When do Data Mining Results Violate Privacy?*

Murat Kantarcıoğlu
Purdue University
Computer Sciences
250 N University St
West Lafayette, IN
47907-2066
kanmurat@cs.purdue.edu

Jiashun Jin
Purdue University
Statistics
150 N University St
West Lafayette, IN
47907-2067
jinj@stat.purdue.edu

Chris Clifton
Purdue University
Computer Sciences
250 N University St
West Lafayette, IN
47907-2066
clifton@cs.purdue.edu

ABSTRACT

Privacy-preserving data mining has concentrated on obtaining valid results when the input data is private. An extreme example is Secure Multiparty Computation-based methods, where only the results are revealed. However, this still leaves a potential privacy breach: Do the results themselves violate privacy? This paper explores this issue, developing a framework under which this question can be addressed. Metrics are proposed, along with analysis that those metrics are consistent in the face of apparent problems.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; H.2.7 [**Database Management**]: Database Administration—*Security, integrity, and protection*

General Terms

Security

Keywords

Privacy, Inference

1. INTRODUCTION

There has been growing interest in privacy-preserving data mining, with attendant questions on the real effectiveness of the techniques. For example, there are discussions about the effectiveness of adding noise to data: while adding noise to a single attribute can be effective [3, 2], the adversary could have much higher ability to recover individual values for multiple correlated attributes [12]. An alternative encryption based approach was proposed in [14]: nobody learns anything they didn't already know, except the resulting data mining model. While [14] only discussed the case for two parties, it has been shown in [10] that this is also feasible for many parties (e.g., rather than providing "noisy" survey results as in [3], individuals provide encrypted survey results that can be used to generate the resulting data mining model.) This is discussed further in Section 4.

However, though these provably secure approaches reveal nothing but the resulting data mining model, they still leave a privacy question open: Do the resulting data mining models inherently violate privacy?

This paper presents a start on methods and metrics for evaluating the privacy impact of data mining models. While the methods are preliminary, they provide a cross-section of what needs to be done, and a demonstration of techniques to analyze privacy impact. Work in privacy-preserving data mining has shown how to build models when the training data is kept from view; the full impact of privacy-preserving data mining will only be realized when we can guarantee that the resulting models do not violate privacy.

To make this clear, we present a "medical diagnosis" scenario. Suppose we want to create a "medical diagnosis" model for public use: a classifier that predicts the likelihood of an individual getting a terminal illness. Most individuals would consider the classifier output to be sensitive – for example, when applying for life insurance. The classifier takes some public information (age, address, cause of death of ancestors), together with some private information (eating habits, lifestyle), and gives a probability that the individual will contract the disease at a young age. Since the classifier requires some information that the insurer is presumed not to know, can we state that the classifier does not violate privacy?

The answer is not as simple as it seems. Since the classifier uses some public information as input, it would appear that the insurer could *improve* an estimate of the disease probability by repeatedly probing the classifier with the known public information and "guesses" for the unknown information. At first glance, this appears to be a privacy violation. Surprisingly, as we show in Section 1.1, given reasonable assumptions on the external knowledge available to an adversary we can *prove* the adversary learns nothing new.

We assume that data falls into three classes:

- **Public Data:**(P) This data is accessible to every one including the adversary.
- **Private/Sensitive Data:**(S) We assume that this kind of data must be protected: The values should remain unknown to the adversary.

*This material is based upon work supported by the National Science Foundation under Grant No. 0312357.

KDD'04, August 22–25, 2004, Seattle, Washington, USA.

- **Unknown Data:**(U) This is the data that is not known to the adversary, and is not *inherently* sensitive. However, before disclosing this data to an adversary (or enabling an adversary to estimate it, such as by publishing a data mining model) we must show that it does not enable the adversary to discover sensitive data.

1.1 Example: Classifier Predicting Sensitive Data

The following example shows that for the "medical diagnosis" scenario above, it is reasonable to expect that publishing the classifier will not cause a privacy violation. Individuals can use the classifier to predict their own likelihood of disease, but the adversary (insurer) does not gain any *additional* ability to estimate the likelihood of the disease.

To simplify the problem, we assume that the classifier is a "black-box": the adversary may probe (use the classifier), but cannot see inside. An individual can use the classifier without any risk of disclosing either their private data or their private result.[1] This represents a best-case scenario: If this classifier violates privacy, then no approach (short of limiting the adversary's access to the classifier) will provide privacy protection.

Formally, suppose $X = (P, U)^T$ is distributed as $N(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \tag{1}$$

where $-1 < r < 1$ is the correlation between P and U. Assume that for n independent samples $(x_1, x_2, \ldots, x_n)$ from $N(0, \Sigma)$, the sensitive data $S = (s_1, s_2, \ldots, s_n)$ can be discovered by a classifier C_0 that compares the public data p_i and the unknown data u_i:

$$s_i = C_0(x_i) = \begin{cases} 1 & \text{if } p_i \geq u_i \\ 0 & \text{otherwise;} \end{cases}, \text{ where:} \tag{2}$$

- each p_i is a public data item that everyone can access,
- the data items denoted by u_i are unknown to the adversary; u_i is only know to the i-th individual,
- each s_i is sensitive data we need to protect, and
- The adversary knows that $X \sim N(0, \Sigma)$, it may or may not know r.

We now study whether publishing the classifier C_0 violates privacy, or equivalently, whether the adversary can get a better estimate of any s_i by probing C_0.

Given the public data p_i for an individual i, the adversary could try to probe the classifier C_0 to get an estimate of s_i as follows. It is reasonable to assume that the adversary has knowledge of the (marginal) distribution that the u_i are sampled from; we can even assume that the adversary knows the joint distribution that $(p_i, u_i)^T$ are sampled from, or equivalently Σ or r. (We will see soon that though the adversary seems to know a lot, he doesn't know anything more about the s_i – this makes our example more surprising). Thus for each individual or for each p_i, the adversary could sample $\tilde{u}_i$ from the conditional distribution of $(U|P)$, he then can use the pairs $(p_i, \tilde{u}_i)^T$ to probe C_0 and get an estimate $\tilde{s}_i \triangleq C_0(p_i, \tilde{u}_i)$. Assuming that the information P was correlated with S, this will give the adversary a better estimate than simply taking the most likely result in S.

However, this assumes the adversary has no prior knowledge. In our medical example, it is likely that the adversary has some knowledge of the relationship between P and S. For example, cause of death is generally public information, giving the adversary a training set (Likely as complete as that used to generate C_0, as for some diseases – Creutzfeldt-Jakob, Alzheimer's until recently – an accurate diagnosis required post-mortem examination, so the training data for C_0 would likely be deceased individuals.)

Given that the adversary has this knowledge, what does the adversary know if we do not publish C_0? Notice that

$$Pr\{S = 1|P = p\} = \Phi(\frac{1-r}{\sqrt{1-r^2}}p) \tag{3}$$

$$= \begin{cases} \geq 1/2, & \text{if } p \geq 0, \\ < 1/2, & \text{otherwise,} \end{cases} \tag{4}$$

where $\Phi(\cdot)$ is the cdf of $N(0, 1)$. According to (3), (or even just based on symmetry), the best classifier the adversary can choose in this situation is:

$$s_i = \begin{cases} 1 & \text{if } p_i > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

Let C_1 denote this classifier.

Next, we study what the adversary knows if we publish the classifier C_0. We even allow the adversary to know r. In this situation, the best classifier the adversary can use is the Bayesian estimator C_2, which is based on the probability of $Pr\{U \leq P|P = p_i\}$:

$$s_i = \begin{cases} 1 & \text{if } Pr\{U \leq P|P = p_i\} > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}, \tag{6}$$

However, notice that

$$\Phi(\frac{1-r}{\sqrt{1-r^2}}p_i) = Pr\{U \leq P|P = p_i\}$$

compare this to (3), we conclude that $C_1 \equiv C_2$.

Thus in this situation, publishing C_0 or even the key parameter r doesn't give the adversary any additional capability, as long as the adversary has no access to the u_i. This enables us to argue that even though C_0 apparently reveals sensitive information, it does *not* actually violate privacy.

1.2 Contribution of this Paper

As the above example demonstrates, determining if a data mining model violates privacy requires knowing many things: What information is sensitive? To whom is it sensitive? What else is known? Whose privacy is at risk? What is an acceptable tradeoff between privacy and the benefit of the data mining result, and how do we measure this tradeoff?

In this paper, we suggest a framework where some of the above questions can be answered. We give precise definitions for privacy loss due to data mining results. A formal analysis of those definitions are provided for some examples, as well as empirical evaluations showing how the models could be applied in real life.

Specifically, in Section 2, we present a model that enables us to discuss these issues in the context of classification. Section 3 presents a metric for privacy loss for one such situation, including examples of when the metric would be appropriate and how the metric could be calculated (analytically or empirically) in specific situations.

[1] This is feasible, for examples see [9].

2. THE MODEL FOR PRIVACY IMPLICATIONS OF DATA MINING RESULTS

To understand the privacy implications of data mining results, we first need to understand how data mining results can be used (and misused). As described previously, we assume data is either *P*ublic, *U*nknown, or *S*ensitive. We now discuss additional background leading toward a model for understanding the impact of data mining results on privacy.

We assume an adversary with access to *P*ublic data, and polynomial-time computational power. The adversary may have some additional knowledge, possibly including *U*nknown and *S*ensitive data for some individuals. We want to analyze the effect of giving the adversary access to a classifier C; specifically if it will improve the ability of the adversary to accurately deduce *S*ensitive data values for individuals that it doesn't already have such data for.

2.1 Access to Data Mining Models

If the classifier model C is completely open (e.g., a decision tree, or weights in a neural network), the model description may reveal sensitive information. This is highly dependent on the model.

Instead, we model C as a "black box": The adversary can request that an instance be classified, and obtain the class, but can obtain no other information on the classifier. This is a reasonable model: We are providing the adversary with *access* to C, not C itself. For example, for the proposed CAPPSII airline screening module, making the classifier available would give terrorists information on how to defeat it. However, using cryptographic techniques we can provide privacy for all parties involved: Nothing is revealed but the class of an instance[9]. (The party holding the classifier need not even learn attribute values.)

Here, we will only consider the data mining results in the form of classification models. We leave the study of other data mining results as future work.

2.2 Basic Metric for Privacy Loss

While it is nice to show that an adversary gains no privacy-violating information, in many cases we will not be able to say this. Privacy is not absolute; most privacy laws provide for cost/benefit tradeoffs when using private information. For example, many privacy laws include provisions for use of private information "in the public interest"[6]. To tradeoff the benefit vs. the cost of privacy loss, we need a metric for privacy loss.

One possible way to define such a metric for classifier accuracy is using the Bayesian classification error. Suppose for data $(x_1, x_2, \ldots, x_n)$, we have classification problems in which we try to classify x_i's into m classes which we labeled as $\{0, 1, \ldots, m-1\}$. For any classifier C:

$$x_i \mapsto C(x_i) \in \{0, 1, \ldots, m-1\}, \qquad i = 1, 2, \ldots, n,$$

we define the classifier accuracy for C as:

$$\sum_{i=0}^{m-1} Pr\{C(x) \neq i | z = i\} Pr\{z = i\}. \tag{7}$$

Does this protect the individual? The problem is that *some* individuals will be classified correctly: If the adversary can predict those individuals with a higher certainty than the accuracy, then the privacy loss for those individuals is worse than expected. Tightening such bounds requires that the adversary have training data, i.e., individuals for which it knows the sensitive value.

2.3 Possible Ways to Compromise Privacy

The most obvious way a classifier can compromise privacy is by taking *P*ublic data and predicting *S*ensitive values. However, there are many other ways a classifier can be misused to violate privacy. We break down the possible forms a classifier that could be (mis)used by the adversary can take.

1. $P \rightarrow S$: Classifier that produces sensitive data given public data. Metric based on accuracy of classification.

$$\sup_i \left(Pr\{C(X) \neq Y | Y = i\} - \frac{1}{n_i} \right) \tag{8}$$

2. $PU \rightarrow S$: Classifier taking public and unknown data into sensitive data. Metric same as above.

3. $PS \rightarrow P$: Classifier taking public and sensitive data into public data. Can adversary determine value of sensitive data. (May also involve unknown data, but this is a straightforward extension.)

4. The adversary has access to *S*ensitive data for some individuals. What is the effect on privacy of other individuals of classifiers as follows.

   (a) $P \rightarrow S$: Can the adversary do better with such a classifier because of their knowledge, beating the expectations of the metric for 1.

   (b) $P \rightarrow U$: Can giving the adversary a predictor for *U*nknown data improve its ability to build a classifier for *S*ensitive data?

We gave a brief example of how we can analyze problem 2 in Section 1.1. The rest of the paper looks at item 4b above, giving both analytical and empirical methods to evaluate the privacy impact of a classifier that enables estimation of unknown values.

3. CLASSIFIER REVEALING UNKNOWNS

A classifier reveals a relationship between the inputs and the predicted class. Unfortunately, even if the class value is not sensitive, such a classifier can be used to create unintended inference channels. Assuming the adversary has t samples from a distribution (P, S), it can build a classifier C_1 using those t samples. Let a_1 be the prediction accuracy of the classifier C_1. Assume a "non-sensitive" classifier $C : P \rightarrow U$ is made available to the adversary. Using C, and the t samples, the adversary can build a classifier $C_2 : P, C(P) \rightarrow S$. Let a_2 be the accuracy of the C_2. If a_2 is better than a_1, then C compromises the privacy of S.

3.1 Formal Definition

Given a distribution (P, U, S), with P public data that everyone including the adversary can access, S sensitive data we are trying to protect (but known for some individuals), and U is data not known by the adversary. A "black-box" classifier C is available to the adversary that can be used to predict U given P. Assume that t samples $((p_1, s_1), \ldots, (p_t, s_t))$ are already available to adversary, our goal is to test whether revealing C increases the ability of the adversary to predict the S values for unseen instances.

First, assume attributes P and U are independent, or more generally, though P and U are dependent, C only contains the marginal information of P. In such cases, classifier C wouldn't be much help to the adversary: as C contains no valuable information of U, we expect that C wouldn't be much more accurate than random guess, and as a result, we expect that the adversary is unable to improve his estimate about S by using C, or formally, the Bayes error for all classifiers using P only should be the same as the Bayes error for all classifiers using $(P, C(P))$.

However, it is expected that C contains information on the joint distribution of P and U (or equivalently the conditional information of $(U|P)$), otherwise C would be uninteresting (no better than a random guess.) The adversary can thus combine C or $C(P)$ with already known information of P to create an inference channel for S, and the prediction accuracy of the newly learned classifier violates privacy.

Formally, given C and t samples from P, S, letting

$$\rho(t) = \rho_{\{t;P,S\}}, \qquad \rho(t;C) = \rho_{\{t;P,C(P),S\}}$$

be the Bayes error for classifiers using P only and using $P, C(P)$ respectively; also, letting

$$\bar{\rho} = \lim_{t\to\infty} \rho(t), \qquad \bar{\rho}(C) = \lim_{t\to\infty} \rho(t;C),$$

we have the following definition:

Definition 1. For $0 < p < 1$, we call the classifier C (t, p)-privacy violating if $\rho(t;C) \leq \rho(t) - p$, and the classifier C is (∞, p)-privacy violating if $\bar{\rho}(C) \leq \bar{\rho} - p$.

The important thing to notice about the above definition is that we measure the privacy violation with respect to number of available samples t. An adversary with many training instances will probably learn a better classifier than one with few training instances.

In this case, the release of the C_1 has created a privacy threat. The main difference between this example and the one given in the Section 1 is that we put a limitation on the number of available examples to the adversary.

3.2 Analysis for Mixture of Gaussians

We now give a formal analysis of such an inference in the case of Gaussian mixtures. Although we gave our definitions for a classifier C, in the case of the Gaussian mixtures, the sensible way to model C is the conditional distribution of some particular attribute based on the other attributes. Note that C can also be viewed as a "black box".

Suppose $X = (P, U)^T$ is distributed as a n-dimensional 2-point mixture $(1-\epsilon)N(0, \Sigma) + \epsilon N(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix}. \tag{9}$$

For a set of t realizations $X = (x_1, x_2, \ldots, x_t)$ (here $x_i = (p_i, u_i)^T$), t sensitive data $S = (s_1, s_2, \ldots, s_t)$ are generated according to the rule:

$$s_i = \begin{cases} 1, & \text{if } x_i \text{ is generated from } N(0, \Sigma), \\ 0, & \text{if } x_i \text{ is generated from } N(\mu, \Sigma). \end{cases} \tag{10}$$

Assume:

- The adversary has access to p_i, and know the marginal distribution of P in detail (this is possible for example for sufficiently large sample size t),
- The adversary has no access to u_i,
- The adversary knows that x_i are from the above 2-point mixture, he knows n, ϵ, μ_1, and Σ_{11}, which can be obtained through the marginal of P, but not μ_2 or any other entries in Σ that can not be obtained through the marginal of P.

We are concerned the following two questions.

1. What is the privacy loss by releasing u_i? In other word, what is the Bayes error when we limit the adversary's to the knowledge to the above assumption.
2. What is the privacy loss by allowing the adversary to know the conditional distribution of $(U|P)$?

Before answering these questions, we work out the Bayes error when only p_i are available and when both p_i and u_i are available. Notice here that, by symmetry, the Bayes error for t samples is the same of univariate Bayes error.

By direct calculation, the Bayes error with only p_i's is:

$$\begin{aligned}\rho(\epsilon, \mu_1, \Sigma_{11}) = (1-\epsilon) Pr\{C_B(p_i) = 1 | s_i = 0\} \\ + \epsilon Pr\{C_B(p_i) = 0 | s_i = 1\}\end{aligned}$$

where C_B is the Bayesian classifier. The Bayes error can be rewritten as:

$$\rho(\epsilon, \mu_1, \Sigma_{11}) \tag{11}$$

$$= (1-\epsilon)\bar{\Phi}\left(\frac{a + \mu_1^T \Sigma_{11}^{-1} \mu_1}{\sqrt{\mu_1^T \Sigma_{11}^{-1} \mu_1}}\right) + \epsilon\bar{\Phi}\left(\frac{a - \mu_1^T \Sigma_{11}^{-1} \mu_1}{\sqrt{\mu_1^T \Sigma_{11}^{-1} \mu_1}}\right) \tag{12}$$

where $a = \log(\frac{1-\epsilon}{\epsilon})$ and $\bar{\Phi}(\cdot)$ is the survival function of $N(0, 1)$.

In comparison, the Bayes error with both p_i's and u_i's is:

$$\begin{aligned}\rho(\epsilon, \mu, \Sigma) = (1-\epsilon) Pr\{C_B(p_i, u_i) = 1 | s_i = 0\} \\ + \epsilon Pr\{C_B(p_i, u_i\} = 0 | s_i = 1).\end{aligned}$$

This can be rewritten as:

$$(1-\epsilon)\bar{\Phi}\left(\frac{a + \mu' \Sigma^{-1} \mu}{\sqrt{\mu' \Sigma^{-1} \mu}}\right) + \epsilon\bar{\Phi}\left(\frac{a - \mu' \Sigma^{-1} \mu}{\sqrt{\mu' \Sigma^{-1} \mu}}\right).$$

We can now answer question 1:

LEMMA 3.1. *Let* $\psi(z) \stackrel{\triangle}{=} (1-\epsilon)\bar{\Phi}(\frac{a+z}{\sqrt{z}}) + \epsilon\bar{\Phi}(\frac{a-z}{\sqrt{z}})$. *Then*

1. $\psi(z)$ *strictly decreases in* z.
2. $\mu_1^T \Sigma_{11}^{-1} \mu_1 \leq \mu^T \Sigma^{-1} \mu$ *with equality if and only if* $\mu_2 = \Sigma_{12}^T \Sigma_{11}^{-1} \mu_1$.
3. *As a result,* $\rho(\epsilon, \mu, \Sigma) \leq \rho(\epsilon, \mu_1, \Sigma_{11})$, *with equality if and only if* $\mu_2 = \Sigma_{12}^T \Sigma_{11}^{-1} \mu_1$.

The proof of Lemma 3.1 is omitted. Lemma 3.1 tells us that, in general, releasing u_i's or any classifier that predicts u_i's will compromise privacy. This loss of privacy can be measured by Bayes error, which has an explicit formula and can be easily evaluated through the function $\psi(z)$.

Next, for question 2, we claim that from the privacy point of view, telling the adversary the detailed conditional distribution of $(U|P)$ is equivalent to telling the adversary all the u_i, in other words, the privacy loss for either situation are exactly the same. To see this, notice that when the adversary knows the conditional distribution of $(U|P)$, he knows the

distribution of S in detail since he already knew the marginal distribution of P. Furthermore, he can use this conditional distribution to sample u_i based on each p_i, the resulting data $s_i = (p_i, \tilde{u}_i)^T$ is distributed as $(1-\epsilon)N(0,\Sigma) + \epsilon N(\mu,\Sigma)$; though s_i's are not the data on our hand, but in essence the adversary has successfully constructed an independent copy of our data. In fact, the best classifier for either case is the Bayesian rule, which classifies s_i's to 1 or 0 according to

$$\epsilon f(x;\mu,\Sigma) \geq (1-\epsilon) f(x;0,\Sigma), \tag{13}$$

here we use $f(x;\mu,\Sigma)$ to denote the density function of $N(\mu,\Sigma)$. Thus there won't be any difference if the adversary know any u_i's of our data set, or just know the conditional distribution of $(U|P)$. This suggests that when S is highly correlated with U, revealing any good method to predict U may be problematic.

3.3 Practical Use

For most distributions it is difficult to analytically evaluate the impact of a classifier on creating an inference channel. An alternative heuristic method to test the impact of a classifier is described in Algorithm 1. We now give experiments demonstrating the use, and results, of this approach.

Algorithm 1 Testing a classifier for inference channels

1: Assume that S depends on only P, U, and the adversary has at most t data samples of the form (p_i, s_i).
2: Build a classifier C_1 on t samples (p_i, s_i).
3: To evaluate the impact of releasing C, build a classifier C_2 on t samples $(p_i, C(p_i), s_i)$.
4: If the accuracy of the classifier C_2 is significantly higher than C_1, conclude that revealing C creates a inference channel for S.

We tested this approach on several of the UCI datasets[4]. We assumed that the class variable of each data set is private, treat one attribute as unknown, and simulate the effect of access to a classifier for the unknown. For each nominal valued attribute of each data set, we ran six experiments. In the first experiment, a classifier was built without using the attribute in question. We then build a classifier with the unknown attribute correctly revealed with probability 0.6, 0.7, 0.8, 0.9, and 1.0. For example, for each instance, if 0.8 is used, the attribute value is kept the same with probability 0.8, otherwise it is randomly assigned to an incorrect value. The other attributes are unchanged.

In each experiment, we used C4.5 with default options given in the Weka package [17]. Before running the experiments, we filtered the instances with unknown attributes from the training data set. Ten-fold cross validation was used in reporting each result.

Most of the experiments look like the one shown in Figure 1 (the credit-g dataset). Giving an adversary the ability to predict unknown attributes does not significantly alter classification accuracy (at most 2%). In such situations, access to the public data may be enough to build a good classifier for the secret attribute; disclosing the unknown values to the adversary (e.g., by providing a "black box" classifier to predict unknowns) does not really increase the accuracy of the inference channel.

In a few data sets (credit-a, kr-vs-kp, primary-tumor, splice, and vote) the effect of providing a classifier on some

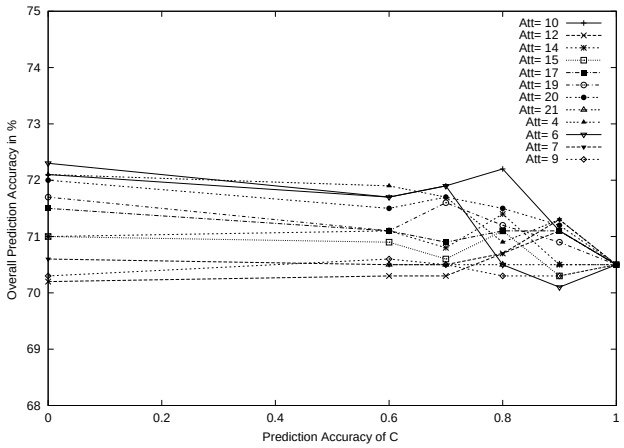

Figure 1: Effect of classification with varying quality estimate of one attribute on "credit-g" data (representative of most UCI data sets.)

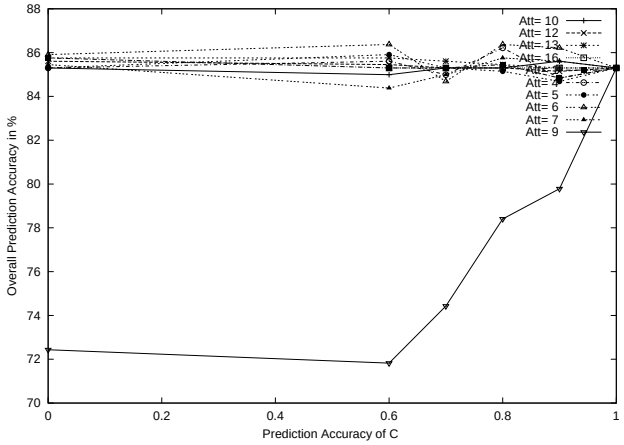

Figure 2: Effect of classification with varying quality estimate of one attribute on "credit-a" data (representative of five UCI data sets.)

attribute increased the prediction accuracy significantly. We discuss the "credit-a" data set as an example of these. If the adversary does not have an access to the 9^{th}(A9) attribute (a binary attribute), it can build a decision tree that infers the secret (class) attribute with 72% accuracy – versus 86% if given all data. This holds even if the adversary is given a classifier (C) that predicts $A9$ with 60% accuracy. However, as shown in Figure 2, if C has accuracy 80% or greater, the adversary can do a significantly better job of predicting the secret (class) attribute.

4. RELATED WORK

Privacy implications of data mining have been pointed out, a survey is given in [8]. To our knowledge, none gave precise definitions for privacy loss due to data mining results.

Considerable research has gone into privacy-preserving data mining algorithms. The goal is to learn a data mining model without revealing the underlying data. There have been two different approaches to this problem. The first is to alter the data before delivery to the data miner so that

real values are obscured, protecting privacy while preserving statistics on the collection. Recently data mining techniques on such altered data have been developed for constructing decision trees[3, 2] and association rules[15, 7]. While [2] touched on the impact of results on privacy, the emphasis was on ability to recover the altered data values rather than inherent privacy problems with the results.

The second approach is based on secure multiparty computation: privacy-preserving distributed data mining[14, 5, 11, 16, 13]. The ideas in this paper compliment this line of work. Privacy-preserving data mining tries to guarantee that nothing is revealed during the data mining process. In our case, we want to make sure that even a limited access to the data mining result does not cause a privacy threat.

The inference problem due to query results has also been addressed in a very different context: Multi-level secure databases. A survey of this work can be found in [1]. This does not address the privacy threat due to the data mining result, and does not directly apply to our problem.

5. CONCLUSIONS

Increases in the power and ubiquity of computing resources pose a constant threat to individual privacy. Tools from privacy-preserving data mining and secure multi-party computation make it possible to process the data without with disclosure, but do not address the privacy implication of the results. We have defined this problem and explored ways that data mining results can be used to compromise privacy. We gave definitions to model the effect of the data mining results on privacy, analyzed our definitions for a Mixture of Gaussians for two class problems, and gave a heuristic example that can be applied to more general scenarios.

We have looked at other situations, such as a classifier that takes sensitive data as input (can sampling the classifier with known output reveal correct values for input?) and privacy compromise from participating in training data. We are working to formalize analysis processes for these situations.

We plan to test our definitions in many different contexts. Possible plans include a software tool that automatically assesses the privacy threat due to the data mining result based on the related training instances and the private data. We also want to augment existing privacy-preserving algorithms so that the output of data mining is guaranteed to satisfy the privacy definitions, or the algorithm terminates without generating results. Finally, we want to be able extend the formal analysis to more complex data models using tools from statistical learning theory.

6. REFERENCES

[1] N. R. Adam and J. C. Wortmann. Security-control methods for statistical databases: A comparative study. *ACM Comput. Surv.*, 21(4):515–556, Dec. 1989.

[2] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 247–255, Santa Barbara, California, USA, May 21-23 2001. ACM.

[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD Conference on Management of Data*, pages 439–450, Dallas, TX, May 14-19 2000. ACM.

[4] C. Blake and C. Merz. UCI repository of machine learning databases, 1998.

[5] W. Du and Z. Zhan. Building decision tree classifier on private data. In C. Clifton and V. Estivill-Castro, editors, *IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining*, volume 14, pages 1–8, Maebashi City, Japan, Dec. 9 2002. Australian Computer Society.

[6] Directive 95/46/EC of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Communities*, No I.(281):31–50, Oct. 24 1995.

[7] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–228, Edmonton, Alberta, Canada, July 23-26 2002.

[8] C. Farkas and S. Jajodia. The inference problem: A survey. *SIGKDD Explorations*, 4(2):6–11, Jan. 2003.

[9] M. Kantarcıoğlu and C. Clifton. Assuring privacy when big brother is watching. In *The 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'2003)*, pages 88–93, San Diego, California, June 13 2003.

[10] M. Kantarcioglu and J. Vaidya. An architecture for privacy-preserving mining of client information. In C. Clifton and V. Estivill-Castro, editors, *IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining*, volume 14, pages 37–42, Maebashi City, Japan, Dec. 9 2002. Australian Computer Society.

[11] M. Kantarcıoğlu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge and Data Engineering*, to appear.

[12] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)*, Melbourne, Florida, Nov. 19-22 2003.

[13] X. Lin, C. Clifton, and M. Zhu. Privacy preserving clustering with distributed EM mixture modeling. *Knowledge and Information Systems*, to appear 2004.

[14] Y. Lindell and B. Pinkas. Privacy preserving data mining. *Journal of Cryptology*, 15(3):177–206, 2002.

[15] S. J. Rizvi and J. R. Haritsa. Maintaining data privacy in association rule mining. In *Proceedings of 28th International Conference on Very Large Data Bases*, pages 682–693, Hong Kong, Aug. 20-23 2002. VLDB.

[16] J. Vaidya and C. Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 206–215, Washington, DC, Aug. 24-27 2003.

[17] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Fransisco, Oct. 1999.