Budget Allocation on Differentially Private Decision Trees and Random Forests

A Thesis Submitted for the Degree of
Doctor of Philosophy

By

Nazanin Borhan

in

School of Software
UNIVERSITY OF TECHNOLOGY SYDNEY
AUSTRALIA
JUNE 2018

© Copyright by Nazanin Borhan, 2018

UNIVERSITY OF TECHNOLOGY SYDNEY
SCHOOL OF SOFTWARE

The undersigned hereby certify that they have read this thesis entitled "**Budget Allocation on Differentially Private Decision Trees and Random Forests**" by **Nazanin Borhan** and that in their opinions it is fully adequate, in scope and in quality, as a thesis for the degree of **Doctor of Philosophy**.

Date: JUNE 2018

Research Supervisors: ______________________
Paul Joseph Kennedy

____
Robin Michael Braun

CERTIFICATE OF ORIGINAL AUTHORSHIP

Date: **JUNE 2018**

Author: **Nazanin Borhan**

Title: **Budget Allocation on Differentially Private Decision Trees and Random Forests**

Degree: **Ph.D.**

I declare that this thesis is submitted in fulfilment of the requirements for the award of Ph.D, in the School of Software, Faculty of Engineering and IT at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed
prior to publication.

Signature of Author

Acknowledgements

I owe my gratitude to all the people who have made this thesis possible and taking the opportunity to thank them has been the best part of writing this dissertation.

First, I would like to express my sincere appreciation to my supervisor, Paul J. Kennedy, for his guidance, patience and financial support over the years. Paul encouraged me to develop my research interests and afforded me a considerable freedom to explore interesting tangents in privacy-preserving data mining and I have greatly benefited from his comments and feedback. Paul, thank you for being a wonderful advisor and for always making time for me when I needed guidance.

I would like to thank my co-supervisor, Robin Michael Braun, for his time and helpful guidance. Additionally, I had the privilege of collaborating with Arik Friedman and Richard Nock, from Data61/CSIRO, on the main chapter of this dissertation (Chapter 3). Richard and Arik are brilliant researchers, and I learned a lot from working with them. Their input helped to improve the quality of this chapter significantly. Richard and Arik, thank you for your help and friendly guidance. I would also like to thank Michele Mooney from Professional Proofreading Services for proofreading this thesis.

My amazing family has shown me constant support and unconditional love, not only through my seemingly never-ending educational journey, but throughout my life. They never stopped believing in me, even when I did, and they helped me push through when I wanted to give up. I feel incredibly blessed to have such a loving and supportive family. I am especially grateful to my mother, who is without doubt the hardest working person I have ever known. She instilled in me an appreciation for education and for hard work and has shown me unconditional love and support in everything that I have done. I am also very grateful to my in-laws, for their love and support, especially during the chaotic final months of my PhD. I would also like to thank my dear friends for their friendship and for providing much-needed distractions that kept me from burning out during my final year.

I owe an enormous debt of gratitude to my husband, Steven Mastwijk, for his constant love, support and patience. Steven, thank you for being a fantastic companion, for picking up my slack at home during deadlines and for never ceasing to believe in me. You encouraged me to keep going during my lowest lows and helped me celebrate my highest points. Thank you for patiently sticking by my side all these years, I love you dearly. I am also thankful to my son, Ilai, who is the pride and joy of my life. You made me a stronger person and gave me tremendous happiness during the last year of my study.

Last, but certainly not least, I gratefully appreciate the financial assistance of Australian government and UTS university by offering me the Australian Postgraduate Award and Women in Engineering and IT Research Excellence scholarships.

To My Husband, ***Steven Mastwijk***

& Our Son, ***Ilai Kian Mastwijk***

Table of Contents

List of Tables

List of Figures

Abstract

Privacy-preserving techniques are necessary to minimize the possibility of identifying and learning sensitive information about individuals from any datasets that have been released or shared. Datasets containing sensitive information on individuals are becoming increasingly public. Although this will support data mining research, information privacy concerns need to be addressed. Many techniques have been developed to preserve the privacy of sensitive public data, with Differential Privacy (DP) being a leading method. Differential privacy, as one of the most important privacy-preserving mechanisms, typically adds noise to prevent the disclosure of individuals' sensitive data. However, adding too much noise can compromise the utility of the output from this mechanism, because the resulting predictions will be too inaccurate. Several techniques can be used to achieve differential privacy in practice, including adding Laplacian noise or Gaussian noise to the output of the computation, or adapting perturbation techniques. Each one of these techniques has been proposed to be applied to the right data mining process and applications. Finding a balance between the privacy of individuals and the utility of the results is one of the biggest challenges in differential privacy.

Differential privacy has received significant attention in the machine learning and data mining communities, in part because the individual privacy guarantee does not

fundamentally contradict the learning goal of *generalizing well* (C. Dwork & Roth, 2014). In supervised learning, where the output of the privacy preservation mechanism is a classifier and the mechanism itself involves a learning algorithm, the minimization of the *classification error* under privacy constraints is not trivial (Chaudhuri, Monteleoni, & Sarwate, 2011; Friedman & Schuster, 2010). Decision tree induction is a textbook case for such a problem. The algorithmic decisions in trees that are made with privacy considerations, have a deep impact on the accuracy of the results. An *improved* privacy-preserved decision tree algorithm, with an order of magnitude of fewer learning samples, can still achieve the same level of accuracy and privacy as the naive implementation.

Although adding Laplacian noise is still one of the main mechanisms to achieve differential privacy in decision trees, a recent line of work has shown that the exponential mechanism leads to better trade-offs between accuracy and privacy on the splitting points of top-down decision tree induction algorithms such as ID3, C4.5 and CART (Blum, Dwork, McSherry, & Nissim, 2005; Friedman & Schuster, 2010). These studies reveal a striking observation: while the splitting criterion of ID3/C4.5 guarantees a better rate of convergence in classification than CART, due to the requirement of fewer interactive queries to reach a given accuracy level, the privacy cost it incurs for a given accuracy level is comparatively larger. This observation is important because the algorithms are greedy; ID3 and C4.5 require fewer interactive queries to reach a given accuracy, but each query incurs a higher privacy cost than is the case for CART.

Many machine learning and minimization tasks, including decision trees, make use of an objective function. At each iteration of a decision tree, a parameter set is defined,

and the objective function returns some scoring value that reflects how good or bad that parameter set is. Then the parameter set is altered, and the process repeats. The process of induction is stopped when the changes in fit become acceptably small and the tree is getting closer to a local or global optima. These stopping rules comprise the rate of convergence of a tree. When an algorithm converges, it has found a parameter set meeting the optimal requirements. From the privacy standpoint, the faster the rate of convergence of a splitting function, the higher the privacy cost it incurs. This inevitably raises the question of whether some splitting functions can achieve two goals simultaneously: fast convergence rates and a small privacy cost. This motivates the research in this thesis to seek other methods to improve the spending of the privacy budget throughout the induction of the decision tree, such as tuning the allocation of the privacy budget across queries.

This thesis presents an adaptive mechanism for allocating the privacy budget, designed for any proper symmetric splitting criterion. In a nutshell, when a tree is splitting nodes, the algorithm probes for the amount of privacy budget to allocate to the split. Some data samples at a node need more privacy budget to build an accurate split, and some are not dependent on spending too much budget. The dynamic budget allocation algorithm will enhance the application of differential privacy on decision trees and forests.

This thesis focuses on the dynamic privacy budget allocation algorithm to have more control over the consumption of the budget in the data mining techniques. There are three contributions in this thesis. Contribution 1 proposes a differentially private budget allocation algorithm on decision tree induction. Contribution 2 extends this schema on random forest algorithm. It is important to consider the adaptive allocation of the budget to the queries that need more or less privacy/utility trade-off. The algorithms are evaluated on synthetic and real datasets for decision trees and random forests. The experiments in this thesis show that the accuracy level of these algorithms are competitively high compared to state-of-the-art models. The budget allocation optimization technique has the potential to be a building block for other data mining techniques and methods. Contribution 3 of this thesis introduces a new splitting criterion for decision trees and evaluates its performance in comparison to other splitting criteria. The result of this evaluation demonstrates that the best splitting criteria for classification under the boosting framework are not necessarily the best to learn in a differentially private setting.

Notations

DP: Differential Privacy

ϵ: Privacy budget Epsilon

CART: Classification and Regression Trees

IG: Information Gain for attribute splitting

Gini Index(G): Gini Index for attribute splitting

Err Score(Err): Misclassification Error Score for attribute splitting

M-Score(M): Matsushita attribute splitting score

ID3: Iterative Dichotomiser 3

RF: Random Forest

PPDP: Privacy Preserving Data Publishing

PPDM: Privacy preserving Data Mining

ACC: Prediction Accuracy

stdev: Standard deviation from the average accuracy prediction

$M(.)$: A learning algorithm

X: A random sample

$\mathcal{D}$: A joint distribution

$\mathcal{S}$: A training set of samples

$\mathcal{S}'$: A neighboring training samples

$\mathcal{T}$: A set of classifiers

h: A (tree) classifier

C: Number of class labels of a tree

$\mathcal{Y}$: Domain of class labels of a tree

ϕ: A permissible function for any symmetric proper loss

ϕ_{IG}: Information Gain criterion

ϕ_{G}: Gini index criterion

ϕ_{Err}: Misclassification Error score

ϕ_{M}: Matsushita's loss

Δ_f: Global sensitivity of function $f(.)$

$\mathcal{L}(h)$: Set of leaves of tree h

m: Number of records in a dataset

m_ℓ: Examples that fall into leaf ℓ

m_ℓ^+: Positive examples reaching ℓ

m_ℓ^-: Negative examples reaching ℓ

d: Dimension of the observations space (number of attributes in the attribute list)

ϵ_{rem}: Remaining privacy budget

t: Adaptive budget threshold