BAYESIAN RANDOM FORESTS FOR HIGH-DIMENSIONAL CLASSIFICATION AND REGRESSION WITH COMPLETE AND INCOMPLETE MICROARRAY DATA

OLANIRAN RIDWAN OYEBAYO

A thesis submitted in fulfillment of the requirement for the award of the Doctor of Philosophy in Science

Faculty of Applied Sciences and Technology
Universiti Tun Hussein Onn Malaysia

NOVEMBER 2018
To my family
ACKNOWLEDGEMENT

I owe my deepest gratitude to Dr. Mohd Asrul Affendi Bin Abdullah who made this thesis possible and supported me in every possible way. Likewise, I am heartily thankful to Dr. Khuneswari A/P Gopal Pillay for her supervision and support.

I also would like to thank the Ministry of Higher Education (MOHE) and Office of Research, Innovation, Commercialization and Consultancy Office (ORICC), Universiti Tun Hussien Onn Malaysia (UTHM) for financially supporting this research under the postgraduate research grant (GPPS) [Vot, U607].

Olaniran Ridwan Oyebayo, UTHM
ABSTRACT

Random Forests (RF) are ensemble of trees methods widely used for data prediction, interpretation and variable selection purposes. The wide acceptance can be attributed to its robustness to high dimensionality problem. However, when the high-dimensional data is a sparse one, RF procedures are inefficient. Thus, this thesis aims at improving the efficiency of RF by providing a probabilistic framework using Bayesian reasoning. The modification comprises of two main modelling problems: high-dimensionality and missing data. These problems were extensively studied within the scope of classification (binary and multi-class) and regression (linear and survival). The new procedure called Bayesian Random Forest (BRF) focuses on modification of terminal node parameter estimation and selection of random subsets for splitting. BRF algorithm combines the strengths of random subset and greedy selection procedures in creating new maximal ordered variable relevance weights. These weights are in turn used to develop new impurity functions for selecting optimal splits for each tree in a forest. BRF works mainly because the maximal weights are computed using a data-driven procedure called bootstrap prior which was shown to satisfy the uniformly minimum variance property under mild regularity conditions. In addition, BRF ensures that important variables are selected at each subset selection step, thus reducing false signals and eventually improving accuracy of models. As a further extension, missing covariates problem was also handled by pre-imputing the variables using Multivariate Imputation by Chain Equation (MICE) before building forests. Performance analysis was achieved using simulated and eighteen real-life classification and regression microarray cancer datasets. Empirical results from the data analysis established appreciable supremacy over RF and several other competing methods.

Keyword: Random Forest, Bayesian Inference, Classification, Regression, Missing Data.
ABSTRAK

Random Forest (RF) adalah ensemble kaedah pokok yang banyak digunakan untuk ramalan data, tafsiran dan tujuan pemilihan yang berubah-ubah. Penerimaan yang luas boleh dikaitkan dengan kekukuhannya kepada masalah dimensi yang tinggi. Walau bagaimanapun, apabila data dimensi tinggi, prosedur RF tidak cekap. Oleh itu, tesis ini bertujuan meningkatkan kecekapan RF dengan menyediakan rangka kerja probabilistik menggunakan penalaran Bayesian. Ia meliputi dua masalah pemodelan utama: data dimensi tinggi dan data yang hilang. Masalah ini dikaji secara meluas dalam skop klasifikasi (binari dan berbilang kelas) dan regresi (linear dan survival).

Ini adalah prosedur baharu yang dikenali sebagai Bayesian Random Forest (BRF) yang memberi tumpuan kepada pengubahsuaian terminal anggaran nod parameter dan pemilihan subset rawak untuk pemisahan atau pembahagian. Algoritma BRF menggabungkan kekuatan subset rawak dan prosedur pemilihan dalam mewujudkan maksimal baharu dengan berat pembolehubah yang relevan. Nilai berat ini seterusnya digunakan untuk membangunkan fungsi baharu untuk memilih pecahan optimum bagi setiap pokok dalam hutan. Kaedah BRF berfungsi degan pemberat maksimum dikira dengan menggunakan prosedur yang didorong oleh data yang dipanggil bootstrap seperti yang telah ditunjukkan untuk memenuhi kawasan varians yang seragam secara minimum di bawah keadaan kekerapan sederhana. Di samping itu, BRF memastikan pembolehubah penting yang dipilih pada setiap subset semasa langkah pemilihan, sekali gus mengurangkan isyarat palsu dan akhirnya meningkatkan ketepatan model.

Seterusnya, masalah kehilangan covariates juga dikendalikan dengan memasukkan pembolehubah menggunakan Imputasi Multivariat melalui Pasamaan Rontai (MICE) sebelum membina hutan. Kaedah telah diuji dengan menggunakan data simulasi dan data pesakit kasah sebenar. Hasil kajian menunjukkan kaedah BRF memberikan hasil yang lebih baik barbanding kaedah RF dan beberapa kaedah lain.

Kata kunci: Hutan Rawak, Kesimpulan Bayesian, Klasifikasi, Regresi, Data Hilang
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td>i</td>
</tr>
<tr>
<td>DECLARATION</td>
<td>ii</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>iv</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>v</td>
</tr>
<tr>
<td>ABSTRAK</td>
<td>vi</td>
</tr>
<tr>
<td>CONTENTS</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF SYMBOLS AND ABBREVIATIONS</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td>xii</td>
</tr>
<tr>
<td>LIST OF PUBLICATIONS</td>
<td>xvi</td>
</tr>
</tbody>
</table>

### CHAPTER 1 INTRODUCTION

1.1 Background of the study  
1.2 Problem statement  
1.3 Research objectives  
1.4 Significance of study  
1.5 Scope and limitation  
1.6 Organization of chapters

### CHAPTER 2 LITERATURE REVIEW

2.1 Introduction  
2.2 Modelling high-dimensional data  
2.3 Variable selection in high-dimensional data analysis  
2.3.1 Variable selection in high-dimensional classification and regression  
2.4 Tree-Based methods for high-dimensional classification and regression  
2.5 Ensemble tree-based methods  
2.5.1 Moving from Bagging to Random Forests
2.6 Missing data issues 23
2.7 Modelling high-dimensional microarray cancer data 26
2.8 Chapter summary 29

CHAPTER 3 TREE-BASED METHODS AND RANDOM FOREST 33
3.1 Introduction 33
3.2 Data description 33
3.3 Review of tree-based models 34
3.4 Review of Classification and Regression Trees (CART) 37
  3.4.1 Regression trees 37
  3.4.2 Classification trees 39
  3.4.3 Survival trees 39
3.5 Review of Random Forests 41
  3.5.1 Missing data algorithms with Random Forest 42
3.6 Review of Bayesian modelling 42
  3.6.1 Prior distribution 43
  3.6.2 Bayesian computation 44
3.7 Performance comparison with Box plot 46

CHAPTER 4 BAYESIAN RANDOM FOREST 47
4.1 Introduction 47
4.2 Justification for updating Random Forests 47
4.3 Bayesian Random Forest for Complete Data 52
4.4 Bayesian Random Forest for Gaussian Response Regression 53
  4.4.1 Priors and posterior specification of Bayesian Random Forest for Gaussian Response Regression 54
  4.4.2 A new weighted splitting of Bayesian Random Forest for linear regression of high-dimensional Data 69
4.5 Bayesian Random Forest for classification 75
  4.5.1 Priors and posterior specification for Bayesian Random Classification Forests 75
  4.5.2 A new weighted splitting of Bayesian Random Forest for classification of high-dimensional data 78
4.6 Bayesian Random Forest for survival 81
LIST OF TABLES

2.1 Review of Frequentist Random Forest extensions 31
2.2 Review of Bayesian Random Forest extensions 31
2.3 Review of Bayesian and Frequentist Missing Data Extensions of Random Forest 32
5.1 Tuning parameters set-up for the various methods used in data analysis. 93
5.2 Average test Root Mean Square Error (ARMSE) over 10-folds cross-validation for Gaussian response. 95
5.3 Average test Root Mean Square Error (ARMSE) over 10-folds cross-validation for Gaussian response. 96
5.4 Average test Root Mean Square Error (ARMSE) over 10-folds cross-validation of the three missing data mechanisms with the proportion of missing observation 0.25, 0.5 and 0.75 for Gaussian response. 99
5.5 Average test RMSE and (Standard error) over 10-folds cross-validation for regression cancer datasets. 101
5.6 Average test Misclassification Error Rate (AMER) over 10-folds cross-validation for binary categorical response. 104
5.7 Average Misclassification Error Rate (AMER) over 10-folds cross-validation of the three missing data mechanisms with the proportion of missing observation 0.25, 0.5 and 0.75 for binary categorical response. 105
5.8 Average test accuracy and (Standard error) over 10-folds cross-validation for binary categorical response cancer datasets. 109
5.9 Average test Misclassification Error Rate (AMER) over 10-folds cross-validation for Multiclass categorical response 111
5.10 Average Misclassification Error Rate (AMER) over 10-folds cross-validation of the three missing data mechanisms with the proportion of missing observation 0.25, 0.5 and 0.75 for multi-class categorical response. 112
5.11 Average test accuracy and (Standard error) over 10-folds cross-validation for multiclass categorical response cancer datasets.

5.12 Average test Root Mean Square Error (ARMSE) over 10-folds cross-validation for survival response.

5.13 Average test Root Mean Square Error (ARMSE) over 10-folds cross-validation of the three missing data mechanisms with the proportion of missing observation 0.25, 0.5 and 0.75 for survival response.

5.14 Average test RMSE and (Standard error) over 10-folds cross-validation for survival cancer datasets.

5.15 Average proportion of relevant variables selected in 10 folds cross validation.

A.1 Datasets with sample size n and number of predictors p
LIST OF FIGURES

2.1 Heat map for the DNA data with 100 rows (genes) shown and 64 columns (samples), ranging from bright green (negative, underexpressed) to bright red (positive, overexpressed). Missing values are grey. Source: James et al. (2013) 28

3.1 CART and Partitions. The first plot shows a CART tree while the second plot shows partitioning into five regions (James et al., 2013). 35

4.1 Probability of selecting relevant variables at varying dimensionality $p$ 52

4.2 Framework of Bayesian Random Forest (BRF) for Classification (binary or multiclass) and regression (linear or survival) of high-dimensional complete data 53

4.3 Simulation plot 1 of the hybrid algorithm for Bayesian Random Forest regression 60

4.4 Simulation plot 2 of the hybrid algorithm for Bayesian Random Forest regression 61

4.5 Single regression tree plot from forest of five trees using Bayesian Random Forest (BRF) hybrid Gibbs and MH procedure. Coloured box corresponds to terminal node or final prediction and white box corresponds to decision node or split point. 62

4.6 Framework of Hybrid Bayesian Random Forest (BRF) for Classification (binary or multiclass) and regression (linear or survival) of high-dimensional incomplete data 90

5.1 Boxplot of test 10-folds cross-validation RMSE of Scenario 1 for Gaussian response. The black middle line in each box represents the median. The dots represents outliers in RMSE results. The outliers in GBM is the highest. 96

5.2 Boxplot of test 10 folds cross-validation RMSE of Scenario 2 for Gaussian response. The black middle line in each box represent the median. 97
5.3 Boxplot of test RMSE for the methods over 10 folds cross-validation under MCAR missing strategy for Gaussian response. The black middle line in each box represent the median. The dots represent outliers in RMSE results.

5.4 Boxplot of test RMSE for the methods over 10 folds cross-validation under MAR missing strategy for Gaussian response. The black middle line in each box represent the median. The dots represent outliers in RMSE results.

5.5 Boxplot of test RMSE for the methods over 10 folds cross-validation under MNAR missing strategy for Gaussian response. The black middle line in each box represent the median. The dots represent outliers in RMSE results.

5.6 Boxplot of test MER for the methods over 10-folds cross-validation for binary categorical response. The black middle line in each box represent the median. The dot represent outliers. The BRF method is more stable at $p = 10000$ with no outliers.

5.7 Boxplot of test MER for the methods over 10-folds cross-validation under MCAR missing strategy for binary categorical response. The dot represent outlier. Adverse effect of deleting missing cases can also be observed with BRF and BART.

5.8 Boxplot of test MER for the methods over 10-folds cross-validation under MAR missing strategy for binary categorical response. The black middle line in each box represent the median. The dot represent outlier. Adverse effect of deleting missing cases can also be observed with BRF and BART.

5.9 Boxplot of test MER for the methods over 10 folds cross-validation under MNAR missing strategy for binary categorical response. The black middle line in each box represent the median. The dot represent outlier. Adverse effect of deleting missing cases can also be observed with BRF and BART.

5.10 Boxplot of test MER of the methods over 10-folds cross-validation for the ten binary categorical response datasets. The black middle line in each box represent the median. The dots represent outlier and it is fewer with BRF.
5.11 Boxplot of test MER for the methods over 10-folds cross-validation for multi-class categorical response. The black middle line in each box represent the median. The dots represent outlier.

5.12 Boxplot of test MER for the methods over 10 folds cross-validation under MCAR missing strategy for multi-class categorical response. The plot shows that RF is better than BRF for incomplete data arising from MCAR.

5.13 Boxplot of test MER for the methods over 10 folds cross-validation under MAR missing strategy for multi-class categorical response. The black middle line in each box represent the median. The plot shows that BRF is better than RF for incomplete data arising from MAR.

5.14 Boxplot of test MER for the methods over 10 folds cross-validation under MNAR missing strategy for multi-class categorical response. The black middle line in each box represent the median. The plot shows that RF is similar to BRF for incomplete data arising from MNAR.

5.15 Boxplot of test MER of the methods over 10 folds cross-validation for the three multiclass categorical response datasets. The black middle line in each box represent the median. The plot shows BRF is the best with minimal median MER for all the datasets.

5.16 Boxplot of test 10 folds cross-validation RMSE of survival response. The plot shows that BRF is the best in terms of minimum median RMSE (middle black line).

5.17 Boxplot of test MER for the methods over 10 folds cross-validation under MCAR missing strategy for survival response. The plot shows that BRF is the best in terms of minimal median RMSE for missing data arising from MCAR.

5.18 Boxplot of test MER for the methods over 10 folds cross-validation under MAR missing strategy survival response. The plot shows that BRF is the best in terms of minimal median RMSE for missing data arising from MAR.

5.19 Boxplot of test MER for the methods over 10 folds cross-validation under MNAR missing strategy survival response. The plot shows that BRF is the best in terms of minimal median RMSE for missing data arising from MNAR.

5.20 Fitted survival function $S(t)$ of various methods to Breast2 and Lymphoma2 cancer datasets.
Boxplot of 10-folds RMSE for the various methods used for survival response. The plot shows that BRF has the least median RMSE in the two real-life dataset used.
## LIST OF APPENDICES

<table>
<thead>
<tr>
<th>Letter</th>
<th>Chapter and Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Chapter 3: Data Description</td>
<td>146</td>
</tr>
<tr>
<td>B</td>
<td>Chapter 3: R code for CART plot</td>
<td>151</td>
</tr>
<tr>
<td>C</td>
<td>Chapter 3: R code for Justification of BRF</td>
<td>152</td>
</tr>
<tr>
<td>D</td>
<td>Chapter 3: R code for Hybrid MH for BRF</td>
<td>153</td>
</tr>
<tr>
<td>E</td>
<td>Chapter 4: R codes Single Tree plot and Regression Forest for 3 covariates</td>
<td>157</td>
</tr>
<tr>
<td>F</td>
<td>Chapter 5: R codes for Simulation and Real-life Results of BRF and competing methods</td>
<td>163</td>
</tr>
</tbody>
</table>
LIST OF SYMBOLS AND ABBREVIATIONS

\[ \lfloor . \rfloor \] - Floor function; the largest integer not greater than (.)
\( \beta_m \) - Tree Node parameter for mean
\( p_m \) - Tree node parameter for proportion
\( R_m \) - Region or partition of a tree
\( p \) - Number of predictors
\( D \) - Dataset with dimension \( n \times p \)
\( \theta \) - Parameter of a model
\( p_m \) - Tree node parameter for proportion
\( n \) - Sample size in data \( D \)
\( n_m \) - Sample size of a tree terminal node
\( Q_m(T) \) - Impurity function of a tree
\( \delta_{im} \) - Censoring indicator for terminal node \( m \)
\( H_m(t) \) - Cumulative Hazard Function of a tree node at time \( t \)
\( \cap \) - Intersection
\( \in \) - element of
\( \pi(\theta_i, \theta) \) - accept/reject probability of MH algorithm
\( \text{mtry} \) - Number of variables randomly selected as candidates for splitting a node
\( \text{nsplit} \) - Number of splits in a tree
\( I \) - Classification or regression tree
\( \sigma^2 \) - Sum of trees variance
\( \Sigma \) - Covariance matrix of tree node parameter
\( J \) - Number of trees
\( B \) - Number of bootstrap for bootstrap prior
\( T_{(k)} \) - Ordered t- statistic
\( F_{(k)} \) - Ordered F- statistic
\( Z_{(k)} \) - Ordered z- statistic
\( \text{BART} \) - Bayesian Additive Regression Trees
\( \text{AIC} \) - Akaike Information Criterion
\( \text{BIC} \) - Bayesian Information Criterion
\( \text{MICE} \) - Multiple Imputation by Chain Equation
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCS</td>
<td>Full Conditional Specification</td>
</tr>
<tr>
<td>MH</td>
<td>Metropolis-Hasting Algorithm</td>
</tr>
<tr>
<td>CMH</td>
<td>Coupled Metropolis-Hasting Algorithm</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forest</td>
</tr>
<tr>
<td>BRF</td>
<td>Bayesian Random Forest</td>
</tr>
<tr>
<td>DRF</td>
<td>Dynamic Random Forest</td>
</tr>
<tr>
<td>CHF</td>
<td>Cumulative Hazard Function</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximization</td>
</tr>
<tr>
<td>CART</td>
<td>Classification And Regression Trees</td>
</tr>
<tr>
<td>BCART</td>
<td>Bayesian Classification And Regression Trees</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov Chain Monte-Carlo</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum Aposteri</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger Ribonucleic Acid</td>
</tr>
<tr>
<td>BP</td>
<td>Bootstrap Prior</td>
</tr>
<tr>
<td>DL</td>
<td>Dirichlet Laplace Prior</td>
</tr>
<tr>
<td>CWD</td>
<td>Case Wise Deletion</td>
</tr>
<tr>
<td>BMA</td>
<td>Bayesian Model Averaging</td>
</tr>
<tr>
<td>BART – BMA</td>
<td>Bayesian Additive Regression Trees using Bayesian Model Averaging</td>
</tr>
</tbody>
</table>
LIST OF PUBLICATIONS


CHAPTER 1

INTRODUCTION

1.1 Background of the study

The advancement in computing application and technology has enhanced the collection of data of unprecedented size and complexity. These datasets usually contain many variables than the available sample size and thus referred to as high-dimensional data. Indeed, many modern scientific investigations require the analysis of high-dimensional data. Such data include microarrays, proteomics, brain images, videos, functional data and high-frequency financial data (Fan et al., 2009; Hernández et al., 2018). This application demand presents many new challenges as well as opportunities for those in statistics and machine learning. Although significant success has been recorded in recent years, there still exists some loopholes in handling such datasets.

One of the common statistical modelling problems involves exploring the relationship between a response variable $Y$ and its associated covariates (or features) $X_1, \ldots, X_p$, based on a predefined sample size $n$. The main property of modern problem defined in the previous paragraph is that the dimensionality $p$ is by far larger than $n$. In mathematical parlance, $p$ can be considered as a function of $n$, which diverges to infinity. The dimensionality overgrows when interactions of the features are considered, which is necessary for many scientific endeavours. For example, in disease classification using microarray gene expression data (West et al., 2001), the number of arrays is usually in the order of tens or hundreds while the number of gene expression profiles is in the order of tens of thousands; in the study of document classification, the sample size (documents or sentences) can be in the order of thousands, but the number
of words can be in the order of millions; in the study of image processing, the sample size (images) can be in the order of hundreds while the number of pixels in each image can be in millions.

The problem of high-dimensional classification and regression has long been studied by statisticians and computer scientists (Cortes and Vapnik, 1995; Vapnik, 2013; Hastie et al., 2010, 2015). Hastie et al. (2010) reported that high-dimensional classification and regression modelling can be achieved in two broad ways: variable (feature) selection and penalized (regularized) regression approaches. Variable selection is an approach to making existing low dimensional data modelling approaches work with high-dimensional data while penalized regression involves imposing some constraint on the dimensionality to achieve a similar objective. The major strength of variable selection approaches is retaining the good properties of low dimensional approaches such as Maximum Likelihood Estimator (MLE) at the expense of inadequacy in covering the complex nature of high-dimensional datasets. Penalized approaches such as (LASSO; Tibshirani (1996), SCAD; Fan and Li (2001)) among others provide an edge on the problem but at the expense of introducing bias in estimation. The common disadvantage of the two approaches is the failure to capture the complex nature of high-dimensional datasets such as (interaction, non-linearity, non-normality, missing or incomplete data) among others (Hernández et al., 2018). One of the robust procedure that has been shown to withstand these issues in low and high-dimensional situations is Classification and Regression Trees (CART) (Breiman et al., 1984; Hwang et al., 2017). The CART is a non-parametric statistical procedure that relaxes the dimensionality assumption and natural adaptation to modelling interaction, non-linearity, and incorporation of missing data using the surrogate splitting procedure.

Amidst the strength of CART regarding simplicity and interpretability lies the instability drawback which often results in loss of accuracy. The late 20th century witnessed the explosion of a new methodological framework for combining several models to build a new holistic model. The approach is termed as ensemble modelling. One of the earliest ensemble approaches within the CART framework is Bagging
(Bootstrap Aggregating) (Breiman, 1996a). The bagging of predictor takes multiple versions of the bootstrap sample (Efron and Tibshirani, 1994) from the training dataset and fits an unpruned CART to each bootstrap sample. The final predictor is obtained by averaging over different versions of the model. This procedure works surprisingly well and outperforms its competitors in most situations. Some intuitive explanations of how and why it works were given in Breiman et al. (1998). This idea has motivated much subsequent work including Random Forests (RF) (Breiman, 2001), a general framework for tree ensembles. RF provides an update to Bagging by replacing the use of all covariates in the splitting step of the CART with random sub sampling of covariates. This approach helps to reduce the correlation between adjacent trees and thus improving the predictive accuracy.

Although Bagging is the bedrock of RF, the foundation of random sampling of the variable space used by RF can be traced to the earlier work of Hoeting et al. (1999) on Bayesian Model Averaging (BMA). BMA treats the model space as random with differing probability of being the choice model. This approach introduces uncertainty to modelling and thus provides more realistic model than RF. Apart from RF and BMA ensemble, there exist another approach called boosting. Boosting methods provides update on the averaging scheme of RF, by reducing impacts of weak trees that contained irrelevant predictors. One of the popular boosting algorithms is Gradient Boosting (Friedman, 2001) and it’s the foundation of most recent Bayesian CART ensemble methods. These methods include Bayesian Additive Regression Trees (BART) (Chipman et al., 2010; Bleich et al., 2014; Pratola, 2016) Smoothed Bayesian Additive Regression Trees (SBART) (Linero, 2018), and Bayesian Additive Regression Tree using Bayesian Model Averaging (BART-BMA) (Hernández et al., 2018) among others. Bayesian procedures have generally been proven to be better than the frequentist or non-Bayesian procedures but with reliance on adequate prior to update the information from the data. In this thesis, we focus on high-dimensional classification and regression of microarray data using Bayesian statistical approaches to RF.
1.2 Problem statement

The complexity of high-dimensional data has necessitated the development of many versions of RF algorithms within the scope of classification and regression modelling. One of the features of high-dimensional dataset is sparsity i.e small number of relevant predictors within the predictor space. This feature is often observed in microarray data where there is fewer number of relevant genes associated with a particular disease outcome (Friedman, 2001; Hernandez et al. 2018). The arbitrary random subsampling of \( \sqrt{p} \) variables in the case of classification or survival and \( p \) variables in the case of regression used by RF fails to capture this feature (Hastie et al., 2010). The unrealistic assumption of RF under this situation is that the predictor space should be well populated with relevant variables (genes) in order to achieve reasonable accuracy. Although, boosting procedure Friedman (2001) specifically addressed this issue by boosting weak trees instead of averaging all the trees as in RF. The approach is robust to increase in dimensionality \( p \) but at the expense of reduced predictive accuracy when compared to RF. The Bayesian modified boosting approach BART does not focus on the problem but only provides an alternative way of estimating a boosting motivated ensemble of CART modelling. In fact, the approach is not robust to increase in dimensionality as reported in Linero (2018). This research therefore present a new framework for RF by introducing a Bayesian motivated greedy search alongside the random search within the splitting step of RF so as to enhance identification and selection of relevant variables.

Another issue with RF modelling of high-dimensional data is estimation of trees descriptive summaries using non-probabilistic approach. RF does not have any model except a step-by-step procedure of prediction. Microarray data are generally posed with small sample size \( n \) issue and this reduces the information that can be obtained from the data. For example, the microarray experiment conducted by West et al. (2001) on the prognosis of breast cancer involves only 49 patients where on each subject 7129 gene expression profiles were recorded. Non-probabilistic approaches such as RF requires reasonable larger sample size to obtain reliable prediction.
Bayesian approach have been found to possess higher statistical power in this situation. The BART procedure was found to be better than RF, GBM and LASSO in the case of binary and linear regression simulation studies conducted by Chipman et al. (2010). Another study by Sparapani et al. (2016) on nonparametric Bayesian survival analysis revealed similar results when compared with the Cox model. In a similar study by Taddy et al. (2011), a new approach was developed called Bayesian Forest (BF). They used posterior of trees instead of bootstrap of trees based on a nonparametric Bayesian model using multinomial draws. BF tried to mimic RF by replacing the bootstrapping procedure by Efron and Tibshirani (1994) by its Bayesian counterpart (Bayesian Bootstrap) (Rubin, 1981). This implies BF focuses on the data generating process of RF but not its parametric summary statistics procedures. The authors also reported that the method is not different from RF in terms of accuracy. Thus, in this thesis, we present a new method of estimating trees descriptive summaries using bootstrap prior Bayesian approach.

The last issue with RF considered in this thesis is missing values or missing data problem that reduces the quality of data during analysis. Microarray data often contain missing data which arises as a result of death or withdrawal of patient from the study. Missing data are popularly handled using the imputation methods (Schafer and Graham, 2002). Also, there exist some crude solutions e.g. complete case and available case analysis as well as hot-deck, predictive distribution substitution, conditional mean and single imputation by mean which usually results in power reduction, inference biased, loss of efficiency and variance underestimation. The dominant approach for handling missing values is Multiple Imputations by Chained Equations (MICE) (Van Buuren et al., 2006; White et al., 2011) also known as imputation by Full Conditional Specification (FCS) (Kapelner and Bleich, 2015). Its superiority to crude and single imputation methods has been established in many publications. Alternatives methods for handling missing values apart from imputation are specially designed methods with built-in procedures to handle missing values. The methods include and not limited to surrogate splitting of CART and proximity imputation of RF (Kapelner and Bleich, 2015). Tang and Ishwaran (2017) reported an
REFERENCES


Ishioka, T. (2013). Imputation of missing values for unsupervised data using the proximity in random forests. In International Conference on Mobile, Hybrid,


networks. Nature medicine. 7(6), 673.


of life and substance abuse in Russia. BMC medical research methodology. 15(1), 71.


computer sciences. 43(6), 1947-1958.


