LOCALIZING FAULTS IN NUMERICAL SOFTWARE USING A VALUE-BASED CAUSAL MODEL

by

ZHUOFU BAI

Submitted in partial fulfillment of the requirements
For the degree of Doctor of Philosophy

Dissertation Advisor: Andy Podgurski

Department of Electrical Engineering and Computer Science
CASE WESTERN RESERVE UNIVERSITY

TBD

Contents

1 Causal Inference Based Fault Localization for Numerical Software with NUMFL 1

1.1 Introduction 1

1.2 BART Model Algorithm 2

1.2.1 A Single Regression Tree model 2

1.2.2 A sum of regression trees model 2

1.2.3 A regularization prior 3

1.2.4 Bayesian Backfitting MCMC Algorithm 3

1.3 BART Model with Causal Inference 3

1.4 EMPIRICAL EVALUATION 6

1.4.1 Limitations 6

1.5 RELATED WORK 6

1.6 CONCLUSION 6

Appendices

List of Tables

List of Figures

Acknowledgements

This dissertation would be impossible without the support of my advisors. I would like to thank my research advisor Dr. Rong Xu for her constant support and guidance. Dr. Xu led me into the field of translational biomedical research and guided me in every project. She shared innovative ideas with me, and contributed a lot of time to make my Ph.D. experience productive and exciting. Everything would be different without her. I would like to thank my advisor and dissertation committee chair Dr. Guo-qiang Zhang, who offered me the opportunity to join CCI, where I met great people and started my research. I appreciate all his insightful discussions and constructive suggestions to improve my dissertation and the overall research.

My sincere thanks also go to the members of my committee, Dr. Jing Li, Dr. Xiang Zhang and Dr. M. Cenk Cavusoglu, for their invaluable feedback, scientific suggestions and insightful discussions, which helped me improve this dissertation. I would like to give special thanks to Dr. Xiang Zhang, Dr. Xiaofeng Ren and Dr. Li Li, who generously offered advices and shared research experiences with me during my Ph.D. study.

I would like to thank all present and past members of CCI and all my friends in the EECS department for their love and friendship through all these years during my Ph.D. I would like to thank my family: my husband Zhuofu Bai, and my parents, for their unconditional love and support.

List of Abbreviations

- OMIM: Online mendelian inheritance in man
- GWAS: Genome-wide association study
- UMLS: Unified medical language system
- CBIR: Content-based image retrieval
- SIFT: Scale invariant feature transformation
- HOG: Histograms of oriented gradients
- SVM: support vector machine
- HPRD: Human protein reference database
- PPI: Protein-protein interaction
- HDN: Human disease network
- CRC: Colorectal cancer
- PD: Parkinson's disease
- DMN: Disease manifestation network
- IMPC: International mouse phenotyping consortium
- FDA: Food and drug administration

Development of computational approaches for medical image retrieval, disease gene prediction, and drug discovery

Abstract

by

Yang Chen

With the deluge of biomedical data, developing computational approaches for data analysis and interrogation has become a key step in translational biomedical research. It is critical to leverage existing data to ask the right question and design algorithms for specific biomedical applications. In this dissertation, I propose using domain knowledge to guide the data gathering, data fusion and algorithm design in solving specific biomedical problems. I demonstrate the strategy with applications in three distinct contexts.

The first application is retrieving disease manifestation images from the web for supporting patients' self-education and decision making. The challenge is threefold: heterogeneous irrelevant web images need to be filtered; the positive examples of disease images contain diverse objects and complex backgrounds; and large amounts of manual efforts in generating training data are unaffordable. We observe that detecting disease-affected abnormal organs may greatly reduce the manual labeling efforts. In our approach, we extract the disease-organ semantic relationships from ontologies to guide the organ detection with pre-trained detectors. Comparing with a standard supervised method, we improve the average precision by 4% while reduce the manual efforts by 85%.

In the second application, we develop three disease-specific models to detect genetic basis for human diseases. For parasitic infectious diseases, we construct a cross-species genetic network to model host-pathogen interactions and analyze the network to predict disease associated genes. We apply the approach on malaria and demonstrate the potential of the top-ranked genes in guiding anti-malaria drug discovery. For multifactorial diseases, we assume that phenotypic similarity reflects common genetic basis between diseases. We explore a new disease phenotype data source in medical ontologies and construct the Disease Manifestation Network (DMN). Then we integrate multiple phenotype networks with genetic networks to predict genes. We apply the approach on Crohn's disease, and demonstrate the translational potential of the predicted genes in drug discovery. Last, we identify the mutual comorbidity for colorectal cancer and obesity in the comorbidity network to detect genetic basis for the link between the two diseases.

Finally, I present a drug repositioning approach combining disease genetics and phenotypic descriptions for mouse genetic mutations. Disease associated genes have the potential to guide drug discovery. On the other hand, the mouse phenotypes provide knowledge on gene functions, which is impossible to be obtained in human. In our approach, we identify disease-specific mouse phenotypes using well-studied disease genes, and search all FDA-approved drugs for the candidates that share similar mouse phenotype profiles with the disease. We used the approach to predict drugs for Parkinson's disease, and demonstrate significantly improvements comparing with a state-of-art approach based on mouse phenotype data. Overall, I demonstrate the effectiveness of the domain knowledge guided computational approaches in concrete biomedical applications.

In summary, I demonstrate the effectiveness of computational algorithms in translational biomedical research. I demonstrate that my computation-based work have great potential in elucidating disease genetic basis, finding innovative drugs, and improving patient health education.

Chapter 1

Causal Inference Based Fault Localization for Numerical Software with NUMFL

1.1 Introduction

The motivation for using Bayesian Additive Regression Trees in statistical fault localization is:

- BART can flexibly fit non-linear response surfaces even with a large number of predictors

- BART does not require the researcher to specify the functional form of the relationship between treatment and outcome

- Program dependence graph can be explained as a causal graph. The casual inference techniques has been proved to be effective in both coverage based fault localization and value based fault localization. BART has good performance in estimating average causal effect of binary treatment and has potential in estimating failure-causing effect of continuous treatment [].

- BART algorithm software is freely available and easy to use [].

The main contributions of this paper are:

The rest of the chapter is organized as follows:

1.2 BART Model Algorithm

The BART model algorithm consists of three pars: a sum of regression trees model, a regularization prior and a fitting algorithm Marcov Chain Mote Carlo (MCMC)

1.2.1 A Single Regression Tree model

Regression tree is one type of decision tree that predicts the value of a target variable based on several input variables. Regression tree handles the situation when the target variable is continuous. Figure **??** shows a single regression tree model. All the interior nodes of a regression tree have decision rules which send the input data set to either left or right side. After the input data set go through the interior nodes and reach the bottom of the tree, the data set is divided into several disjoint subgroup. Each group of data is represent by a leaf node.

A single regression tree model is denoted as: $y = g(T, R, M) + \epsilon$

Here R denotes a binary regression tree consisting a set of interior node decision rules and a set of terminal nodes, and let $M = \left\{ \mu_1, \mu_2, ..., \mu_b \right\}$ denotes a set of parameter value associated with each of the b terminal nodes of R. Each terminal node represent a regression model of outcome Y on Treatment T

1.2.2 A sum of regression trees model

BART model

1.2.3 A regularization prior

1.2.4 Bayesian Backfitting MCMC Algorithm

1.3 BART Model with Causal Inference

For binary treatment, BART model primarily estimate failure-causing effects such as $E(Y(1)|\boldsymbol{X}=\boldsymbol{x})-E(Y(0)|\boldsymbol{X}=\boldsymbol{x})=E(Y|T=1,\boldsymbol{X}=\boldsymbol{x})-E(Y|T=0,\boldsymbol{X}=\boldsymbol{x})=f(1,\boldsymbol{(}x))-f(0,\boldsymbol{(}x))$. The algorithm contains the following steps:

1. Fit BART model using MCMC algorithm to full sample

2. Get posterior prediction for each unit by setting the treatment variable value $T=1$ and keep confounding variable value unchanged..

3. Get posterior prediction for each unit by setting the treatment variable value $T=0$ and keep confounding variable value unchanged.

4. Calculate the difference between the posterior predictions for each unit.

5. Estimate failure-causing effect by averaging all the differences of posterior predictions.

In step 1, the fitted BART model characterized the causal relationship between treatment T and outcome Y given the confounding $\boldsymbol{X}$. We can use the fitted model to predict the outcome Y under different $(T,\boldsymbol{X})$ conditions. For a untreated unit $(T=0,\boldsymbol{X}=\boldsymbol{x})$, if we set the treatment variable to 1 and then input $(T=1,\boldsymbol{X}=\boldsymbol{x})$ into the BART model, the output is the estimated outcome for that unit in treated condition. Similarly, we can estimate the outcome of a treated unit in untreated condition by inputing $(T=0,\boldsymbol{X}=\boldsymbol{x})$ into the BART model. Thus, in step2, the BART model is used to predict outcome for each unit at observed treatment condition. In step 3, the fitted BART model is used to estimate posterior predictions for each unit at counterfactual condition. Thus, the difference calculated in step 4 is the causal effect estimation for each unit. The average of these differences is failure-causing effect causal effect which is estimated in step 5.

If the treatment variable is continuous, the causal effect of treatment T on outcome Y is characterized by a function $r_e(T)$, which is called dose-response function in medical research. For the failure-causing effect, the dose response functions of continuous treatments are usually non-linear. For example, in Chapter**??**, NUMFL use parameterized quadratic model and double linear model to approximate the dose response function within subclasses and get reasonable well result. But the parameterized model has limitations. It usually requires user to make assumptions to specify form of the dose-response function. For example, in NUMFL, we make some assumptions ($Assumptioin 1, 2, 3$). The $Assumptions 3$ is: executions with large absolute treatment errors have larger output errors than executions with small values of treatment errors. Although this assumption is often holds in numeric programs, it is not always to be true. In this case, the dose response curve of treatment T on outcome Y is likely to deviate from the quadratic model $Y = \zeta {T_e}^2 + \eta T_e + c$, which may result in a bad estimation of failure-causing effect. Comparing to parametric model used in NUMFL, BART model does not require user to make assumptions or specify the functional form of the dose response function. The sum of trees structure of BART model can approximate the non-linearity of the DRF during the training phone with MCMC.

failure-causing effect estimation is a challenge for BART model. In NUMFL, the failure-causing effect is characterized by the coefficient of the regression model. But in BART model, the sum of trees structure does not such parameters can directly characterize the ACE. To solve this problem, we propose to estimate the treatment causal effect at each unit in the sample. The average of the treatment causal effect on all sample units is the estimated ACE. The causal effect of treatment T on outcome Y for a single unit can be estimated by increasing the treatment variable value of the unit and see how outcome changes. For example, assume a unit i has treatment variable $T = t$ and confounding variables$\boldsymbol{X} = \boldsymbol{x}$, we can use the fitted BART model to estimate the outcome of the unit $Y = y$. Then we increase the treatment variable to $t' = t+\varepsilon$, here ε is a small number. We input T' and $\boldsymbol{X}$ into the fitted BART model and get the estimated posterior $Y = y'$. Then the estimated causal effect of T on Y for unit i is $|y - y'|$. The algorithm is as follows:

1. Fit BART model using MCMC algorithm to full sample

2. Increase the treatment variable value t of each unit by ε. The new treatment value $t' = t + \varepsilon$ and the original confounding variables forms a new data set.

3. Calculate the difference between the posterior predictions of each unit in original data set and the posterior predictions of each unit in the new data set.

4. Estimate failure-causing effect by averaging all the differences of posterior predictions.

In the above algorithm, the BART model fitted in first step is used to approximate a function $Y = f(T, \boldsymbol{(}X\boldsymbol{)})$ that specify the dose-response function of the continuous treatment T and confounding variables $\boldsymbol{X}$. Step 2 and Step 3 estimate the treatment causal effect for each sample unit. Step 4 estimate the failure-causing effect of treatment on outcome.

One problem in step 2 is how to choose the value of the small number ε. To estimate the casual effect, the increased treatment variable value $t' = t + \varepsilon$ should be a reasonable value. However, treatment variables in different numerical expressions have different scale of values, it is hard to find a constant ε which can make $t + \varepsilon$ be a reasonable value for all the treatments. To address this problem, we use the method illustrated by Figure**??**. First, we sort the sample data set in increasing order according to the value of the treatment variables before step 2. In step 2, the increased treatment variable value t' is equal to the treatment variable value t of the next unit in the list. The last unit in the list will be discarded. Thus, for a data set with n observational units, we will have a new data set of $n-1$ units after increasing the treatment variable. In this method, the increased value of treatment variable is reasonable, because it is belong to one of the observation unit in the data set.

The Algorithm of BART model based statistical fault localization is shown in Figure:

1.4 EMPIRICAL EVALUATION

1.4.1 Limitations

1.5 RELATED WORK

1.6 CONCLUSION