

Performance Measurements for Privacy Preserving Data Mining

Nan Zhang, Wei Zhao, and Jianer Chen

Department of Computer Science, Texas A&M University,
College Station, TX 77843, USA

{nzhang, zhao, chen}@cs.tamu.edu

Abstract. This paper establishes the foundation for the performance measurements of privacy preserving data mining techniques. The performance is measured in terms of the accuracy of data mining results and the privacy protection of sensitive data. On the accuracy side, we address the problem of previous measures and propose a new measure, named “effective sample size”, to solve this problem. We show that our new measure can be bounded without any knowledge of the data being mined, and discuss when the bound can be met. On the privacy protection side, we identify a tacit assumption made by previous measures and show that the assumption is unrealistic in many situations. To solve the problem, we introduce a game theoretic framework for the measurement of privacy.

1 Introduction

In this paper, we address issues related to the performance measurements of privacy preserving data mining techniques. The objective of privacy preserving data mining is to enable data mining without violating the privacy of data being mined.

We consider a distributed environment where the system consists of a data miner and numerous data providers. Each data provider holds one private data point. The data miner performs data mining tasks on the (possibly perturbed) data provided by the data providers. A typical example of this kind of system is online survey, as the survey analyzer (data miner) collects data from thousands of survey respondents (data providers). Most existing privacy preserving algorithms in such system use an randomization approach which randomizes the original data to protect the privacy of data providers [1, 2, 3, 4, 5, 6, 8].

In this paper, we establish the foundation for analyzing the tradeoff between the accuracy of data mining results and the privacy protection of sensitive data. Our contribution can be summarized as follows.

- On accuracy side, we address the problem of previous measures and propose a new accuracy measure, named “effective sample size”, to solve this problem. We show that our new measure can be upper bounded without any knowledge of the data being mined and discuss when the bound can be met.
- On privacy protection side, we show that a tacit assumption made by previous measures is that all adversaries use the same intrusion technique to invade privacy. We address the problems of this assumption and propose a game theoretic formulation which takes the adversary behavior into consideration.

2 System Model

Let there be n data providers C_1, \dots, C_n and one data miner S in the system. Each data provider C_i has a private data point (e.g., transaction, data tuple, etc.) x_i . We consider the original data values x_1, \dots, x_n as n independent and identically distributed (i.i.d.) variables that have the same distribution as a random variable X . Let the domain of X (i.e., the set of all possible values of X) be V_X , and the distribution of X be p_X . As such, each data point x_i is i.i.d. on V_X with distribution p_X .

Due to the privacy concern of data providers, we classify the data miners into two categories. One category is honest data miners. These data miners always act honestly in that they only perform regular data mining tasks and have no intention to invade privacy. The other category is malicious data miners. These data miners would purposely compromise the privacy of data providers.

3 Related Work

To protect the privacy of data providers, countermeasures must be implemented in the data mining system. Randomization is a commonly used approach. It is based on an assumption that accurate data mining results can be obtained from a robust estimation of the data distribution [2]. Thus, the basic idea of randomization approach is to distort individual data values but keep an accurate estimation of the data distribution.

Based on the randomization approach, the privacy preserving data mining process can be considered as a two-step process. In the first step, each data provider C_i perturbs its data x_i by applying a randomization operator $R(\cdot)$ on x_i , and then transfers the randomized data $R(x_i)$ to the data miner. We note that $R(\cdot)$ is known by both the data providers and the data miner. Let the domain of $R(x_i)$ be V_Y . The randomization operator $R(\cdot)$ is a function from V_X to V_Y with transition probability $p[x \rightarrow y]$. Existing randomization operators include random perturbation operator [2], random response operator [4], MASK distortion operator [8], and “select-a-size” operator [6].

In the second step, a honest data miner first employs a distribution reconstruction algorithm on the aggregate data, which intends to reconstruct the original data distribution from the randomized data. Then, the honest data miner performs the data mining task on the reconstructed distribution. Various distribution reconstruction algorithms have been proposed [1, 2, 6, 4, 8]. Also in the second step, a malicious data miner may invade privacy by using a private data recovery algorithm. This algorithm is used to recover individual data values from the randomized data supplied by the data providers.

Clearly, any privacy preserving data mining technique should be measured by its capability of both constructing the accurate data mining results and protecting individual data values from being compromised by the malicious data miners.

4 Quantification of Accuracy

In previous studies, several accuracy measures have been proposed. We classify these measures into two categories. One category is application-specified accuracy measures

[8]. Measures in this category are similar to those in systems without privacy concern and are specific to a data mining task (e.g., classification, association rule mining, etc.). The other category is general accuracy measures. Measures in this category can be applied to any privacy preserving data mining systems based on the randomization approach. An existing measure in this category is information loss measure [1], which is in proportion to the expected error of the reconstructed distribution.

We remark that the ultimate goal of the performance measurements is to help the system designers to choose the optimal randomization operator. As we can see from the privacy preserving data mining process, the randomization operator has to be determined before any data is transferred from the data providers to the data miner. Thus, in order to reach its goal, a performance measure must be estimated or bounded without any knowledge of the data being mined. As we can see, the application-specified accuracy measures depend on both the reconstructed data distribution and the performance of data mining algorithm. The information loss measure depends on both the original distribution and the reconstructed distribution. Neither measure can be estimated or bounded when the original data distribution is not known. Thus, previous measures cannot be used by the system designers to choose the optimal randomization operator.

We propose effective sample size as our new accuracy measure. Roughly speaking, given the number of randomized data points, the effective sample size is in proportion to the minimum number of original data points needed to make an estimate of the data distribution as accurate as the distribution reconstructed from the randomized data points. The formal definition is stated as follows.

Definition 1. Given randomization operator $R : V_X \rightarrow V_Y$, let \tilde{p} be the maximum likelihood estimate of the distribution of x_i reconstructed from $R(x_1), \dots, R(x_n)$. Let $\tilde{p}_0(k)$ be the maximum likelihood estimate of the distribution based on k variables randomly generated from the distribution p_X . We define the effective sample size r as the minimum value of k/n such that

$$D_{\text{Kol}}(\tilde{p}_0(k), p_X) \leq D_{\text{Kol}}(\tilde{p}, p_X) \quad (1)$$

where D_{Kol} is the Kolmogorov distance [7], which measures the distance between an estimated distribution and the theoretical distribution¹.

As we can see, effective sample size is a general accuracy measure which measures the accuracy of the reconstructed distribution. We now show that the effective sample size can be strictly bounded without any knowledge of p_X .

Theorem 1. Recall that $p[x \rightarrow y]$ is the probability transition function of $R : V_X \rightarrow V_Y$. An upper bound on the effective sample size r is given as follows.

$$r \leq 1 - \sum_{y \in V_Y} \min_{x \in V_X} p[x \rightarrow y]. \quad (2)$$

Due to space limit, please refer to [9] for the proof of this theorem.

¹ Other measures of such distance (e.g., Kuiper distance, Anderson-Darling distance, etc) can also be used to define the effective sample size. The use of other measures does not influence the results in this paper.

5 Quantification of Privacy Protection

In previous studies, two kinds of privacy measures have been proposed. One is information theoretic measure [1], which measures privacy disclosure by the mutual information between the original data x_i and the randomized data $R(x_i)$ (i.e., $I(x_i; R(x_i))$). This measure was challenged in [5], where it is shown that certain kinds of privacy disclosure cannot be captured by the information theoretic measure. The other kind of privacy measure can be used to solve this problem [5, 10, 2]. In particular, the privacy breach measure [5] defines the level of privacy disclosure as $\max_{x, x' \in V_X} p[x \rightarrow y]/p[x' \rightarrow y]$ for any given $y \in V_Y$. This measure captures the *worst case* privacy disclosure but is (almost) independent of the average amount of privacy disclosure.

Note that the data miner has the freedom to choose different intrusion techniques in different circumstances. As such, the privacy protection measure should depend on two important factors: a) the privacy protection mechanism of the data providers, and b) the unauthorized intrusion technique of the data miner. However, previous measures do not follow this principle. Instead, they make a tacit assumption that all data miners will use the same intrusion technique. This assumption seems to be reasonable as a (rational) data miner will always choose the intrusion technique which compromises the most private information. However, as we will show in the following example, the optimal intrusion technique *varies* in different circumstances. Thereby, the absence of consideration of intrusion techniques results in problems of privacy measurement.

Example 1. Let there be $V_X = \{0, 1\}$. The original data x_i is uniformly distributed on V_X . The system designer needs to determine which of the following two randomization operators, R_1 and R_2 , discloses less private information.

$$R_1(x) = \begin{cases} x, & \text{with probability 0.70,} \\ \bar{x}, & \text{with probability 0.30.} \end{cases} \quad R_2(x) = \begin{cases} 0, & \text{if } x = 0, \\ 1, & \text{if } x = 1, \text{ with probability 0.01,} \\ 0, & \text{if } x = 1, \text{ with probability 0.99.} \end{cases}$$

In the example, we have $I(x; R_1(x)) \gg I(x; R_2(x))$. Due to the information theoretic measure, R_2 discloses less privacy. However, R_2 discloses more privacy due to the privacy breach measure. The reason is that if the data miner receives $R_2(x_i) = 1$, then it can always infer that $x_i = 1$ with probability of 1. We now show that whether R_1 or R_2 discloses more private information actually *depends* on the system setting. In particular, we consider the following two system settings.

1. The system is an online survey system. The value of x_i indicates whether a survey respondent is interested in buying certain merchandise. A malicious data miner intends to make unauthorized advertisement to data providers with such interest.
2. The system consists of n companies as the data providers and a management consulting firm as the data miner. The original data x_i contains the expected profit of the company which has not been published yet. A malicious data miner may use x_i to make investment on a high-risk stock market. The profit from a successful investment is tremendous. However, an unsuccessful investment results in a loss five times greater than the profit the data miner may obtain from a successful investment.

In the first case, an advertisement to a wrong person costs the data miner little. As we can see, R_1 discloses the original data value with probability of 0.7, which is greater than that of R_2 (0.501). Thus, R_2 is better than R_1 in the privacy protection perspective.

In the second case, the data miner will not perform the intrusion when R_1 is used by the data providers. The reason is that the loss from an incorrect estimate of x_i is too high to risk. As we can see, when R_1 is used, the expected net benefit from an unauthorized intrusion is less than 0. However, the data miner will perform the intrusion when R_2 is used. The reason is that when $R_2(x_i) = 1$, the data miner has a fairly high probability (99%) to make a successful investment. If a randomized data $R_2(x_i) = 0$ is received, the data miner will simply ignore it. As such, in this case, R_1 is better than R_2 in the privacy protection perspective.

As we can see from the example, the data miner will choose different privacy intrusion techniques in different system settings (in the above example, there is an intruder-or-not selection), which will result in different performance of randomization operators. Thus, the system setting has to be considered in the measurement of privacy disclosure.

In order to introduce the system setting and the privacy intrusion technique to our privacy measure, we propose a game theoretic framework to analyze the strategies of the data miner (i.e., privacy intrusion technique). Since we are studying the privacy protection performance of the randomization operator, we consider the randomization operator as the strategy of the data providers.

We model the privacy preserving data mining process as a non-cooperative game between the data providers and the data miner. There are two players in the game. One is the data providers. The other is the data miner. Since we only consider the privacy measure, the game is zero-sum in that the data miner can only benefit from the violation of privacy of the data providers. Let S_c be the set of randomization operators that the data providers can choose from. Let S_s be the set of the intrusion techniques that the data miner can choose from. Let u_c and u_s be the utility functions (i.e., expected benefits) of the data providers and the data miner, respectively. Since the game is zero-sum, we have $u_c + u_s = 0$. We remark that the utility functions depend on both the strategies of the players and the system setting.

We assume that both the data providers and the data miner are rational. As such, given a certain randomization operator, the data miner always choose the privacy intrusion technique which maximizes u_s . Given a certain privacy intrusion technique, the data providers always choose the randomization operator which maximizes u_c . We now define our privacy measure based on the game theoretic formulation.

Definition 2. *Given a privacy preserving data mining system $G \langle S_s, S_c, u_s, u_c \rangle$, we define the privacy measure l_p of a randomization operator R as*

$$l_p(R) = u_c(R, L_0), \quad (3)$$

where L_0 is the optimal privacy intrusion technique for the data miner when R is used by the data providers, u_c is the utility function of the data providers when R and L_0 are used.

As we can see, the smaller $l_p(R)$ is, the more benefit is obtained by the data miner from the unauthorized intrusion. Let σ be the ratio between the benefit obtained by

a malicious data miner from a correct estimate and the loss of it from an incorrect estimate. A useful theorem is provided as follows.

Theorem 2. *Let there be $\max_{x_0 \in V_X} \Pr\{x_i = x_0\} = p_m$ in the original data distribution. We have $l_p(R) = 0$ if the randomization operator $R : V_X \rightarrow V_Y$ satisfies*

$$\max_{y \in V_Y} \frac{\max_{x \in V_X} p[x \rightarrow y]}{\min_{x \in V_X} p[x \rightarrow y]} \leq \frac{1 - p_m}{\sigma p_m}. \quad (4)$$

Please refer to [9] for the proof of this theorem.

6 Conclusion

In this paper, we establish the foundation for the measurements of accuracy and privacy protection in privacy preserving data mining. On accuracy side, we address the problem of previous accuracy measures and solve the problem by introducing an effective sample size measure. On privacy protection side, we present a game theoretic formulation of the system and propose a privacy protection measure based on the formulation. Our work is preliminary, and there are many possible extensions. We are currently investigating using our performance measurements to derive the optimal trade-off between accuracy and privacy which can be achieved by the randomization approach.

References

1. D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 247–255. ACM Press, 2001.
2. R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 439–450. ACM Press, 2000.
3. W. Du and M. Atallah. Privacy-preserving cooperative statistical analysis. In *Proceedings of the 17th Annual Computer Security Applications Conference*, page 102, Washington, DC, USA, 2001. IEEE Computer Society.
4. W. Du and Z. Zhan. Using randomized response techniques for privacy-preserving data mining. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 505–510, New York, NY, USA, 2003. ACM Press.
5. A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222, New York, NY, USA, 2003. ACM Press.
6. A. Evfimievski, R. Srikant, R. Agarwal, and J. Gehrke. Privacy preserving mining of association rules. *Inf. Syst.*, 29(4):343–364, 2004.
7. F. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253).
8. S. Rizvi and J. Haritsa. Maintaining data privacy in association rule mining, 2002.

9. N. Zhang, W. Zhao, and J. Chen. On the performance measurement for privacy preserving data mining. technical report, 2004.
10. Y. Zhu and L. Liu. Optimal randomization for privacy preserving data mining. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 761–766, New York, NY, USA, 2004. ACM Press.