Utility of Privacy Preservation for Health Data Publishing

Lengdong Wu, Hua He, Osmar R. Zaïane
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada
{lengdong, hhe, zaiane}@ualberta.ca

Abstract

In the medical field, we are amassing phenomenal amounts of data. Because of understandable ethical and legal responsibility to maintain the privacy, many techniques of anonymization have been proposed to provide means of data publishing without jeopardizing privacy. The strictness of the techniques is putting in question the utility of the health data after severe anonymization.

1. Introduction

Health research is central to the advancement of health care, which imperatively requires access to health data. In recent years, several significant privacy preserving techniques have been proposed to protect individual's privacy for health data publishing. However, although such techniques can prevent privacy leakage, they also significantly hinder the data utility for research purposes. In this paper we investigate several rigorous anonymization techniques with novel criterions to evaluate the data utility.

2. Privacy Preservation Technique

A typical health data table includes the basic personal information as well as their sensitive information. All these attributes can be categorized into three classes [3]: identifier, quasi-identifier (QI) and sensitive attributes. According to the *HIPPA*[1] regulation, the removal of all identifier is required. Furthermore, the *k*-Anonymity [1, 3] technique is designed to avoid re-linking attacks by generalizing the QI attribute values into an equivalence class [3], thus the re-linking attack cannot distinguish a certain individual from other records in the equivalence class. However, *k*-anonymity is vulnerable to homogeneity attacks [2] due to auxiliary clew from sensitive attributes.

To defend the defect of *k*-anonymity, ℓ-diversity [2] requires that the sensitive attribute values in each equivalence class should be as diverse as possible, and each class should have at least ℓ well-represented sensitive attribute values. This requirement on sensitive attributes adds an extra protection layer over *k*-anonymity. When a health data table satisfies the ℓ-diversity principle, the adversary who can breach the *k*-anonymity, still needs to exclude the (ℓ-1) possible sensitive values. However, ℓ-diversity simply makes sensitive attribute values numerically diverse, it is still vulnerable to similarity attack [4] or skewness attack [5], which can utilize semantic similarity information leakage in sensitive attributes. The breach can be serious when the number of sensitive attribute categories is small. The *t*-closeness [5] technique is later proposed to solve this issue, and it requires the sensitive values to be semantically diverse, so that the distribution of sensitive values in each equivalence class is close to the overall distribution of the table.

3. Experiments Evaluation

The three important anonymization techniques are effective in protecting data privacy. However, there is a risk that they lower the utility of the data in the context of health research. Therefore the balance between the data utility for scientific research and the privacy preservation for health data is highly importance. We evaluate the utility loss as follows:

(1) Utilize an effective machine learning method, Support Vector Machine (SVM) [6], to examine the utility value through the measure of accuracy after anonymization. The lower the accuracy, the less utility value can be preserved.

(2) Evaluate the similarity between anonymized and original table based on Earth Mover's Distance (EMD) [7].The greater the distance is, the more utility is lost.

[1] Notice of Addresses for Submission of HIPAA Health Information Privacy Complaints Federal Register, Vol. 68, No. 54, March 20, 2003

3.1. Datasets and Experimental Setup

We use two census-based datasets, the Adult dataset and the IPUMS dataset, publicly available from the UCI Machine Learning repository [2]. We choose attributes *age, education, gender, race, marriage* and *hometown* as QI attributes, and use *salary* as the sensitive attribute. Our experiments apply the common settings that are not too strict to make the anonymized data completely unusable.

The datasets are divided into the training and test sets randomly in three fold cross validation sets. We apply SVM on both original and anonymized data. By comparing the classification accuracy, we can evaluate to what degree the anonymized data could lose utility.

To compare the similarity of tables with EMD, we first map data into an underlying ordered space, then we rank the data and finally calculate the rank distance between two columns. This approach can overcome the underlying data representation limitations of the EMD algorithm.

3.2 Experiment Results

Table 1 and Table 2 present the comparisons of the accuracies on the Adult dataset. The result shows that there is a significant difference in terms of accuracy between evaluations on the original data and anonymized data. Significant drops, 7% in accuracy and 18% in F-measure, can be observed and the reason is due to the inadvertent obfuscation of pertinent information necessary for building the classification model.

Table 1: Experiment Results with Adult Dataset

	PRECISION	*RECALL*	*F-MEASURE*	*ACCURACY*
Original	82.5%	82.2%	82.3%	87.1%
k-Anonymity	65.0%	75.3%	64.8%	70.3%
ℓ-diversity	56.7%	75.3%	64.7%	70.2%
***t*-closeness**	54.4%	75.4%	64.8%	70.2%

Table 2: Experiment Results with IPUMS Data

	PRECISION	*RECALL*	*F-MEASURE*	*ACCURACY*
Original	87.2%	87.9%	86.3%	87.9%
k-Anonymity	73.1%	83.5%	70.8%	75.3%
ℓ-diversity	66.8%	82.9%	62.7%	68.2%
***t*-closeness**	58.4%	83.2%	60.6%	66.2%

Table 2 also shows that there is a noticeable drop in terms of accuracy for the classifier when using data from ℓ-diversity and *t*-closeness as compared to *k*-anonymity, indicating an additional loss beyond *k*-anonymity. Moreover, ℓ-diversity provides better precision while *t*-closeness results in better accuracy, meaning the supposition that being stricter than ℓ-diversity, *t*-closeness is not necessarily responsible for more utility loss. Indeed, ℓ-diversity and *t*-closeness are built on distinct adjustments of *k*-anonymity, and thus they both inherit directly from *k*-anonymity.

In the experiment with EMD, it is observed that the distance values of each column in the three anonymized tables are almost on the same level for the Adult dataset, suggesting although ℓ-diversity and *t*-closeness are stricter than *k*-anonymity, the distribution are similar. This result is consistent with the one obtained using SVM. In results on IPUMS dataset, the difference for certain columns is modest, indicating that the data skew caused by a more strict principle is also acceptable. However, for some columns the distance value growth is large, thus the data skew in these columns is serious, which indicates the low data utility. If these columns are important for research in certain health care areas, a looser anonymization strategy should be considered.

4. Conclusion

In this paper we examine the issue of health data utility after three anonymization techniques. We utilize two practical measurement methods and focus on the necessity of finding a balance between the privacy protection and data utility for data publishing. By evaluating the utility loss of three important privacy preservation techniques with SVM and EMD, we show that today's privacy preservation techniques can significantly jeopardize the data utility due to the highly strict protection principles they impose.

References

[1] Bayardo, etc. 2005. Data privacy through optimal *k*-anonymization.In *ICDE'05*, pp.217-228.

[2] Machanavajjhala, A.,etc. ℓ-diversity: privacy beyond *k*-anonymity. In *Transactions on Knowledge Discovery from Data,* 2007

[3] Sweeney, L. 2002. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems.* 10(6):571–588.

[4] Truta, T. M. and Vinay, B. 2006. Privacy protection: p-sensitive k-anonymity property. In *PDM'06.*

[5] Li, N., etc. 2007. t-closeness: privacy beyond k-anonymity and l-diversity. In *ICDE'07*. pp. 106-115.

[6] Drucker, etc. Support Vector Regression Machines. In *NIPS'07*. pp. 155–161, MIT Press.

[7] Rubner, Y., etc. 2000. The earth mover's distance as a metric for image retrieval. Int. J. Computer Vision, 40(2):99–121.

[2] http://archive.ics.uci.edu/ml/machine-learning-databases/