



Privacy-Preserving Model Selection

Mission

A staggering amount of data, much of it sensitive, is distributed among a variety of data owners, collectors, and aggregators. Modern methods of data analysis provide the power to extract useful knowledge from this data. However, privacy concerns may prevent different parties from sharing their data with others. A major challenge is how to realize the utility of this distributed data while also protecting data privacy. Privacy-preserving data analysis provides data analysis algorithms in which the goal is to compute or approximate the output of one or more particular algorithms applied to the joint data, without revealing anything else that is sensitive about the data.

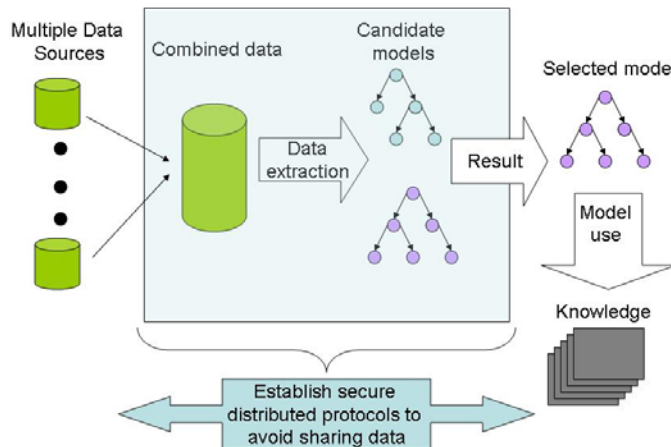
However, the data analyst's task rarely starts and ends with running a particular data analysis algorithm. In particular, a data analyst seeking to model some data will often run a number of different kinds of data analysis algorithms and then perform some kind of model selection to determine which of the resulting models to use. Our project studies the notion of privacy-preserving model selection. Our initial work assumes a very specific kind of vertical partitioning in which one party holds all the data except the class labels, and a second party holds all the class labels. In this setting, we are able to show how to perform model selection using cross validation in a privacy-preserving manner, without revealing the parties' data to each other. Our solution enables the parties to privately determine the best among a number of candidate models for the data, thereby extending the privacy of the data from the initial model computation through to the model selection step.

Outreach

Project participants organized two workshops related to securing private information: "Mathematical & Computational Methods for Information Security" held at Texas Southern University on December 7, 2007 and Data Privacy at Rutgers University on February 4 - 7, 2008.

Collaborator(s):

- Rebecca Wright
- Zhiqiang Yang
- Sheng Zhong



Funded by: US Department of Homeland Security and the National Science Foundation

Combining data from multiple distributed sources offers richer potential for analysis and observation but may compromise privacy. We are developing protocols that will enable us to extract information from distributed data, while protecting privacy by avoiding the need to share data.

