Privacy Preservation in Crowdsourced Health Research

Hiromi Arai
Information Technology Center
The University of Tokyo
2-11-16 Yayoi, Bunkyo-ku
Tokyo, Japan
arai@dl.itc.u-tokyo.ac.jp

Abstract

Crowdsourced health research is a growing field toward achieve personalized healthcare. Health research often confronts with the small N problem; difficulty for statistical inference due to small sample size. Crowdsourced science leverages the power of the mass scale and provides benefits. However, privacy concerns often prevent data sharing. Privacy preservation data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. This talk gives overview of recent PPDM techniques. We also discuss how to utilize PPDM for health research.

Introduction

Crowdsourced health research is emerging in science. Personalized service organizations collect and use large amount of personal data. For example, openSNP for personal genomics, Facebook for social studies and purchase history for recommendation systems. These crowdsourced personal data collections are a significant resource for cohort studies.

However, there are still some problems that make data holders hesitate to share their data. One big problem is the privacy of the participants.

We focus on privacy preservation data mining (PPDM) techniques (Agrawal and Srikant 2000; Lindell and Pinkas 2000) to solve privacy problems in data sharing. PPDM is a novel research direction in data mining and statistical databases. Many techniques for data mining while preserving privacy have been proposed.

In this study, we introduce several data sharing models for crowdsourced health research. Then, we review PPDM techniques for these models and discuss their suitablity and further works.

Data Sharing Models for Crowdsourced Health Research

We consider several data sharing models for crowdsourced health research in a privacy-preserving way. Let us consider database owners and users as parties. Each database owner collects and use a personal health data such as lifelogs, medical records, etc. Each user wants to use data or data-based services for healthcare research or applications. To encourage participation in databases, the database owners want to operate their data in a privacy-preserving way.

Let $I = \{1, \ldots, n_I\}$ be a set of data owners. Let D_i be the database that the data owner i holds. Assume that D_i contains personal sensitive data and the data owner i wants to preserve is. Let $U = \{1, \ldots, n_U\}$ be a set of users.

The first data sharing model is for data publishing. In this model, the owner i publishes his data D_i or function values of the data $f(D_i)$ to public for data utilization. This model assumes the publishing of the research results in journals or websites, and publishing of cencus statistics. This model includes not only publishing datatable or statistical values, but also the results of data analysis such as classifiers, regression models, etc.

Second is a database-querying model. Assume the user u requests the query function $f_u(D_i, q_u)$ to the data owner i. We assume similarity searches, recommendation, diagnosis such as healthcare advice or genetic diagnosis, etc. Here, we consider that the query q_u contains user u's private information. For example, similar patient recommendation uses a user's medical records. The genetic diagnosis uses a user's genetic information.

Third is a data integration model. Assume the data owners I wants to integrate their data to enhance the knowledge quality they can obtain from their data. For example, a data mining result $f(\cup_i D_i)$ is expected to be better than $f(D_i)$ when each D_i holds a small number of samples. We assume database integration repositories such as The Cancer Genome Atlas (TCGA), iDash (Ohno-Machado et al. 2012).

Privacy Preserving Data Mining

Here we discuss whether privacy preserving data mining techniques can be applied to the data sharing models introduced in the previous section.

For the data publishing model, output privacy techniques can be applied. These techniques sanitize data D_i or functional value of it $f(D_i)$ so as to achieve their privacy standards. Anonymization is one approach for privacy preserving database publishing by de-identifying individuals in the dataset. k-anonymity (Sweeney 2002) is one of the major anonymization methods. The quasi-identifiers are a pieces of

information that are not unique identifiers, but can be identifying information when combined. k-anonymization generalizes or erase the quasi-identifiers so that any individual cannot be distinguished from at least $k-1$ individuals in the dataset. Data generalization or surppression are used for this technique.

Randomized approach is another way to preserve privacy in data disclosure. Differential privacy (Dwork et al. 2006) is often used for data publishing. Differential privacy is a condition of the data publishing mechanism that gives a privacy guarantee for the input private data. To satisfy differential privacy, generally randomized mechanism is used (McSherry and Talwar 2007).

Some people consider that the statistical data derived from personal data does not breach privacy when the sample size is large enough. Some reports show that there are privacy risks even in statistical data. For example, a part of personal genome information is possibly obtained from GWAS reports (Homer et al. 2008; Wang et al. 2009; Im et al. 2012). Note that privacy risk detection is still an immature research area. Systematic understanding about privacy risk is an open problem.

For the database-querying model, multi partiy computation (MPC) techniques (Goldreich 2004) can be used to preserve a user's privacy. In MPC, all participants jointly execute cryptographic protocol to obtain some statistical data while their inputs are kept secret. In other words, I can obtain $f(D_1, \ldots, D_I)$ without revealing $D_1, \ldots, D_I$. However, MPC requires more computational cost and time compared with computation with plain texts. For this reason it may be unrealistic in the case of large scale data.

Query auditing techniques (Nabar et al. 2008) can be used to check database-privacy preservation. This is the process of examining past data publishing to detect disclosure of private data. Due to computational complexity of this problem, query auditing has been studied only against simple aggregate queries.

For the data integration model, MPC is one solution to obtain statistical data of integrated data while preserving a databases privacy. However, there is a problem that MPC becomes difficult for large scale data.

Emsemble or data aggregation is another solution for data integration. If the data-mining for integrated data is desired and statistical data such as classifiers are available, , emsemble learning (Rokach 2010) is one solution. For the case that statistical data should be private, a differentially-private emsemble learning framework has been proposed recently (Sarwate et al. 2014).

Discussion

We propose some private data sharing models for croudsourced health research and review several privacy preserving techniques for these models. The privacy preserving techniques are helpful to preserve privacy, but these require computational cost or diminish the utility of data by randomization, generalization or surppression. Overall, privacy protection in data utilization is still under discussion. There have been several privacy standards and privacy preserving techniques proposed, but we have to discuss and examine which one is well suitable to healthcare research.

References

Agrawal, R., and Srikant, R. 2000. Privacy-preserving data mining. *ACM Sigmod Record* 29(2):439–450.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography* 265–284.

Goldreich, O. 2004. *Foundations of cryptography: Basic applications*. Cambridge University Press.

Homer, N.; Szelinger, S.; Redman, M.; Duggan, D.; Tembe, W.; Muehling, J.; Pearson, J. V.; Stephan, D. A.; Nelson, S. F.; and Craig, D. W. 2008. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics* 4(8):e1000167.

Im, H. K.; Gamazon, E. R.; Nicolae, D. L.; and Cox, N. J. 2012. On sharing quantitative trait gwas results in an era of multiple-omics data and the limits of genomic privacy. *The American Journal of Human Genetics* 90(4):591–598.

Lindell, Y., and Pinkas, B. 2000. Privacy preserving data mining. In *Advances in CryptologyCRYPTO 2000*, 36–54. Springer.

McSherry, F., and Talwar, K. 2007. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. 48th Annual IEEE Symposium on*, 94–103. IEEE.

Nabar, S.; Kenthapadi, K.; Mishra, N.; and Motwani, R. 2008. A survey of query auditing techniques for data privacy. *Privacy-Preserving Data Mining* 415–431.

Ohno-Machado, L.; Bafna, V.; Boxwala, A. A.; Chapman, B. E.; Chapman, W. W.; Chaudhuri, K.; Day, M. E.; Farcas, C.; Heintzman, N. D.; Jiang, X.; et al. 2012. idash: integrating data for analysis, anonymization, and sharing. *Journal of the American Medical Informatics Association: JAMIA* 19(2):196–201.

Rokach, L. 2010. Ensemble-based classifiers. *Artificial Intelligence Review* 33(1-2):1–39.

Sarwate, A. D.; Plis, S. M.; Turner, J. A.; Arbabshirani, M. R.; and Calhoun, V. D. 2014. Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation. *Frontiers in neuroinformatics* 8.

Sweeney, L. 2002. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems* 10(5):571–588.

Wang, R.; Li, Y. F.; Wang, X.; Tang, H.; and Zhou, X. 2009. Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the 16th ACM conference on Computer and communications security*, 534–544.