Anonymizing medical data: developing a generalization algorithm that optimizes p-privacy

Shannon Kroes Matthijs van Leeuwen Mart Janssen

June 24, 2019

Context

Researchers and medical institutions are obliged to protect the privacy of individuals from whom data has been collected. As a result, methods have been developed that aim to *anonymize* data such that they can be shared without the risk of breaching privacy. One popular method is *generalization*, which replaces each value in the data by a generalized, larger range of values. For example, instead of reporting that an individual is 50 years old, the generalized data might state that this individual is between 45 and 55 years old. As this tends to create more overlapping values among individuals, this will make it more difficult to identify individuals in the data, and hence to extract sensitive information.

Problem

In order to generalize a data set, it needs to be decided which values should be grouped together, e.g., how intervals can be best segmented. Several algorithms have been developed to this end; both bottom-up and top-down approaches have been considered [4]. These algorithms need a measure of privacy, so that privacy gain can be balanced to the loss of information. The most commonly used measure of privacy is *k-anonymity* [3], which quantifies privacy by computing the minimum number of individuals k that have the exact same values for all variables in the data set. This concept has frequently been criticized though, as it does not ensure that sensitive information cannot be extracted [2]. For example, consider a data set on a group of patients containing the information that all 50-year-old men have been diagnosed with cancer. Knowing an individual's age and gender will be sufficient to infer the diagnosis based on this data set. Several measures have been developed since to address these shortcomings, such as l-diversity [2] and t-closeness [1]. These methods are designed to evaluate entire (generalized) *data sets*, and are not suitable to quantify the privacy of *individuals*. As a consequence, these measures of privacy are difficult to incorporate into generalization algorithms, which may be why there are no generalization algorithms available that optimize any of these entities.

Aim of the project

During the course of this project a generalization algorithm will be developed based on p-privacy. Using this recently developed measure, privacy of individuals can be easily evaluated on an individual level, and thus optimization will be more straightforward compared to the aforementioned methods. We have previously shown how the measure can be used to compare different generalizations and the extension to an automated process is therefore very appealing, especially considering the fact that no such algorithm is currently available in the literature.

Realisation

A generalization algorithm will be developed and implemented. This algorithm will output a generalized version of a (medical) data set that optimizes p-privacy, while balancing the information loss resulting from generalization. The algorithm will be applied to different publicly available data sets to test its performance, and a report will be written on the results.

This project is a collaboration between the Dutch blood bank *Sanquin*, LUMC, and LIACS. If you have any questions please feel free to send an email to S.kroes@sanquin.nl.

References

[1] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[2] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering*, page 24. IEEE, 2006.

[3] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[4] Ke Wang, S Yu Philip, and Sourav Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *Fourth IEEE International Conference on Data Mining*, pages 249–256. IEEE, 2004.