Secure Training of Extra Trees Classifiers over Continuous Data

Chaitali Choudhary[*], **Martine De Cock**[*†], **Rafael Dowsley**[‡], **Anderson Nascimento**[*], **Davis Railsback**[*]
[*]School of Engineering and Technology, UW Tacoma
{cc201,mdecock,andclay,drail}@uw.edu
[†] Dept. of Applied Mathematics, Computer Science and Statistics, Ghent University
martine.decock@ugent.be
[‡] Department of Computer Science, Bar-Ilan University
rafael@dowsley.net

Abstract

Existing secure Multi-Party Computational (MPC) protocols for the training of decision trees over distributed data are only capable of handling categorical attributes. This is an enormous restriction on the practicality of their use, as attributes in data sets used in practice are often numerical. The standard "in the clear" algorithm to train decision trees on real-valued data sets requires sorting training examples for each feature at each node to find an optimal cut point – a prohibitively expensive operation in MPC. In this paper we propose an alternative method for securely training tree-based models on data with continuous attributes. Namely, the secure training of extremely randomized trees ("Extra Trees"). In addition to randomizing feature choices – as is done in random forests training – feature value thresholds are chosen randomly as well, thereby removing the need for sorting. We implement our solution in the semi-honest majority setting with additive secret sharing based on pre-distributed correlated randomness. To the best of our knowledge, our solution is the very first that privately trains a decision tree based model with continuous attributes where the overall complexity depends only linearly on the size of the entire training data set – contrary to sorting solutions.