probability of disclosing the private

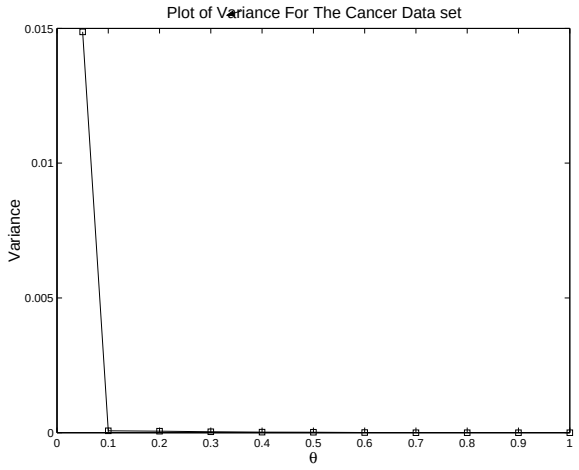

Figure 5: Plot of variance for the cancer data set

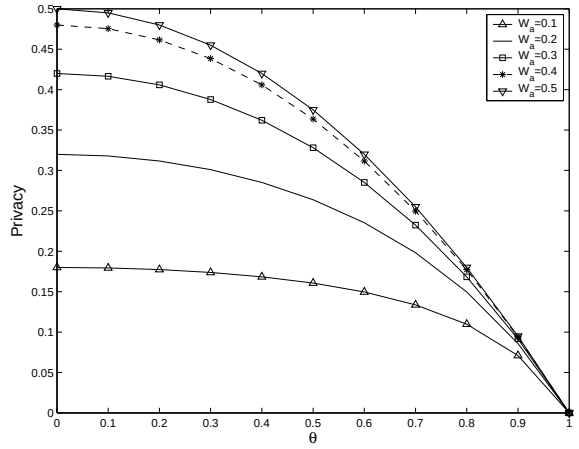

Figure 6: Privacy plot

data is decreasing. Therefore the privacy level of the private data increases.

- When private data (A) distribution approaches to uniform ($W_a = 0.5$), the privacy level is increasing. Since the uniform distribution will make the original data recoverability be the lowest.

6 Discussion

6.1 The Unrelated-Question Model

There are two types of estimation models for randomized response technique [11]. One is the related-question model where two questions are related and answers for the two questions are opposite (i.e., one question is "Is it true that you have attribute K?" and the other is "Is it true that you do not have attribute K?"). The second type is the unrelated-question model where two questions are unrelated as we discussed in this paper. A multi-variant randomized response technique (MRR) based on the related-question model to deal with multiple attributes has been proposed in [3]. Support for the latter type comes from research showing [6] that the respondents might be more cooperative provided that they could reply to one of the two questions where one question is totally unrelated to the private attribute. Actual survey results were reported to illustrate that the unrelated-question model did increase the respondent's probability of telling the truth. Theoretical framework provided in [5] proved that the unrelated-question model actually reduced the resultant variance. Figure 7 further compares naive Bayesian classification results of two models on the *Adult* data set (limited by the space, the similar results of *Breast Cancer* data set is not shown here).

We observe that there are three advantages of using unrelated-question model:

1) When $\theta = 0.5$, although this θ value provides the highest privacy for related-question model, the related-question model cannot be applied. The unrelated-question model does not suffer from this restriction.

2) For the related-question model, the results for $\theta \in [0, 0.5)$ are similar to the results for $\theta \in (0.5, 1]$, therefore, the actual privacy level is limited to half of the whole possible domain [0,1]. But the privacy level for unrelated-question model can take all the values in the whole possible domain.

3) From the Figure 7, we can see that unrelated-question model provides better results than the related-question model when θ is close to 0.5. The best privacy in related-question model is achieved as θ close to 0.5. However, as Figure 7 shows, the results of related-question model for these values of θ are not as good as the results of the unrelated-question model.

6.2 How to Implement a Web-based Randomized Response Scheme

In the original paper on randomized response technique [11], the scheme was paper-based. However, our goal is to implement the proposed scheme in a web-based system.

Assume the server is maintained by the data collector. The interface for respondents is depicted in Figure 8. For the sake of simplicity, we only show single attribute case (the scheme for the multi-variant (multiple attributes) case can be similarly derived). Each respondent uses a randomization device. For each question, the respondent generates a random number. If the number is greater than the pre-defined threshold (θ), the respondent clicks the corresponding *Yes/No* button. For instance, if the answer to a private question is Yes and the answer to personal question is No. The respondent generates a random number (r) using her randomization device. If $r > \theta$, she clicks the Yes button, else she clicks the No

θ	0.51	0.6	0.7	0.8	0.9
Mean	0.66	0.81	0.82	0.82	0.82
Variance	0.0054	0.0002	0.0001	0	0

(a) Related-Question Model

θ	0.5	0.51	0.6	0.7	0.8	0.9
Mean	0.81	0.81	0.82	0.82	0.82	0.82
Variance	0.0001	0.0001	0.0001	0.0001	0	0

(b) Unrelated-Question Model

Figure 7: The comparison on the adult data set

button. Since the data collector (server side) does not know the random number generated by the respondent, he cannot know whether the respondent answers a private or a personal question. Note that both data collector and respondents know θ, but each random number generated by each respondent is known only by the respondent herself.

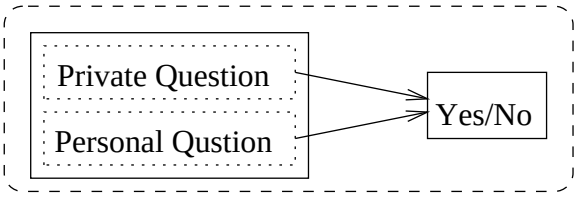

Figure 8: Web-based randomized response scheme

6.3 How to Increase the Data Privacy Level

The proposed scheme (Equation 1) is considered a single-group scheme since all the attributes are grouped together. For each record in the collected data set, if the data collector somehow finds out the respondent tells the private information about one attribute (question), the data collector then knows the respondent tells the private information about other attributes as well. To enhance data privacy, respondents can divide all the attributes into two or more groups (all the respondents should group the attributes in the same way, e.g., one respondent lets attribute $A_1(Y_1)$ and $A_2(Y_2)$ to be in the group 1, then other respondents also let attribute $A_1(Y_1)$ and $A_2(Y_2)$ to be in the group 1). They then apply the multi-variant randomized response techniques for each group *independently* such that knowing information about attributes in one group cannot hurt the privacy of the information for the other group.

What happens when the data collector somehow finds out the answer to a personal question, e.g., *Do you live near a lake?*. In this case, the data collector still does not know whether the respondent answered this personal question, or the private question associated with it. So even the knowledge of an answer to a personal question does not necessarily compromise the privacy of the answer to a private question.

7 Concluding Remarks

In this paper, we have presented a method to build naive Bayesian classifiers using multi-variant randomized response technique. The experimental results show that when we select an appropriate randomization parameter θ, we can get fairly accurate classifiers comparing to the classifiers built from the undisguised data. The proposed multi-variant unrelated-question model can be used not only for naive Bayesian classification, but also can be utilized in many other privacy-preserving data mining computations, such as decision tree induction, Bayesian classification, probabilistic-based clustering. As future work, we will apply the proposed scheme to other data mining problems.

Acknowledgements

The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada and of the Communications and Information Technology Ontario for this research.

References

[1] R. Agrawal and R. Srikant, "Privacy-preserving data mining", in *Proceedings of The ACM SIGMOD Conference On Management of Data*, Dallas, Texas, USA, 2000.

[2] L. F. Cranor, J. Reagle and M. S. Ackerman, "Beyond concern: Understanding net users' attitudes about online privacy", AT&T Labs-Research, Apr, 1999.

[3] W. Du and Z. Zhan, "Using randomized response techniques for privacy-preserving data mining", in *Proceedings of The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 24-27, Washington, DC, USA, Aug, 2003.

[4] A. Evfimievski, R. Srikant, R. Agrawal, and J. E. Gehrke, "Privacy preserving mining of association rules", in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 2000.

[5] B. G. Greenberg, A. A. Abul-Ela, W. R. Simmons and D. G. Horvitz, "The unrelated question randomized response model: Theoretical framework", *the American Statistical Association*, vol. 64, pp. 520-539, 1969.

[6] D. G. Horvitz, B. V. Shah, and R. Walt, "The unrelated question randomized response model", in *Social Statistics Section Proceedings of the American Statistical Association*, pp. 65-72, 1967.
[7] P. Langley, W. Iba, and K. Thompson, "An analysis of Bayesian classifiers", *National Conference on Artificial Intelligence*, pp. 223-228, 1992.
[8] Y. Lindell and B. Pinkas, "Privacy preserving data mining", in *Advances in Cryptology-CRYPTO'00*, LNCS 1880, pp. 36-54, Spinger-Verlag, 2000.
[9] S. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining", in *Proceedings of the 28th VLDB Conference*, Hong Kong, China, 2002.
[10] J. Vaidya and C. Clifton, "Privacy-preserving association rule mining in vertically partitioned data", in *Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 23-26, July 2002.
[11] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias", *The American Statistical Association*, vol. 60, no. 309, pp. 63-69, Mar. 1965.
[12] A. C. Yao, "Protocols for secure computations", in *Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science*, 1982.
[13] Z. Zhan and S. Matwin, "Privacy-preserving data mining in electronic surveys", in *The 4th International Conference on Electronic Business*, Beijing, China, pp. 5-9, Dec. 2004.

Justin Zhan will join the Heinz School of Information Networking Institute in Carnegie Mellon University as a faculty member in Fall 2006. His research interest contains privacy and security issues in data mining, network security and wireless network security.

Stan Matwin is a professor at the School of Information Technology and Engineering, University of Ottawa, where he directs the Text Analysis and Machine Learning (TAMALE) lab. His research is in machine learning, data mining, and their applications, as well as in Privacy-Enhancing Technologies. Former president of the Canadian Society for the Computational Studies of Intelligence (CSCSI) and of the IFIP Working Group 12.2 (Machine Learning). Member of the Board of the Centre for Communications and Information Technology of the Ontario Centre of Excellence, he is an Ontario Champion of Innovation. Program Committee Chair and Area Chair for a number of international conferences in AI and Machine Learning. Member of the Editorial Boards of the Machine Learning Journal, Computational Intelligence Journal, Journal of AI Research, and the Intelligent Data Analysis Journal.