A Differentially Private Random Decision Forest using Reliable Signal-to-Noise Ratios

Sam Fletcher & Md Zahidul Islam
safletcher@csu.edu.au & zislam@csu.edu.au
School of Computing and Mathematics, Charles Sturt University, Australia

Abstract

When dealing with personal data, it is important for data miners to have algorithms available for discovering trends and patterns in the data without exposing people's private information. Differential privacy offers an enforceable definition of privacy that can provide each individual in a dataset a guarantee that their personal information is no more at risk than it would be if their data was not in the dataset at all. By using mechanisms that achieve differential privacy, we propose a decision forest algorithm that uses the theory of Signal-to-Noise Ratios to automatically tune the algorithm's parameters, and to make sure that any differentially private noise added to the results does not outweigh the true results. Our experiments demonstrate that our differentially private algorithm can achieve high prediction accuracy.

Introduction

Privacy is becoming an increasingly important concern when performing data analysis, and ϵ-differentially privacy address that concern by offering a definition of privacy that hides each individual's presence in a dataset by a user-defined ϵ amount whenever the data is queried. A data analyst will only be given a β budget worth of privacy loss that they are allowed to induce on the data, and so it's important to spend the budget wisely. Harnessing several strategies, we offer a way to build a decision forest (i.e a collection of decision trees, see Figure 1) with a limited privacy budget.

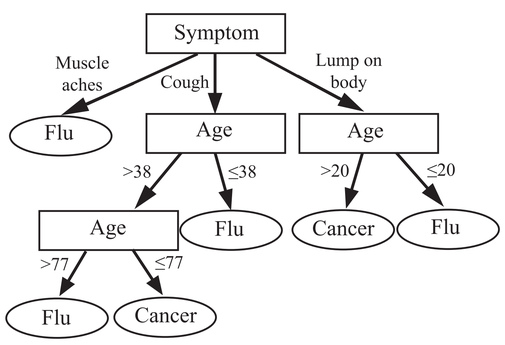

Figure 1: Decision trees are made up of nodes, where each node splits the data into subsets. A tree starts from a top "root" node, and ends in "leaf" nodes. Future records are predicted to have the most common label of the records that matched the same leaf.

Signal-to-Noise Ratio

SNR is a comparison between some sort of measurement (signal) and the background noise accompanying that measurement. In our case, the noise is intentionally added with the Laplace Mechanism. The SNR of a measurement can be expressed as

$$\text{SNR} = \frac{\mu}{\sigma} \tag{1}$$

where μ is the mean of the signal mean and σ is the std dev of the noise.

Signal Averaging

Signal averaging quantifies the intuition that if noise can increase or decrease a signal with equal probability, then summing multiple signals together will result in a total that is less noisy:

$$\text{SNR} = \frac{\mu}{\sigma} = \frac{\Sigma_x^X \mu_x}{\sqrt{|X|\sigma^2}} \tag{2}$$

where X is the set of signals, and $|X|$ is the size of that set. When rewritten in terms of our scenario, SNR becomes:

$$\text{SNR} = \frac{\epsilon \Sigma_L^{\text{Leafs}} \left(\Sigma_c^C L_c\right)}{|C|\sqrt{2 \times |\text{Leafs}|}} \tag{3}$$

where *Leafs* is the set of all leaf nodes that are descendants of the current node, and L_c is the count for each label in leaf L.

Our Algorithm

1. Based off the size of the privacy budget β, the size of the dataset D and the domain sizes of the attributes A, automatically tune the following parameters:
   - τ, the number of trees in the decision forest, which in turn dictates the ϵ spent per tree.
   - θ, the minimum size for nodes in the trees.
2. Build a randomized decision forest using τ and θ (i.e. the nodes are randomly created).
3. Query the dataset using the Laplace Mechanism to learn the labels in each leaf.
4. Prune away nodes with SNR < 1, using *signal averaging* for non-leaf nodes.
5. Find the node with the largest majority of a label (i.e. the most confident node) in each path from the root node to a leaf node, in each tree.
6. Predict the label of future records by voting on the most confident predictions made by each tree.

θ is defined as $\theta = \frac{\sqrt{2}|C|}{\epsilon}$, where $|C|$ is the number of labels and $\epsilon = \frac{\beta}{\tau}$. τ is defined as the largest number, up to the number of attributes in the data, that satisfies $\theta < \frac{|D|}{\delta^2}$, where where δ is the average domain size of the attributes. These parameters are derived from the SNR theory.
By removing nodes with too high a ratio of noise, and then making predictions based off the most confident nodes out of all the trees, we can guarantee that any predictions made about future records are made using label counts that not only have high confidence, but also outweigh any noise that might have been added to them.

The Laplace Mechanism

The Laplace Mechanism can be used to inject a small amount of randomness into queries in order to achieve ϵ-differential privacy. A query Q satisfies ϵ-differential privacy if it outputs $y + Lap(1/\epsilon)$, where $y \in Y : Q(D) \rightarrow Y$ and $Lap(x)$ is an i.i.d. random variable drawn from the Laplace distribution with mean 0 and scale x (i.e. variance $2x^2$).

Results

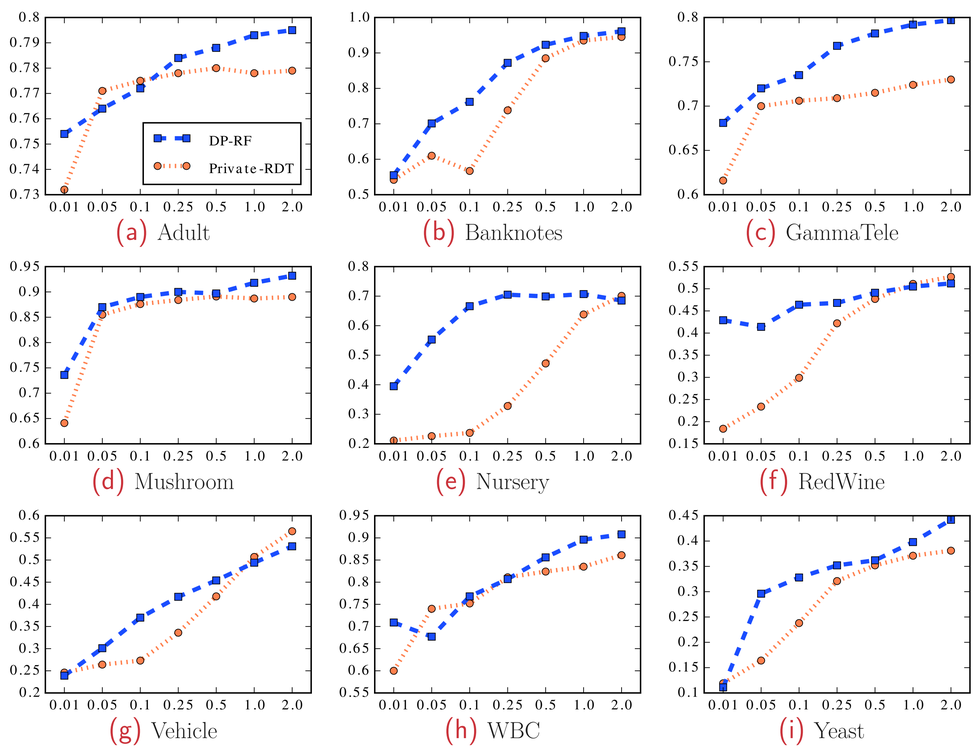

Figure 2: Comparing the prediction accuracy of our technique (DP-RF, blue) to Jagannathan's 2012 Private-RDT (red) at different ϵ values, using 9 different datasets. The y axis represents the prediction accuracy of the classifier and the x axis represents β.