Data61 PhD opportunity: Privacy-Preserving Active Learning for Data Linkage

A PhD opportunity with a potential full scholarship is available for those interested in research involving privacy-enhancing technologies and data linkage.

Project description

Data linkage is the process of identifying matching data points (or records) from multiple different databases, which is required in a variety of applications, for example, cybersecurity and health analytics applications. This process is often prone to data errors and variations and is significantly dependent on domain expertise for manual labelling of records (also known as clerical review) for accurate linkage. Moreover, machine learning algorithms for data linkage require training data that is not generally available in practical applications and thus require manual labelling. Manual labelling or classification is challenging in real-world applications and has several constraints to be addressed: Cost-constrained: Manual labelling often incurs a high cost to label a large number of samples. Privacy-constrained: The growing privacy and confidentiality concerns preclude the sharing of data with human experts for labelling that can reveal personal identifiable information about individuals represented by the data. Fairness-constrained: Supervised machine learning-based data linkage algorithms can learn to ignore poor performance on a small (minority) group if it can exploit knowledge about the majority population from the training data, potentially leading to unfair outcomes.

In this project, we aim to study privacy-preserving active learning subject to the three constraints. Privacy-preserving active learning techniques should not only reduce the cost of labelling by selecting informative data samples to be labelled but also need to reduce the privacy risk of re-identification by sequentially and interactively revealing part of the information from data samples while meeting fairness objectives. To the best of our knowledge, no work has so far addressed the active learning problem with all three constraints, which we believe is important for practical data linkage applications. The main objectives and outcomes of the project (which we aim to publish in high-impact conferences/journals on data privacy) are:

- Study and compare existing cost-effective and interactive algorithms for selecting data samples for manual labelling.
- Design and develop novel active learning algorithms using privacy-enhancing technologies such as Differential privacy mechanisms combined with masking functions.
- Design fairness-aware algorithms for privacy-preserving active learning.
- Compare and evaluate (both theoretically and empirically) the trade-off between privacy, the accuracy of linkage, fairness of linkage, and cost provided by the algorithms.

Skills/Capability required for the project

- Bachelor's degree in Computer Science or relevant field.
- Programming experience in Python and deep learning tools such as TensorFlow or PyTorch.
- Knowledge in privacy-enhancing technologies and machine learning is preferable.

For more details and how to apply, please visit the following link
https://jobs.csiro.au/job/Various-Data61-PhD-Scholarships/796808000/?locale=en_GB