

Preserving Privacy in Data Mining via Importance Weighting

Charles Elkan

Department of Computer Science and Engineering
University of California, San Diego,
La Jolla, CA 92093-0404
elkan@cs.ucsd.edu

Abstract. This paper presents a fundamentally new approach to allowing learning algorithms to be applied to a dataset, while still keeping the records in the dataset confidential. Let D be the set of records to be kept private, and let E be a fixed set of records from a similar domain that is already public. The idea is to compute and publish a weight $w(x)$ for each record x in E that measures how representative it is of the records in D . Data mining on E using these importance weights is then approximately equivalent to data mining directly on D . The dataset D is used by its owner to compute the weights, but not revealed in any other way.

1 Introduction and Framework

Suppose that a hospital possesses data concerning patients, their diseases, their treatments, and their outcomes. The hospital faces a fundamental conflict. On the one hand, to protect the privacy of the patients, the hospital wants to keep the dataset secret. On the other hand, to allow science to progress, the hospital wants to make the dataset public. This conflict is the issue addressed by research on privacy-preserving data mining. How can a data owner simultaneously both publish a dataset and conceal it?

In this paper, we propose a new approach to resolving this fundamental tension between publishing and concealing data. The new approach is based on a mathematical technique called importance weighting that has proved to be valuable in several other areas of research [Hastings, 1970]. The essential idea is as follows. Let D be the set of records that the owner must keep confidential. Let E be a different set of records from a similar domain, and suppose that E is already public. The owner should compute and publish a weight $w(x)$ for each record x in E . Given x in E , its weight is large if x is similar to the records in D , while its weight is small otherwise. Data mining on E using the weights will then be approximately equivalent to data mining on D . The owner uses D privately to compute the weights, but never reveals D in any way.

Note that the proposed approach is non-interactive. The owner chooses E , and normally there will only be one dataset E and one set of weights. Data users, and adversaries, know the contents of E , but do not get to select E in any way.

A learning algorithm can be any statistical analysis method, including any supervised or unsupervised data mining method. The class of learning algorithms that we consider consists of methods that use the dataset D indirectly, by computing averages over it. Suppose that D contains records of type X , meaning that each record in D is an element of the space X . For example, if each record in D is a real-valued vector of length k , then $X = \mathbb{R}^k$. Let b be any real-valued function of X , i.e. $b : X \rightarrow \mathbb{R}$. The empirical average over D of b is

$$\bar{b} = \frac{1}{|D|} \sum_{x_i \in D} b(x_i).$$

Assume that the samples x_i are drawn independently and identically distributed (iid) from a fixed probability distribution f over the space X . The empirical average is then an estimate of the expectation of b over X given f .

The point of a learning algorithm, or of any statistical analysis, is to induce some property of the distribution f , rather than merely to obtain a measurement of the dataset D . The goal of an insider (that is, a data owner), is to allow outsiders to estimate properties of f without revealing D . In general, properties of f are expectations. An insider can estimate the expectation of the function $b(x)$ using the empirical distribution:

$$E[b(x)|x \sim f(x)] \doteq \frac{1}{|D|} \sum_{x_i \in D} b(x_i).$$

Different learning algorithms require estimates of different expectations relative to the distribution $f(x)$. The question is, how can the data owner allow outsiders to estimate all these expectations without revealing the specific x_i records in D to them?

Note that the framework adopted here for learning is the statistical queries model of [Kearns, 1998]. The operation of a learning algorithm is divided into two parts. One part is an algorithm that takes as given the availability of measurements of population averages. The other part is a procedure for obtaining estimates of these measurements. The importance sampling method proposed here is a privacy-preserving procedure for obtaining these estimates.

2 Preserving Privacy via Importance Weighting

Let f be any probability distribution over X and let b be any real-valued function over X . The definition of the expectation of b with respect to f is

$$E[b(x)|x \sim f(x)] = \int_{x \in X} b(x)f(x)dx.$$

Now let g be a different probability distribution over X , and consider the following equations:

$$\int b(x)f(x)dx = \int b(x)f(x)\frac{g(x)}{g(x)}dx = E[b(x)\frac{f(x)}{g(x)}|x \sim g(x)].$$

In words, the expectation of b with respect to the distribution f is equal to the expectation of $b(x)f(x)/g(x)$ with respect to the distribution g . This result is sometimes called the importance sampling identity [Hastings, 1970, Press, 2004] and has been used in recent research on sample-selection bias and covariate shift [Shimodaira, 2000, Smith and Elkan, 2007, Cortes et al., 2010]. The ratio $w(x) = f(x)/g(x)$ is called the importance of x .

For the importance sampling identity to be true, $g(x)$ must be positive for all x such that $f(x)$ is positive, i.e. the support of g must be a superset of the support of f . This condition is required to avoid division by zero. If $f(x) = g(x) = 0$ for any x , one can define $f(x)/g(x) = 0$ for that x .

In order to apply importance weighting to achieve privacy-preserving data mining, let the confidential dataset D be a random sample from the distribution f over X . A statistical query concerning D is an expectation

$$E[b(x)|x \sim f(x)] = \int b(x)f(x)dx$$

that would be estimated by the owner of D as

$$\frac{1}{|D|} \sum_{x_i \in D} b(x_i).$$

Now, let E be a random sample from a different distribution g over X that is known to an outsider. The fact that D and E are samples over the same space X means that, in database terminology, they have the same schema. The outsider can then estimate $E[b(x)|x \sim f(x)]$ as

$$\frac{1}{|E|} \sum_{x_i \in E} b(x_i)w(x_i).$$

In order to compute this estimate, for any b , the outsider does not need any access to D . The outsider does need to know the weights $w(x_i)$. However, these weights are the same for all b . The weights can be computed by the data owner based on its knowledge of D and E , and then published, once and for all.

The importance-weighting approach to privacy-preserving data mining is based on the assumption that an appropriate dataset E exists and is public. There are several possibilities for how E might exist. First, E might be a dataset that was revealed previously, perhaps inadvertently. At the time E was revealed, there was a breach of privacy, but now, one might as well use E for future research that does not breach privacy any further. Second, E might consist of information about individuals who have given consent for data concerning them to be published. For example, a hospital can gather consent for data release from some patients, and then compute weights that allow the consenting patients to be representative of all patients. Third, E might even consist of artificially generated synthetic data.

The method proposed above is different from two previous approaches that may appear similar at first sight. Some recent research has considered how to publish a version D' of D that preserves privacy yet is such that functions

in a certain class have similar values on D' and D [Blum et al., 2008]. In the proposal above, no new dataset D' is created or published; instead an existing dataset E is reused, and only scalar weights are published. Other research has considered how to answer subset-sum queries interactively in a differentially private way [Blum et al., 2005]. A subset-sum query asks the data owner to evaluate a function on a subset of D . The proposal above is non-interactive: the data owner does not answer queries at all. The data owner merely designates a dataset E and publishes one set of weights, once and for all.

3 How to Compute Importance Weights

For each data point x_i in E , its importance weight $w(x_i)$ is the ratio of the probability density of x_i according to two different distributions. Both distributions are over the space X , which in general has high dimensionality; the dimensionality is the length of the x_i vectors. Estimating high-dimensional densities is difficult at best, and often infeasible [Scott, 1992]. Fortunately, we can estimate the ratio $w(x)$ without estimating $f(x)$ and $g(x)$, as follows.

Let F be the combined dataset $D \cup E$ where samples from D are extended with the label $s = 1$ and samples from E are extended with the label $s = 0$. Suppose that we use F to learn a model of $p(s = 1|x)$. Then,

$$p(s = 1|x) = \frac{p(x|s = 1)p(s = 1)}{p(x)}$$

by Bayes' rule. Therefore,

$$p(s = 1|x) = \frac{f(x)p(s = 1)}{f(x)p(s = 1) + g(x)p(s = 0)} = \frac{1}{1 + \frac{g(x)p(s=0)}{f(x)p(s=1)}}.$$

As above, let $w(x) = f(x)/g(x)$ and let $r = p(s = 0)/p(s = 1)$. We can derive

$$w(x) = \frac{r}{1/p(s = 1|x) - 1}. \tag{1}$$

The equation above lets us write each weight $w(x)$ as a deterministic transformation of $p(s = 1|x)$. The equation is correct as a statement of probability theory. Its practical usefulness depends on being able to estimate the probability $p(s = 1|x)$ for each x in the dataset E . Fortunately, in general we can learn to estimate these probabilities accurately. To do so, we apply a supervised learning method that yields well-calibrated conditional probability predictions to the union of the D and E datasets. The simplest method with this property is logistic regression, but many other appropriate methods exist also [Zadrozny and Elkan, 2001].

To clarify, only the data owner knows both datasets D and E . Using these, the owner trains the model $p(s = 1|x)$, and applies this model to each example x_i in E . The owner then computes and publishes $w(x_i)$ for each of these examples,

non-interactively, using Equation (1). Outsiders know the dataset E and the published numerical weights. They are not given access to anything else.

Each dataset D and E is treated as a random sample from a corresponding population. The two populations may be more similar or less similar. If the populations happen to be identical, then it will be the case that $w(x_i)$ equals the same constant for all x_i in E . In separate research, we have developed a variant of regularized logistic regression that allows for lower and upper bounds L and U for predicted probabilities:

$$0 < L \leq \min_x p(s = 1|x) < \max_x p(s = 1|x) \leq U < 1.$$

If D and E come from indistinguishable populations, then the new variant of logistic regression will in principle learn that $L = U$.

The data owner does not need to publish the model used to compute weights (only the numerical weights themselves). However, if an adversary happened to know this model, it could compute $p(s = 1|x)$ for any data record x . If this value is high, then x is more typical of records in D than of records in E . But that simply means x is representative of the population from which D is drawn. The adversary cannot conclude that x actually appears in D .

The approach just explained to estimate the ratio of two probability densities, without needing to estimate the two densities individually, is something of a folk result. Variations of it have been discovered and used independently several times [Zadrozny, 2004, Smith and Elkan, 2007, Tsuboi et al., 2009]. To be useful in practice, the approach requires careful regularization.

4 Research Questions

The importance-weighting approach has two major drawbacks. The first obvious issue is that an appropriate dataset E must already exist and be public. A less obvious issue is that the dataset E may be too good, that is too similar to D . Suppose for the sake of argument that E is a superset of D . Then the ideal weights will be $w(x) > 0$ for each x in D , and $w(x) = 0$ for each x that is in E but not in D . In general, weights will be high for examples in E that are representative of D . An adversary will know this, but cannot conclude that a record in E with a high weight appears “as is” in D .

The following research questions need answers. They are related to each other, so the order in which they are stated here is somewhat arbitrary and does not reflect their relative importance. The first questions concern importance weighting in general, while later ones are specifically related to privacy.

1. Is logistic regression the best supervised learning method for the data owner to use to estimate $p(s = 1|x)$, or does a better alternative exist? What variety of smoothing or regularization is best?
2. How should D and E be divided into training, validation, and test sets for the purpose of computing weights?

- For some x_i in E , the estimated value of $w(x_i)$ will be large. These x_i will have disproportionate influence in estimates of $E[b(x)]$ for all functions b . How can large values of $w(x_i)$ be avoided, while still maintaining correctness?

The variant of logistic regression with lower and upper bounds mentioned above should help answer the last question above, because Equation (1) implies that bounds on $p(s = 1|x)$ correspond to bounds on the importance weights.

The following research questions ask what theoretical guarantees concerning information disclosure can be proved for the importance-weighting approach. We conjecture that under some conditions, the approach can be proved to satisfy the definition of differential privacy [Dwork, 2008]. Intuitively, the more general a statistical query is, the higher the accuracy with which it can be answered using E and the published weights. Queries that in fact ask about a single potential record in D will only be answered with very low precision.

The specific research questions are the following:

- Characterize the uncertainty in estimates

$$E[b(x)|x \sim f(x)] \hat{=} \frac{1}{|E|} \sum_{x_i \in E} b(x_i)w(x_i)$$

by computing confidence intervals. What are the quantities on which these intervals depend?

- Intuitively, knowing the expectation of an indicator function such as

$$b(x) = I(\text{lastname}(x) = \text{Obama})$$

destroys privacy, whereas a function such as

$$b(x) = I(\text{age}(x) \geq 40)$$

is irrelevant to privacy. Provide a formal definition of privacy-destroying and privacy-irrelevant functions.

- Show that if $b(x)$ is privacy-destroying then the uncertainty in its estimated expectation is high, while if $b(x)$ is privacy-irrelevant then the uncertainty is small.
- Show that publishing the dataset E with the weights $w(x_i)$ for x_i in E satisfies the definition of differential privacy for D .
- Under interactive models of differential privacy, the number of queries allowed must be sublinear in the size of D . When learning importance weights, is this “privacy budget” relevant? If so, how can one avoid exceeding it?

There is a simple intuitive argument why differential privacy is guaranteed for the importance-weighting approach. The only information that is computed and revealed from the confidential dataset D is a single logistic regression function. And regularized logistic regression can be trained while respecting differential privacy [Chaudhuri and Monteleoni, 2008, Chaudhuri and Sarwate, 2009]. Intuitively, if the published weights $w(x)$ are approximately unchanged whether or not any particular record is included in D or excluded from D , then the importance-weighting approach satisfies differential privacy. And because the weights are

based on logistic regression, they do satisfy differential privacy. Making this argument precise is a priority for continued research.

Acknowledgments. The author is grateful to anonymous referees for comments that helped in clarifying the ideas of this position paper.

References

- [Blum et al., 2005] Blum, A., Dwork, C., McSherry, F., Nissim, K.: Practical privacy: the SuLQ framework. In: Proceedings of the 24th ACM Symposium on Principles of Database Systems, pp. 128–138. ACM Press, New York (2005)
- [Blum et al., 2008] Blum, A., Ligett, K., Roth, A.: A learning theory approach to non-interactive database privacy. In: Proceedings of the 40th Annual ACM Symposium on Theory of Computing, pp. 609–618. ACM Press, New York (2008)
- [Chaudhuri and Monteleoni, 2008] Chaudhuri, K., Monteleoni, C.: Privacy-preserving logistic regression. In: Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS), pp. 289–296 (2008)
- [Chaudhuri and Sarwate, 2009] Chaudhuri, K., Sarwate, A.D.: Privacy constraints in regularized convex optimization. Arxiv preprint arXiv:0907.1413 (2009)
- [Cortes et al., 2010] Cortes, C., Mohri, M., Riley, M., Rostamizadeh, A.: Sample selection bias correction theory. In: Algorithmic Learning Theory, pp. 38–53. Springer, Heidelberg (2010)
- [Dwork, 2008] Dwork, C.: Differential privacy: A survey of results. In: Agrawal, M., Du, D.-Z., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008)
- [Hastings, 1970] Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1), 97–109 (1970)
- [Kearns, 1998] Kearns, M.: Efficient noise-tolerant learning from statistical queries. *Journal of the ACM* 45(6), 983–1006 (1998)
- [Press, 2004] Press, W.H.: How to use Markov chain Monte Carlo to do difficult integrals (including those for normalizing constants) (2004), Draft working paper available at <http://www.nr.com/whp/workingpapers.html>
- [Scott, 1992] Scott, D.W.: Multivariate density estimation: Theory, practice, and visualization. Wiley-Interscience, Hoboken (1992)
- [Shimodaira, 2000] Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* 90(2), 227–244 (2000)
- [Smith and Elkan, 2007] Smith, A., Elkan, C.: Making generative classifiers robust to selection bias. In: Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 657–666. ACM Press, New York (2007)
- [Tsuboi et al., 2009] Tsuboi, Y., Kashima, H., Bickel, S., Sugiyama, M.: Direct Density Ratio Estimation for Large-scale Covariate Shift Adaptation. *Journal of Information Processing* 17, 138–155 (2009)
- [Zadrozny, 2004] Zadrozny, B.: Learning and evaluating classifiers under sample selection bias. In: Proceedings of the 21st International Conference on Machine Learning, pp. 903–910. ACM Press, New York (2004)
- [Zadrozny and Elkan, 2001] Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In: Proceedings of the 18th International Conference on Machine Learning, pp. 609–616. Morgan Kaufmann, San Francisco (2001)