View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Memorial University Research Repository

Using the Machine Learning Approach to Predict Patient Survival from High-Dimensional Survival Data

by

© *Wenbin Zhang*

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Master of *Science*

Department of *Computer Science*

Memorial University of Newfoundland

October 2015

St. John's Newfoundland

Abstract

Survival analysis with high dimensional data deals with the prediction of patient survival based on their gene expression data and clinical data. A crucial task for the accuracy of survival analysis in this context is to select the features highly correlated with the patient's survival time. Since the information about class labels is hidden, existing feature selection methods in machine learning are not applicable. In contrast to classical statistical methods which address this issue with the Cox score, we propose to tackle this problem by discretizing the survival time of patients into a suitable number of subgroups via silhouettes clustering validity. To cope with patient's censoring, we use "k-nearest neighbor" based on clinical parameters that are truly associated with survival time. These are selected using penalized logistic regression and the penalized proportional hazards model with the EM algorithm. They are then used to estimate censored survival time. Next, the estimated class label is combined with feature selection to identify a list of genes that are correlated with the survival time and classifiers are applied to this subset of genes to determine which subtype is present in a future patient. By doing so, we expect that the identified subgroups are not only biologically meaningful but also differ in terms of survival. The effectiveness and efficiency of the proposed method are demonstrated through comparisons with classical statistical methods on real-world datasets and simulation datasets.

Acknowledgements

First and foremost, my deepest gratitude goes to my supervisor, Dr. Jian Tang, for the consistent and illuminating instruction he offered me throughout my masters program. He introduced me to this research and helped me navigate a path through it. This research would not have been possible without his enlightening advice, patience, and encouragement.

The Department of Computer Science, the Department of Mathematics and Statistics, and Memorial University have provided me with much appreciated financial, academic, and technical support. I thank Dr. Yildiz Yilmaz, Dr. Minglun Gong and Dr. Shuigeng Zhou for helping me pass through troubled waters in both my life and studies. I extend my gratitude to a number of people for their help and caring at various stages of this research, in particular, to Nolan White, Dwayne Hart, Donald Craig, Aaron Casey, Liuhua Zhang, Spencer Xu, Grace Chen, Edward O'Neill and Wenjie Chen.

Finally, my love and special thanks go to my beloved family for their unconditional support and encouragement through both the highs and lows of my academic pursuits, whether I was abroad or at home. I express my sincere gratitude to everyone I have had the pleasure of meeting and getting to know over the past three years as a result of this undertaking. Best wishes to all these people.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Cancer, the leading cause of death in Canada, accounted for 30% of all deaths in 2015 [12]. When a patient is diagnosed with cancer, a number of clinical parameters are used to assess the patient's survival profile. While a certain form of cancer, lung cancer for instance, is often thought of as a single disease, growing evidence suggests that there are multiple subtypes of a specific cancer disease that occur with clinically significant differences in survival [57]. One possible explanation is that two seemingly alike tumors are actually completely different diseases at the molecular profile of the tumor [33, 16, 58]. With the aim to ultimately improve the clinical management of cancer disease, researchers have sought to specify the subtypes of newly diagnosed patients, especially when those subtypes are associated with patient's survival time, or elicit different prognoses and responses to certain therapies.

After collecting the survival information of a group of cancer patients with the

same clinical diagnosis, the survival prognosis can be predicted by studying the patients' survival profiles. Figure 1.1 illustrates the survival information obtained from the application of the method proposed in this thesis to one dataset, in which all patients have the same clinical lung cancer diagnosis [5]. As can be seen in Figure 1.1a, patients with this type of cancer are at a high risk with the median survival time around 30 months, that is, only 50% of patients are expected to survive beyond 30 months. This type of fatal cancer must be treated aggressively, although aggressive treatments have potentially serious side effects. However, Figure 1.1b indicates that there exists another subtype of this cancer, which is distinguished by a difference at the molecular level of the tumor.

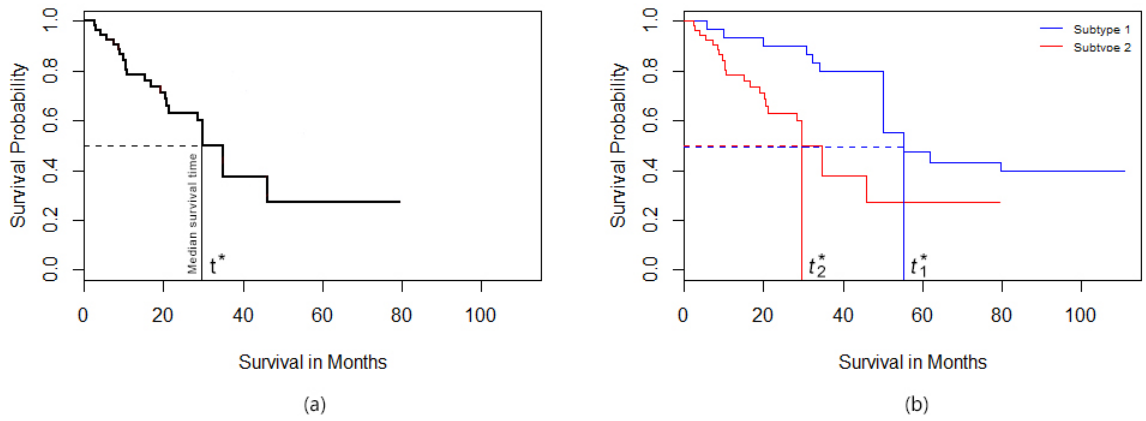

Figure 1.1: Kaplen-Meier survival plots for (a) one identified subgroup and (b) both stratified subgroups of the lung cancer dataset.

Compared to patients with the previous subtype, subjects with this alternative

subtype have a considerably improved long-term survival rates, with a median survival time of around 55 months. Patients with this less aggressive cancer can be treated with milder medications and still have excellent outcomes. In addition, patients diagnosed with the same type of cancer have different responses to different treatments; they may also respond differently to the same treatment [79, 56]. Therefore, a specific type of cancer is not a single disease, and there is an urgent need to identify the subtypes of cancers and diagnose patients accordingly.

Cancer subtypes provide clues into patient disparities with respect to survival time, and can help in designing more targeted treatment strategies and more effective therapies. In recent years, a number of methods have been proposed for diagnosing patients with a particular tumor subtype when the subtypes are already known [80]. When neither the subtypes nor the number of subtypes are known in advance, the issue of identifying such subtypes becomes much more complicated. The scientific community has been using clinical information to develop techniques to identify cancer subtypes in the diagnosis of future patients [47, 22, 11, 6]. However, this problem still remains a largely open research question and further research is required [76].

To date, all efforts towards solving this problem fall into a class of statistical procedures [4, 3, 5], and there is rarely study on identifying subtypes associated with survival using the techniques proposed by computer scientists in the machine learning community. Statistical procedures have achieved varying degrees of success. However, their effectiveness diminishes in coping with high-dimensional data as selecting relevant features irrespective of redundancy jeopardizes generalization capability [58]. Moreover, selecting relevant features by iteratively fitting a univariate Cox

proportional hazards model is time-consuming, especially in high-dimensional setting.

The aforementioned research sparks our interest in exploring the feasibility of applying the machine learning approach, in the context of survival data, to identify subtypes of cancer and to use this knowledge to diagnose future patients. Compared with the methods from statistical perspective, feature selection in machine learning possesses not only high relevancy, but also low redundancy with respect to the phenotype of interest. Furthermore, most of them are model free, which allows for wide applicability and easy implementation. This thesis describes a new approach that utilizes both gene expression data and clinical data to conduct feature selection and survival prediction from the machine learning perspective. The main goal is to open a range of possibilities for future work on designing a more powerful tool for diagnosing and treating cancer from a different class of techniques.

1.2 Research Questions

The main objective of this research is to address issues surrounding the prediction of patient survival from high-dimensional survival data. Since this research lies beyond the traditional research paradigm that falls into the class of statistical procedures, it leads to some fundamental research questions related to the value of the proposed methodology:

How can machine learning approaches be applied to predict patient survival from high-dimensional survival data?

The value of diagnosing and prognosing the subtypes of different cancers is well-established [69, 8, 23], especially when neither the subtypes nor the number of

subtypes are known [47, 25, 81]. There is an extensive body of literature on survival analysis with high-dimensional data for selecting significant features, identifying cancer subtypes, and predicting future observations [43, 22, 11]. To the best of my knowledge, all reported methods fall into the class of statistical category.

The foundation of these methods in selecting significant features but also evaluating future patients is the Cox score derived from the well-known and widely used Cox proportional hazards model [74]. The Cox score quantifies how well a feature predicts survival by fitting a univariate Cox proportional hazards model for each individual feature, regardless of the availability of class label. Using the Cox score as a basis, statisticians employ a variety of methods to identify significant features that are likely to be associated with survival and all these methods exclusively involve computing Cox score or the variants of Cox score[76, 43]. Although feature selection in machine learning is model-free and highly efficient for high-dimensional data, there is a limited number of related studies on high-dimensional survival data. One obstacle is that feature selection in machine learning necessitates the dependence on class label. Therefore, finding a way to determine the hidden class label is necessary. This research question will be answered in Chapter 3.

What is the feasibility of using the machine learning approach for high-dimensional survival data?

In this study, we developed procedures to select significant features and to stratify newly diagnosed patients into identified subtypes from the machine learning perspective. As such, this work can be classified as an empirical study in survival analysis for high-dimensional data. In high-dimensional survival data studies, the central aim is to develop tools to diagnose different subtypes of cancer, including

subtypes that are already known to exist and those that are unknown, in order to improve the clinical management of cancer disease [80, 8, 1]. Therefore, one of the research questions of this thesis is to explore the feasibility of the proposed approach on the measures of efficiency and effectiveness. In this research, "efficiency" refers to the time requires for feature selection methods to identify a list of significant genes, whereas "effectiveness" gauges the significance of the selected genes and the quality of the survival prediction of future patients. This research question will be addressed by an empirical study described in Chapter 4.

1.3 Approach

The task of identifying cancer subtypes involves the discovery or identification of survival classes or meaningful groups of objects that hold vital implications for survival time. Technically speaking, the objectives are to identify a set of latent class memberships that are associated with the phenotype of interest. Within the identification process, the initial step of many existing methods is to subset the feature space into a relevant subset of features. In this research, a diagnostic procedure that makes use of both the subsetted gene expression data and the clinical data of previous patients was employed to predict survival of future patients.

The first aspect we considered was selecting an informative subset of features from an existing feature space. However, unlike statistical variable selection method, since the information about class label is hidden, the existing feature selection in machine learning is not applicable. We tackled this problem by discretizing the survival time of patients into a suitable number of subgroups via silhouettes clustering

validity [66]. The second step was to employ machine learning classifiers to diagnose future patients [61]. Although several different subtypes of a certain form of cancer might exist, if the prognosis for all patients is the same regardless of patient survival, the subgroups predicted in future patients may not differ in terms of survival. We therefore used class labels relied on clinical data along with feature selection, to identify a list of genes that were significantly associated with survival. Classifiers were applied to this subset of genes to predict the subtype for a future group of patients. Finally, empirical evaluations were conducted on both publicly available datasets and simulation datasets in order to explore the efficiency and effectiveness of the specific design choices in comparison to current statistical approaches. Overall, a practical research methodology was employed to predict the survival of future patients. This methodology employs techniques mainly from the machine learning perspective and is in contrast to approaches that fall into the class of statistical category.

1.4 Organization of the Thesis

The remainder of this thesis is organized as follows:

In Chapter 2, the basic concepts of survival analysis will be presented. This includes descriptions of existing methodologies and related techniques used in this thesis, as well as an overview of statistical methods for the analysis of high-dimensional survival data in the literature.

In Chapter 3, the proposed approach for analyzing high-dimensional survival data from the machine learning perspective is described. Selecting significant features and predicting identified subtypes in the future patients with the help of machine learning

approach are discussed, respectively.

In Chapter 4, the results of a real-world data analysis and a simulation study using our approach are discussed in detail.

The thesis concludes in Chapter 5 with a summary of the study's contributions, and suggestions for future research.

Chapter 2

Survival Analysis with High-Dimensional Data

This chapter provides an introduction to some concepts essential to survival analysis, including censored data, survival functions, and models used to estimate survival processes. A review of statistical methods for the analysis of high-dimensional survival data and related techniques follows.

2.1 Survival Analysis

2.1.1 Survival Data

The main outcome addressed in survival analysis is the time at which an event of interest occurs. Events of interest can refer to any experience had by an individual: their response to treatment, their recurrence of or recovery from a disease, or any designated experience of interest that occurs in an individual. The span of the event

of interest, starting from the beginning of the follow-up of an individual in the study until an event occurs, is called the survival time. Survival time can be considered as either a negative or positive experience. For instance, the duration spent in remission following surgical removal of a primary tumor is a negative event, while returning home after a stay in the hospital due to an infection following cardiac surgery is regarded as positive.

The purpose of analyzing survival data is to determine the proportion of cured patients, define the possible effects of covariates, and predict future survival [43]. If the event of interest occurs in all individuals of the study cohort, usually referred to as uncured or susceptible, survival analysis does not give advantage over other methods of analysis. The benefits of survival analysis occur when only a portion of individuals have experienced the event by the end of the study period. This results in access to only a subset of survival information from the study cohort, a phenomenon called censoring.

Figure 2.1 illustrates the concepts of censoring. Suppose we are interested in the time spent in remission by a group of cancer patients, following surgical removal of a primary tumor. In this example, the event of interest is relapse from remission, and the survival time is from the beginning of the patient's follow-up until relapse. Censoring occurs in this example because there are insufficient data to determine the exact duration of remission. For example, Patients 2, 3, 5 and 6 entered the study at different times and their survival times vary; however, their survival times are not censored simply because they all relapsed from remission before the end of the study. As long as the participant experiences the event of interest before the end of the study, he/she belongs to the category not censored. On the other hand, Patient 1 withdrew

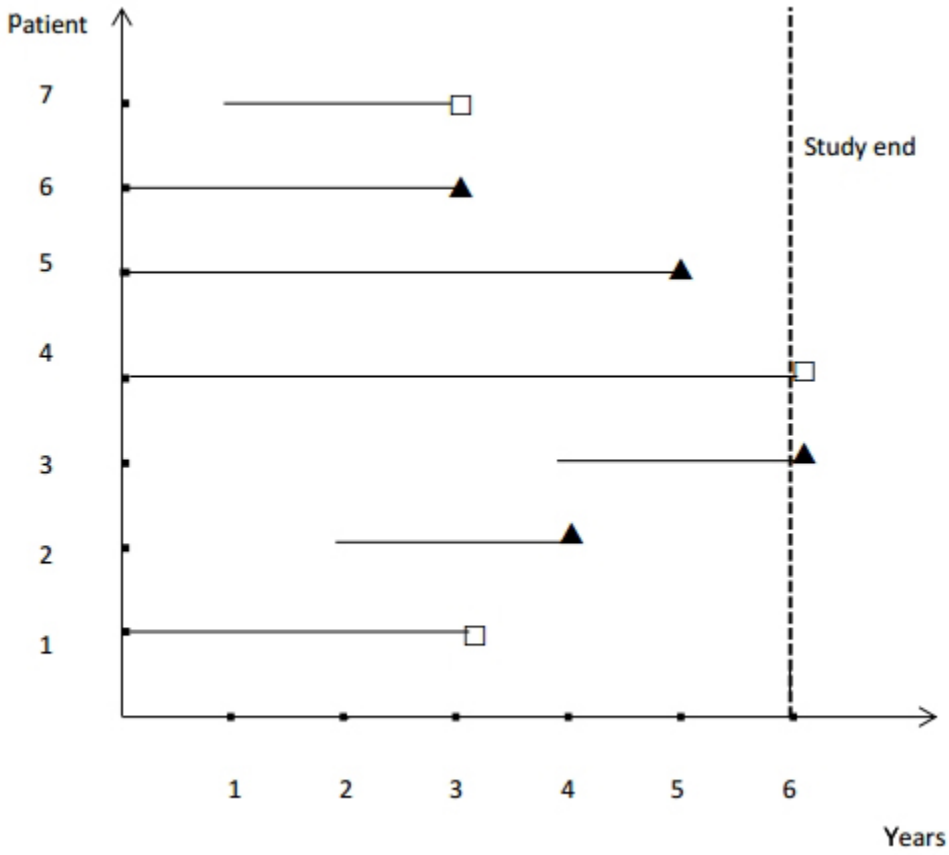

Figure 2.1: Dashed vertical line is the end of study period, ▲ = relapse from remission, □ = censored.

from the study due to an event that was not of interest, making further follow-up impossible; Patient 4 did not experience the relevant outcome by the end of the study; and Patient 7 was lost to follow-up during the study period. Therefore, their exact survival times in the study period are unknown, and as such they are all censored, the censored survival times for Patient 1, 4 and 7 are 3 years, at least 6 years, and 2 years, respectively. These three examples serve to illustrate how censoring usually occurs.

There are three types of censored event: right, left and interval censored. Right censored events are those that continue beyond the end of the follow-up period. This case applies to Patient 4, who entered the study at the beginning and was monitored until the sixth year (i.e., the end of the study) without the event occurring. Patient 4's survival time is at least six years. Suppose, however, that Patient 4 entered the study sometime after surgery and experienced relapse from remission before the end of the study. This is a left censored event, which means that the actual survival time is less than or equal to the observed time of recurrence. Alternatively, assume several years after the study ends we are interested in the study again. If Patient 4 experienced the event of interest within the time interval or was lost to the follow-up, Patient 4 is then considered to be interval censored. This means the true time is within a known time interval [38]. Although left and interval censoring do exist, most survival data is right censored. We refer to only right censored data in the discussion to follow. In addition, independent censoring, random censoring, and non-informative censoring are three assumptions taken into account in the analysis of survival data, that is, the event and censoring time for each patient are independent and the reason for censoring is not specified [43].

2.1.2 Basic Tools for Survival Analysis and Related Techniques

Throughout this thesis, we often use the survival analysis terms Kaplan-Meier, Cox proportional hazards model, Mixture cure model and so forth interchangeably. We now turn our attention to describing these basic tools for survival analysis. Kalbfleish and Prentice give an excellent review on the subject and their many applications [39]. Related techniques used in the current work are also presented in the section.

2.1.2.1 Terminology and Notation

Three different variables T, t and δ are widely used in survival data. Each individual's survival time is a random variable commonly defined as T, while t denotes any specific value of the random variable T, that is, any value of the event of interest. Since T denotes time, its value is always greater than or equal to zero. This is the case for t as well. The binary random variable δ is assigned 1 or 0 to indicate the absence or presence of censorship, respectively. When a person has not experienced the event by the end of the study, has been lost to follow-up or withdraws from the study, their survival time is censored, and δ equals 0. In the absence of censorship, the patient experiences the event before the end of the study, and δ equals 1 [43].

Survival data is usually considered and modeled in terms of two quantitative terms: the survivor function and the hazard function, denoted by *S(t)* and *h(t)*, respectively [64]. The former gives the probability that an individual will survive from the start time to a specific future time, t. The survivor function is essential to survival analysis because obtaining survival probabilities for different values of t

provides a crucial summary of information from the time patients enter the study to the event of interest. Depending on these values, we can gain a general view of the study cohort's survival experience. Theoretically, the value of *S(t)* decreases from 1 to 0 when t increases from 0 to infinity. That is becasue at the beginning of the follow-up (i.e., $t = 0$), no one has experienced the event yet, so $S(0) = 1$. When t approaches infinity, that is, when the study period increases without limitation, every study participant would theoretically experience the event of interest, so $S(\infty) = 0$. When dealing with actual survival data, however, the study period is finite in length. Furthermore, an individual may withdraw from the study or may be lost to follow-up during the study period, so not everyone will necessarily experience the event. In other words, the value of *S(t)* may be greater than zero rather than decreasing all the way to zero at the end of the study. Figure 2.2 shows a typical graph of a survival function.

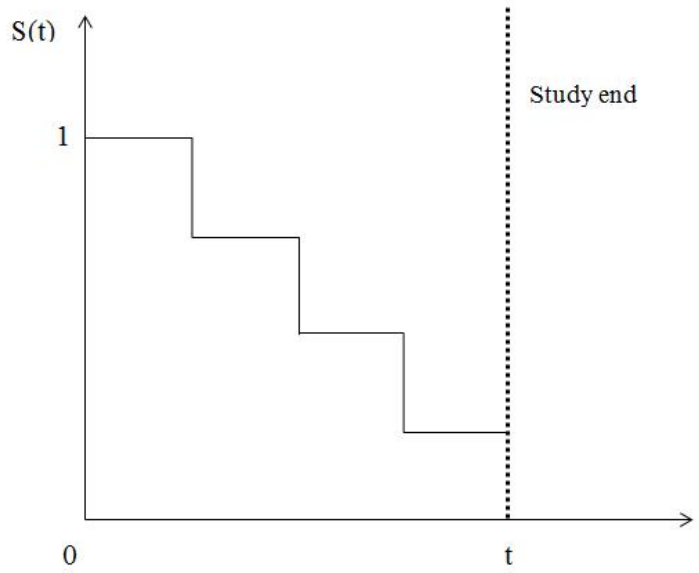

Figure 2.2: An illustration of the survival function $S(t)$

In contrast to the survival function, the hazard function investigates the incident event rate, which is the instantaneous potential per unit time which happens an event under observation at a specific time, t. The value of the hazard function can range from 0 to infinity, and can either be increasing, decreasing or remain constant (i.e., stability is maintained throughout the study). These three types of survival functions are depicted in Figure 2.3, respectively, in lines a, b and c. Line a might be expected for patients not responding to treatment and patients recovering from surgery can be represented by line b. Line c occurs when a person continues to be healthy throughout the study period [59].

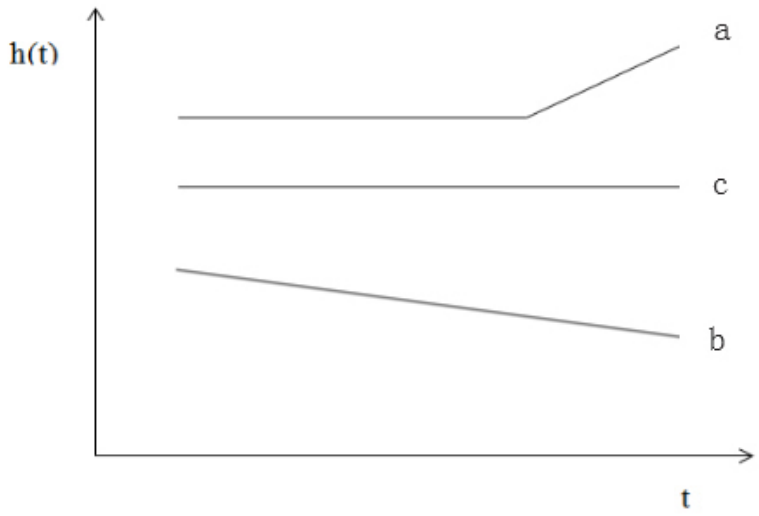

Figure 2.3: Graphs depicting $h(t)$ as (a) an increase function, (b) a decrease function, and (c) a constant function.

There is a clearly defined relationship between the survival functions and hazard function. Knowing the formula for *S(t)*, the corresponding formula for *h(t)* can be determined, and vice versa. The general formula is expressed as follows:

$$\mathrm{S}\left(\mathrm{t}\right) = \exp\left[-\int_{0}^{t} h\left(u\right) du\right] \tag{2.1}$$

$$\mathrm{h}\left(\mathrm{t}\right) = -\left[\frac{dS\left(t\right)/dt}{S\left(t\right)}\right] \tag{2.2}$$

In survival analysis, the Kaplan-Meier (KM) model is frequently used to plot and interpret survival data. After we generate KM curves, the log-rank test can be conducted to help us assess the information revealed by the KM curves and test their equivalency. Other popular test methods include the Cox proportional hazards model and mixture cure model, to name a few. We will discuss the background and use of each of these methods in the following sections.

2.1.2.2 Kaplan-Meier Model

The Kaplan-Meier (KM) model is a nonparametric test for estimating the survival probability from survival data [40, 59]. It is typically used to estimate the proportion of a cohort, both censored and uncensored, who survive from the start time to a specified future time *t*. Equation 2.3 is the general formula for the Kaplan-Meier model:

$$\mathrm{S}\left(t_j\right) = \mathrm{S}\left(t_{j-1}\right) \times P_r(\mathrm{T} > t_f \mid \mathrm{T} \geq t_f) = \mathrm{S}\left(t_{j-1}\right)\left(1 - \frac{d_i}{n_j}\right) \tag{2.3}$$

Recall that *S(t)* gives the probability that a patient survives from the time of origin to a specific time, t. Suppose that N individuals have the following survival times in non-descending order: $t_1 \leq t_2 \leq t_3 \leq ... \leq t_{n-1} \leq t_n$. Equation 3 gives the probability of surviving over the interval t_{j-1} *to* t_j. A formal way to express the formula is that $S(t_j)$, the survival probability of being alive at t_j, is the multiplication of $S(t_{j-1})$, those who survive at the last previous time, and $(1 - \frac{d_i}{n_j})$, those who survive over the interval between them, where n_j is the number of individuals alive just before t_j, and d_j is the number of events at t_j. A decrease in the number n_j during the interval may result from either the subject experiencing the event of interest or the subject being censored within the interval period. Censored individuals need to be taken into account in the total number of individuals available at t_j.

Figure 2.4 shows as an example of the KM step function plot for a set of survival data that we generated from simulating the distribution of the lung cancer dataset [5]. This generated dataset includes each patient's curing and censoring status as well as their corresponding failure time. By selecting the appropriate parameter values, the data was generated with a higher cure rate to better illustrate survival data with a cured proportion. A univariate was generated with binary values 1 and 0 to represent two different groups. This is relevant, for instance, when comparing treatment versus placebo in a clinical trial. The details describing data generation in this thesis, including simulation data used for testing the later proposed method, will be covered in section 4.2.1.

As we can see from Figure 2.4, the estimated relapse-free survival curve from the "Feature1 Equals 1" group is always above or at least at the same level as the

"Feature1 Equals 0" group. This indicates that the survival probability of patients from the former group is higher than that of the latter group, that is, the former group experiences more effective treatment. Figure 2.4 also indicates that both curves level off at a value substantially greater than 0 after a period of follow-up, which means that some patients will not experience the recurrence of their disease after undergoing treatment. Therefore, an innegligible portion of the population in both groups may never experience the event of interest.

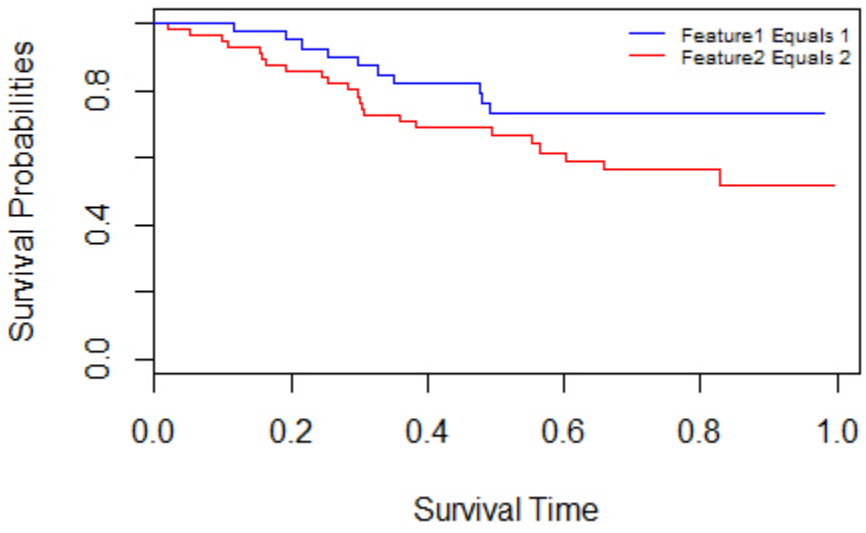

Figure 2.4: KM plot for survival data with cured patients

Alternatively, if we generated data with patients who experience the event of interest only, the KM plot would appear as shown in Figure 2.5. Although the estimated relapse-free survival curve from the "Feature1 Equals 1" group is always

at the same level or above the "Feature1 Equals 0" group, both curves level off at 0, which means both groups experience recurrence of their disease in the end.

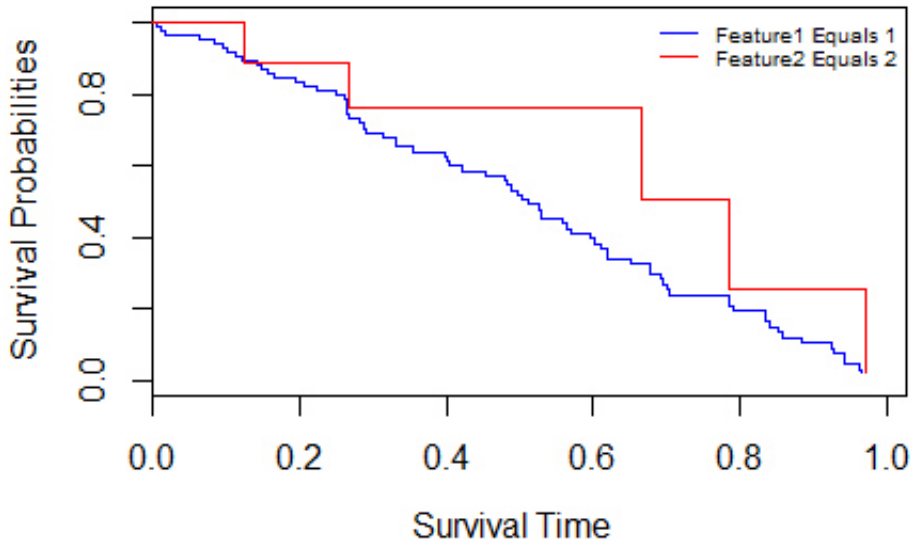

Figure 2.5: KM plot for survival data without cured patients

The KM plot depicted in Figure 2.4 and 2.5 represent two common scenarios. The Kaplan-Meier method graphs survival curves for different patient groups, and the width of the gap between groups suggests the degree of visual difference or inequivalence. To quantify the difference, the log-rank test can be applied to statistically calculate the overall population survival difference [2].

2.1.2.3 The Cox Proportional Hazards Model

A number of statistical methods have been proposed for modeling survival analysis data. Among them, the Cox proportional hazard model and mixture cure model are widely used. In this section, we will describe the Cox proportional hazards model, abbreviated to Cox model or PH model, and demonstrate its application [20]. The mixture cure model is discussed in next section. The Cox PH model is usually written in the form given in Equation 2.4:

$$\mathrm{h}\left(t, x\right) = h_0\left(t\right) \times e^{\sum_{i=1}^{p} \beta_i x_i} \tag{2.4}$$

where $x = (x_1, x_2, \ldots, x_p)$

The formula describes the instantaneous event rate at a given time t, where t is determined by the hazard function and specification of a set of covariates denoted by X. The term $h_0\left(t\right)$ is called the baseline hazard function and occurs when all the x_i values are equal to zero. The exponential part of the formula only contains the time-independent variable X, which does not change in value as the time varies. One reasonable way to estimate the regression coefficients β is to apply the partial likelihood estimation [17]. The function can be written as follows:

$$L\left(\beta\right) = \prod_{r \in D} \frac{\exp\left(\beta^T X^r\right)}{\sum_{i \in R_r} \exp\left(\beta^T X^i\right)} \tag{2.5}$$

In Equation 2.5, D is the set of indices of the failures, R_r is the set of indices of the individuals at risk at time $t_r - 0$ including the censoring individuals, and r is the

index of the failure at time t_r. In order to estimate β, we first fit the Cox PH model then maximize the partial likelihood function. It is worth noting that this approach cannot be used directly to estimate β with respect to the small n (the number of observations) large p (the number of covariates) problem, that is, when the number of covariates exceeds the number of observations [4, 13]. This setting is referred to as "high-dimensional".

2.1.2.4 The Mixture Cure Model

The unstated assumption behind the Cox Proportional Hazard model is that all patients will eventually experience the event of interest, given a sufficient follow-up time. Remarkable advances in medicine, however, have increasingly made lethal diseases curable. Also, long-term survivors are statistically regarded as cured when an innegligible proportion of patients will never experience disease recurrence (i.e., an estimated Kaplan-Meier survival curve will reach a plateau after a certain time) [46]. In the case where a large percentage of the cohort does not experience recurrence, standard survival models fail to provide a good understanding of the survival process [30].

In contrast to standard survival models, the mixture cure model, first introduced by W. Boag [9], accounts for the possibility that some patients will be free of recurrence. This model is often selected when standard survival models are inadequate, and it provides insight into factors that affect susceptibility and recurrence. The mixture cure model can be expressed as follows:

$$\mathrm{S\,(t|X,\ Z)} = \mathrm{Pr\,(T > t\ |}x,\ z) = \ \mathrm{P\,(z)}\ \ + (1 - \mathrm{P\,(z)})\,\mathrm{S\,(t|x)} \tag{2.6}$$

where $P(z)$ models the proportion of non-susceptible patients, conditional upon z (i.e., the probability of being cured) and is usually referred as the "incidence"; and $\mathrm{S}(\mathrm{t}|\mathrm{x})$ denotes the survival probability of the uncured patients given x and is referred to as "latency". Note that x and z are the observed values of two covariate vectors that affect the survival function. Although we use different covariate notations, identical covariates are allowed for these two components.

The standard formula of the mixture cure model in Equation 2.6 can be regarded as the combination of the cure-rate function and the survival function. The left side of the equation (the incidence segment) models the probability of the patients being cured and the remainder (the latency segment) is the survival distribution of uncured patients. These two segments help us differentiate the study cohort as individuals who will remain free of disease in the long term and those who are out of remission within the study period.

Various models for the incidence segment have been proposed, such as the logistic regression model, log-log model and probit model. Similar to the proportional hazard model, the latency segment can be derived from the Weibull distribution and the Exponential distribution, among others [43]. The Weibull model in particular has been found to provide a good description of many types of lifetime data and is widely used in biomedical applications [48]. In addition, the logit link function is typically used to model the effect of z. The logistic regression model and Weibull model are formulated as follows, where $\lambda_i = e^{\beta x}$:

$$P(Z) = \frac{exp\left(\sum_{i=1}^{p} \gamma_i z_i\right)}{1 + exp\left(\sum_{i=1}^{p} \gamma_i z_i\right)} \tag{2.7}$$

$$\mathrm{S}\left(t_i|x_i\right) = \exp\left(-\lambda_i t_i^{\beta}\right) \tag{2.8}$$

2.1.2.5 The Lasso Method for Variable Selection

When fitting models, the "best" subset of variables has to be determined. There are many different variable selection methods and these fall into two broad categories: (1) the discrete method, for example the stepwise selection [15]; and (2) the continuous method, such as the shrinkage-based method [82]. This research employs the shrinkage-based method that uses L_1 penalty, that is, the Least Absolute Shrinkage and Selection Operator (Lasso) [73], to select important factors for the mixture cure model considering its stable and efficient merits.

Consider the typical survival data setup: we have data (y_i, x_i, δ_i), $i = 1, \ldots, n$, where x_i represents the vector of predictors and y_i is the observed survival time. The event is complete if $\delta_i = 1$ and is right censored if $\delta_i = 0$. Take Equation 2.5 as an example and denote the log partial likelihood by $\mathrm{l}\left(\beta\right) = logL\left(\beta\right)$. Lasso is used to solve the original problem by adding a constraint as the penalty,

$$argmin\, l\left(\beta\right),\ subject\ to\ \sum |\beta_i| \leq t \tag{2.9}$$

or equivalently to solve the problem,

$$argmin\ \left\{\ l\left(\beta\right)\ +\ n\lambda \sum |\beta_i|\right\} \tag{2.10}$$

where t, or alternatively λ, is a tuning parameter. Note if $\sum |\beta_i| > t$, the solutions to Equation 2.5 are the usual partial likelihood estimates. If $\sum |\beta_i| \leq t$, however, some coefficients shrink to 0. The strategy for solving Equation 2.9 is to express the usual Newton-Raphson update as an iterative reweighted estimation step, and then replace the weighted estimation step by a constrained weighted estimation procedure [36]. The procedure is outlined as follows:

1. Start with $\mathrm{E} = \{\mathrm{j}_0\}$ where $\delta_{\mathrm{j}_0} = \text{sign}\left(\hat{\beta}^0\right)$, and $\hat{\beta}^0$ is the ordinary estimate.
2. Find $\hat{\beta}$ to minimize the $l(\beta)$ subject to $G_E\delta \leq t \cdot 1$.
3. If $\sum |\beta_i| \leq t$, then stop; otherwise proceed to step 4.
4. Add $\delta_j = \text{sign}\left(\hat{\beta}\right)$ to G_E, that is, let $G_E = \begin{pmatrix} \delta_j^T \\ \delta_i^T \end{pmatrix}$, and return to step 2.

Note that if an unconstrained minimization was instead used in step 2, this procedure would be identical to the usual Newton-Raphson algorithm for maximizing the partial likelihood [17].

In terms of the tuning parameter t, data-driven methods, such as generalized cross validation (GCV) can be used to determine the best t value [29]. The GCV statistic is

$$GCV(t) = \frac{1}{N} \frac{-l_t}{N\left[1 - p(t)/N\right]^2} \tag{2.11}$$

where $p(t)$ is the effective parameters and l_t is the log-partial likelihood for the constrained fit with constraint t. Intuitively, the GCV criterion inflates the negative log partial likelihood by a factor that involves $p(t)$. As Chapter 3 will reveal, Lasso is used as the penalty function to generate the new penalized likelihood function as the clinical parameter selection.

2.1.2.6 Discretizing Continuous Features

After the discussion of basic survival analysis tools, a fundamental technique, discretization, of this work is covered in this section. Discretization is usually performed to search for the width, or the boundaries of the arity of intervals given the range of values of a continuous attribute. In doing so, a set of landmarks that partition the range of values are identified. A typical discretization process is shown in Figure 2.6. It usually starts with optionally sorting data in ascending or descending order with respect to the continuous values of the variable to be discretized. After the sorting step, landmarks are chosen among the whole dataset to either divide the range into intervals or merge adjacent intervals according to some evaluation function equipped with a stopping criterion; the evaluation function measures class coherence, and the iteration process is terminated when the number of inconsistencies or the misfit measure is deemed to be below a given tolerance. The stopping criterion involves a trade-off: lesser artity gives a better understanding but lower accuracy, and more arity is accompanied by a poorer understanding but higher accuracy. Researchers in the machine learning community have introduced numerous evaluation functions, and an overview of discretization algorithms can be found in [26].

Previous work on continuous feature discretization can be categorized into

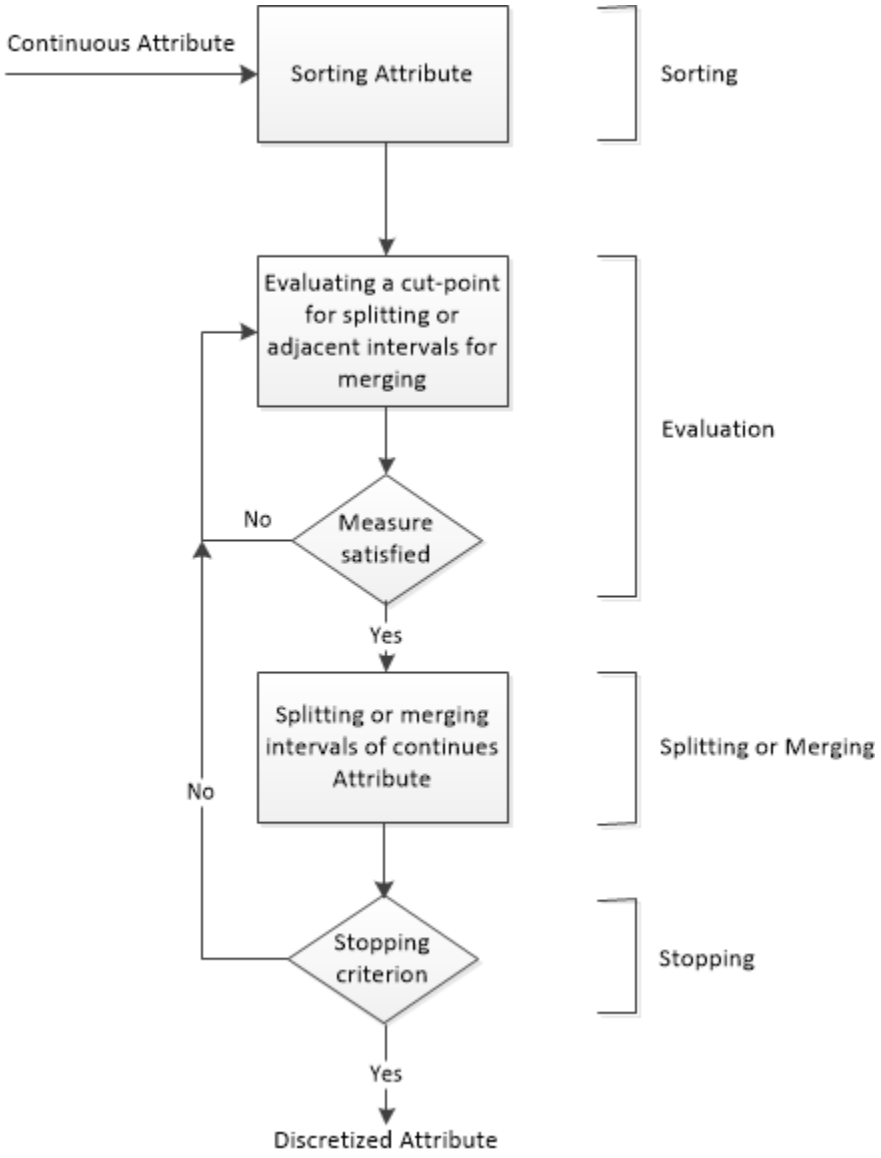

Figure 2.6: The typical discretization process

unsupervised discretization methods, such as equal width interval binning (EW), or supervised discretization methods such as ChiSplit, which maximizes the chi-square criterion applied to the two sub-intervals adjacent to the splitting point [45]. The difference between the two types of methods is the presence of a class label in the discretization process. Alternatively, discretization methods, namely global or local, can be identified by different axes. Binning, for example, is a global discretization method that produces a mesh over the entire continuous instance space. Local discretization methods perform the partition at localized regions of the instance, as exemplified by C4.5 [49]. Other dimensions of discretization methods are direct or incremental, static or dynamic, and top-down or bottom up [35]. Table 2.1 provides a two-dimensional summary of some representative discretization methods classified by the global/local and supervised/unsupervised axes.

Table 2.1: Summary of discretization methods

	Supervised	Unsupervised
Global	1RD(Holte 1993) ChiMerge(Kerber 1992) Supervised Monothetic Contrast Criterions D-2(Catlett) Adaptive Quantizers StatDisc(Richeldi & Rossotto) Global Entropy Minimisation(Ting)	Equal width interval binning Equal frequency interval binning Supervised Monothetic Contrast Criterions
Local	C4.5(Quinlan 1993) Vector Quantization(Kohonen) ConMerge Entropy based hierarchical method Entropy Minimisation (Fayyad & Irani)	K-means clustering

We now describe in detail two unsupervised discretization methods (equal width interval binning and k-means clustering), and two supervised discretization methods

(Holte's 1R discretizer and recursive minimal Entropy Partition). The simplest discretization method, equal width interval binning (EW), determines the minimum and maximum values of the continuous attribute, and then divides the range of observed values into k bins of equal width discrete intervals, where k is the user-defined parameter. Equal width interval binning has been widely applied as a means of producing nominal values from continuous features. The process is irrespective of instance class information, and each feature is partitioned into a sub-range independent of the other attributes. Thus, EW is an unsupervised as well as a global discretization method. The obvious weakness of the EW method is that it is vulnerable to the distribution of attributes with heavy-tailed or outliers [14].

The majority of systems using unsupervised methods carry out global discretization with the exception of another common discretization technique: k-means clustering (KM) developed by MacQueen [52]. It produces intervals that are applied to sub-partitions of the instance space. KM is a non-hierarchical procedure that partitions observations into k non-overlapping bins using an algorithm that produces groups of objects with a high degree of similarity within each group and a low degree of similarity between groups. A more robust variant application of k-means clustering is k-medoids clustering, which uses the most centrally located objects as the representative of a cluster instead of shifting centroids according to the computing of the mean of each cluster [41]. The most common realization of k-medoid clustering is the Partitioning Around Medoids (PAM) algorithm [63]. It starts by arbitrarily selecting k objects as the medoids and associating each remaining object to the closest medoid, then iteratively swapping each medoid and non-medoid if the recomputed configuration decreases until a pre-defined criterion is met. We

employ this probabilistic naturalized method to discretize the continuous phenotype of interest in clinical data. How to find the optimal number for k to perform clustering will be covered in section 3.2.

Holte describes an error-based supervised discretization method using error counts to refine the partition breakpoint estimation of each bin [37]. This method, referred to here as 1RD, sorts the attribute into ascending order and greedily divides the feature into bins, each containing only one instance of a particular class. The inherent danger of such a scheme is that it may lead to one bin for each instance. To circumvent this problem, the algorithm is constrained to formed bins, excluding the upper most bin, with a minimum number of instances of a particular class. Any specific bin, therefore, can comprise more than one class label and boundaries will not be continually divided, leading to overfitting. The partition moves to the right to add an observed value to a particular bin until it contains at least six instances of a class label, and it continues until the instance to be considered is not part of the majority class label. Empirical analysis of 1RD on a number of classification tasks suggests that a minimum bin size of six performs the best [26].

The last discretization method we will consider is based on the algorithm of Fayyad and Irani [31], which uses entropy measures to evaluate candidate splitting points with the Minimum Description Length (MDL) as the stopping rule. This supervised algorithm utilizes the class information entropy of candidate partitions to select bin boundaries for discretization. Class information entropy is a measurement of the unpredictability of information content, and it measures the amount of information needed to identify which classes of a set that a particular instance belongs [10]. The concept of entropy will be covered in section 3.3 with regards to feature selection.

Information entropy considers one large interval containing all known values of a feature and selects a binary discretization boundary by finding a single boundary that minimizes the entropy function over all possible boundaries. This entropy function is then applied recursively to split both of the partitions into smaller subintervals until the stopping criterion MDL is achieved, thus leading to a discretization of the continuous attribute into multiple intervals. This method is performed as a step of data preprocessing process (see section 4.1.2) to discretize the value of each gene expression profiling into discrete intervals.

2.2 Current Statistical Methods for the Analysis of High-Dimensional Survival Data

Over the past decade, the revolution in biomedical technologies has changed the face of biomedical research [79, 8, 23, 67]. Biologists are now able to conduct more experiments at the same time in less time-consuming ways, which has resulted in the increasing availability of genetic data, a better understanding of the biological mechanics of diseases such as cancer, and opportunities for secondary uses of heath information.

As a result of this breakthrough in modern genomics technology, many studies on survival have emerged [79, 47, 23]. In the context of survival data, the number of features (p), for instance gene expression profiling, significantly exceeds the number of observations (n), namely patients. This type of data requires considerable amendment in order to apply the aforementioned classical statistical methods [4, 13]. In addition

to biological data, clinical data, including survival outcomes of the same observations, are also available. Collett provides a good overview and a variety of examples on survival data [19]. Many studies have resulted in the successful identification of previously unknown subtypes of cancer as well as stratifying newly diagnosed patients into subtypes based on short- or long-term prognoses and predicting survival time [47, 22, 11, 6].

A number of existing approaches in the literature are applicable for distinguishing patients into subtypes, which are associated with patient survival time and response to treatment, based on survival data in high-dimensional setting are presented below. These methods fall into three broad categories: unsupervised learning techniques that consider only microarray data; supervised learning approaches, which are exclusively based on clinical data; and, more recently, semi-supervised procedures that take both microarray and clinical data into account for the determination of cancer subtypes.

2.2.1 Terms and Notation

High-dimensional data is complicated by a number of factors, such as latent class label structure and the small n large p problem for microarray data concerning the same cancer diagnosis [27]. In addition to the terminology presented thus far, several other useful terms and notations can be defined. Let $\boldsymbol{X}$ denote an $n \times p$ matrix on a sample of n observations with p features each. Each observation $X^i \in R^p$ is a $p \times 1$ vector of features, and the associated survival information (i.e., survival time $s_i = (t_i, \delta_i)$, tumor size, patient gender, etc.) is also accessible. The survival time, t_i, is the time from disease diagnosis to the last follow-up. If the observation is

complete, $\delta_i = 1$, and if it is censored, $\delta_i = 0$; that is, $\delta_i = 1$ if observation i failed at time t_i, and if $\delta_i = 0$, the observation i survived to at least time t_i or was lost to follow-up. In the context of the current survival analysis, subgroup C_i can be defined as the subtypes associated with patient survival time. The objective of these approaches therefore depends on the available data $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{S})$ in order to identify the subgroup membership, $\boldsymbol{C}$, which is both biologically meaningful and correlated with clinical outcome. In contrast to the unsupervised learning technique and supervised learning approach based on either $\boldsymbol{X}$ or $\boldsymbol{S}$ to guide the identification of subgroups, a semi-supervised procedure takes advantage of both $\boldsymbol{X}$ and $\boldsymbol{S}$ to determine subgroup discovery.

2.2.2 Unsupervised Approach

Generally speaking, the problem with unsupervised learning lies in attempting to identify the latent structure in unlabeled data. With regards to the analysis of high-dimensional survival data, the objective is to unseal the concealed gene expression profile structure that define cancer subtypes, which differ at the molecular level and are associated with survival outcome. This knowledge is then used to diagnose a future group of patients in terms of their subgroups. Survival information is not taken into account during the determination of subgroup membership (i.e., the subgroups are identified using only the gene expression data). Once subgroup membership, $\boldsymbol{C}$, has been determined according to the training dataset, X_0, a classifier can be trained according to the assigned subgroups $C_1, C_2, ..., C_n$ in combination with testing data set, X_1, to predict the subgroup membership of a future group of patients. Figure 2.7

depicts the general procedures of unsupervised learning techniques.

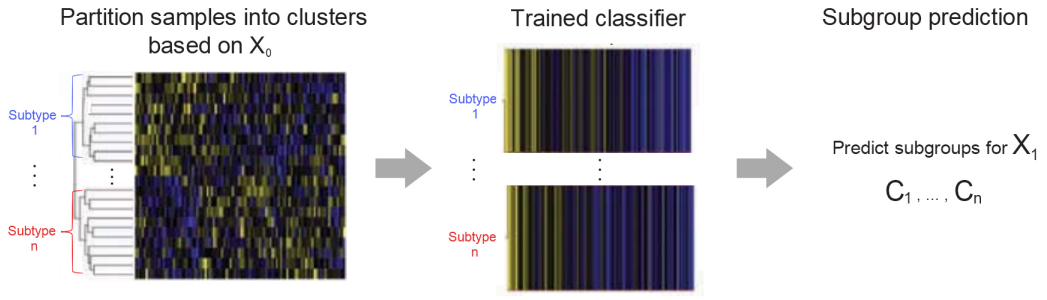

Figure 2.7: The general steps that constitute unsupervised learning approach

Approaches to unsupervised learning include hierarchical clustering [28], k-means clustering [72], and model based clustering [32] to name a few. For an overview of unsupervised learning techniques, see Clifford et al [18] or Gordon [34]. Of these techniques, hierarchical clustering has been used to identify cancer subtypes associated with survival outcome in a number of different studies [25].

The basic principle behind hierarchical clustering is to use a metric of similarity between each individual observation to group observations into different clusters. The measure of similarity is based on all or a selected subset of the features chosen independent of the phenotype of interest. Thus, objects in the same cluster are similar, and they are dissimilar to different clusters. A clustering dendrogram is then used to define subtypes of patients. Figure 2.8 illustrates the clustering dendrogram for a publicly available renal cell dataset [85], which is also used to test our proposed approach in Chapter 4. The p-value for the log-rank test is 0.0479, which is predictive of survival (at $p = 0.05$). However, there is no guarantee that the subtypes this

approach obtains will always be correlated with the clinical outcome because survival information is ignored when determining the subgroup membership.

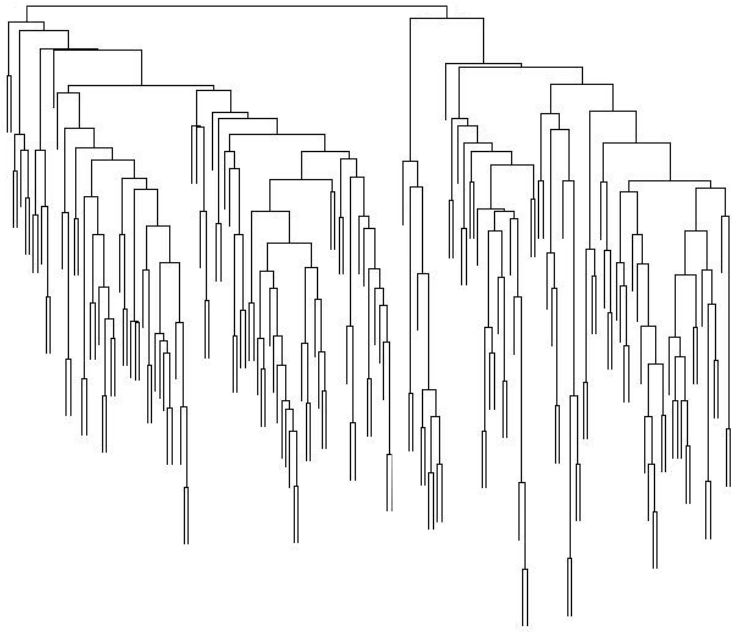

Figure 2.8: Hierarchical clustering dendrogram of renal cell data

2.2.3 Supervised Approach

The second approach for identifying cancer subtypes that differ in survival progress is supervised learning. Supervised learning techniques use clinical data exclusively in the determination of subgroup membership. In the context of survival time, observations can be partitioned into a "low-risk" or a "high-risk" subgroup based on whether patients are still alive at a certain follow-up time. The determined subgroups $C_1, C_2, ..., C_n$ can be used in training a classifier that predicts the subgroup

membership of a future group of patients. Alternatively, information such as whether a patient's tumor has metastasized or other available clinical thresholds can be used to develop procedures to predict subgroup membership [11]. This approach has been used to identify cancer subtypes in a number of studies [80, 81]. Figure 2.9 depicts the overall process for a supervised learning approach.

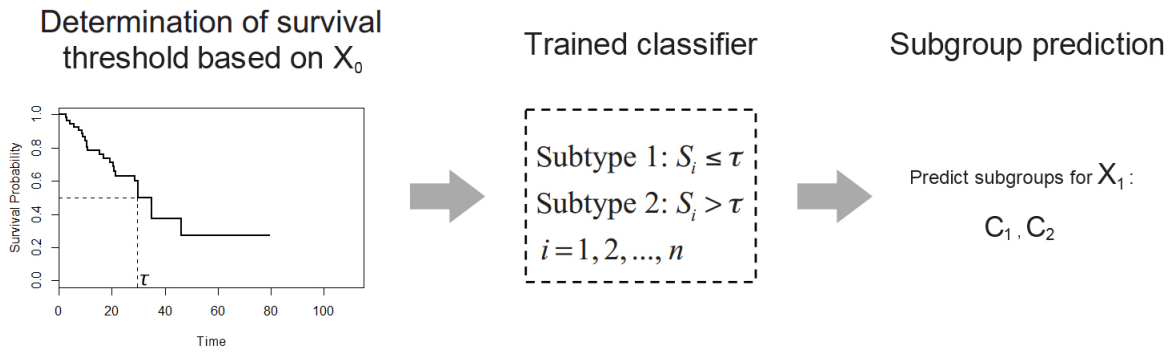

Figure 2.9: The overall process for a supervised learning approach

Once observations in the training data have been labeled with subgroup membership, (i.e., $C_1, C_2, ..., C_n$), this information is used to diagnose future patients. For instance, patients are partitioned into a low-risk or high-risk subgroup based on the observed median survival time as the clinically relevant threshold. The trained median survival time function can then be used for mapping future instances into one class or the other. An optimal scenario occurs when new examples are correctly assigned to their actual subgroups. This requires the trained classifier to generalize from the training data and apply its knowledge to an unseen situation in a "reasonable" way. However, since supervised learning by definition does not involve X guiding the identification of subgroups, the resulting subgroups may lack reasonable biological meaning.

Suppose there are two pre-specified subtypes associated with the same cancer diagnosis, and patients with subtype 1 live somewhat longer than those with subtype 2, as shown in Figure 2.10. Since there is a significant overlap between the two subgroups with respects to survival time, simply assigning observations to the low-risk or high-risk group based on the median survival time would result in an incorrect determination of the subgroups for unseen patients. Therefore, the diagnosis of any future patients based on this model would be questionable. More accurate prediction can be made by including microarray data in the determination of subgroup and by developing a model that can predict which subtype is present in a future diagnosis.

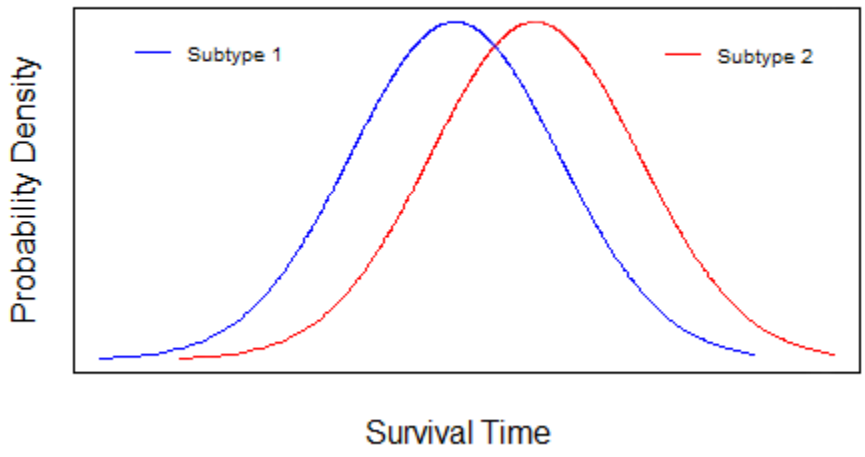

Figure 2.10: Two subtypes of cancer diagnosis with significant overlap

2.2.4 Semi-Supervised Approach

As previously stated, the exclusive use of either microarray data or clinical data in unsupervised and supervised methods for the identification of subgroup inhibits both the discovery of biologically meaningful subgroups and the likelihood of accurately predicting survival outcome. To overcome these difficulties, the semi-supervised approach uses both X and S to guide the determination of cancer subtypes facilitating the identification of biologically meaningful classes that are also correlated with clinical outcome. Here, four universal steps shared by semi-supervised approaches are first discussed, followed by descriptions of two canonical semi-supervised learning techniques, namely, semi-supervised clustering [3] and the semi-supervised risk index method [5].

2.2.4.1 The General Approach to Semi-Supervised Learning

The four universal steps of semi-supervised learning are shown in Figure 2.11.

1. **Data splitting:** The first step to any semi-supervised learning approach is preprocessing the raw data. After this step, the data is partitioned into two sets: a training set tasked with inferring the function, and a testing set used for validating whether the mapping of new examples based on the inferred function is "reasonable". Sometimes the partition is a natural one; this occurs when data is obtained as a collection of training and testing data. In the vast majority of cases, however, data is collected as one complete set without having a training and testing set in mind. Therefore, the random division of data into calibration and validation sets is needed. The random splitting should preserve

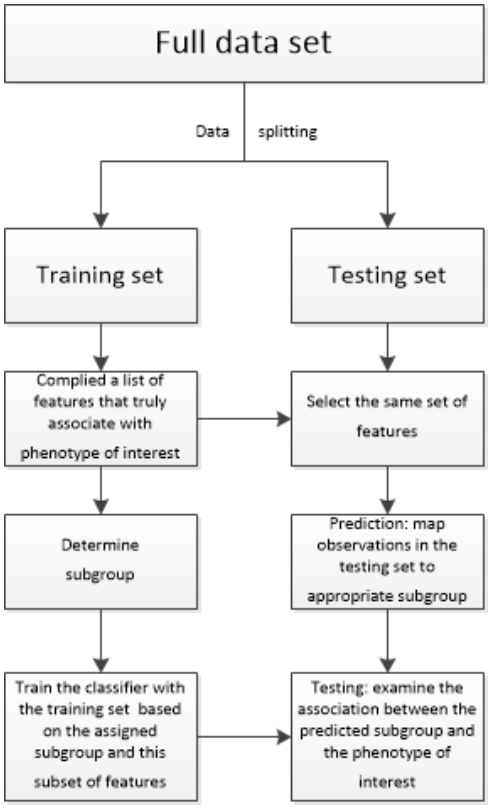

Figure 2.11: The general procedure of semi-supervised learning

the overall class distribution of the data, although conventions vary in terms of the proportion of samples allocated to calibration and validation sets. A half-and-half allocation is preferred because it balances poor predictive performance results from overfitting and an inadequate number of samples in the validation set [61]. In Chapter 4, this is how we split data for the purpose of calibration and validation.

2. **Subgroup assignment based on supervision:** The purpose of this phase is to assign subgroups to the observations in the training set obtained from the previous data splitting step. This action distinguishes the semi-supervised approach from supervised and unsupervised methods. A list of significant features is first identified apropos of the phenotype of interest, and the assignment of subgroups is based on this subset of features. By making use of both X and S to determine subgroups, the identified subgroups are both biologically meaningful and associated with the primary phenotype of interest.

3. **Determining class subgroup in the testing set:** The function inferred from the training set is then used for mapping "unseen" examples in the testing set into the identified subgroups. The two canonical semi-supervised learning techniques apply different inferred functions. The Clustering-Cox method employs the "nearest shrunken centroids" techniques [75] to categorize future observations into the appropriate subgroups. On the other hand, the Risk Index approach is based on percentile cut-off points to classify future examples into suitable subgroups. Specifically, the risk index is created from the cumulative effects of all the identified features' univariate Cox score of each observation.

These two methods will be discussed in detail in sections 2.2.4.2 and 2.2.4.3.

4. **Testing the association with phenotype:** In an ideal scenario, the inferred function correctly assigns new examples in a testing set to their intrinsic subgroups. This requires the inferred function to generalize from the training data and apply this knowledge in a "reasonable" way. The level considered reasonable can be determined by testing the association between subgroup prediction of examples in the testing set and the phenotype of interest. Methods for testing the association will depend on the nature of the primary phenotype. In this thesis, the phenotype is a time-to-event outcome (i.e., survival time), thus a log-rank test and resultant p-value can be used [2].

2.2.4.2 Clustering-Cox Method

Bair and Tibshirani proposed a semi-supervised approach to identify cancer subtypes and predict patient survival [3]. Unlike the fully unsupervised method, where all features are selected for clustering regardless of the phenotype, this method's guiding principle is to apply unsupervised clustering techniques to a list of identified genes that correlate with the primary phenotype of interest. In the case of survival data, the method is carried out as follows.

First, a subset of genes actually associated with survival must be compiled. The most widely used and straightforward way to identify features that individually correlate with survival is to use univariate Cox scores [43]. For each feature, a univariate Cox proportional hazards model is fitted. The feature selection with Cox score (FSCS) quantifies the correlation between gene expression level and patient

survival as follows:

$$S_j = \left(\frac{dl(0)}{d\beta_j}\right) \Bigg/ \left(\frac{d^2l(0)}{d\beta_j^2}\right)^{\frac{1}{2}} = \frac{\sum_{r\in D}\left(x_j^r - \frac{1}{n_r}\sum_{i\in R_r} x_j^i\right)}{\left[\sum_{r\in D}\frac{1}{n_r}\sum_{i\in R_r}\left(x_j^i - \frac{1}{n_r}\sum_{k\in R_r} x_j^k\right)^2\right]^{\frac{1}{2}}} \tag{2.12}$$

where x_j represents each feature, and the other notations are identical to those described in Equation 2.5 (see section 2.1.2.3).

A large S_j indicates that feature j predicts survival well. A positive value of the Cox score suggests that over-expression of that gene is correlated with increased survival, and a negative value indicates decreased survival. Univariate Cox scores are calculated for each gene in the expression data, and only genes with a Cox score that exceeds a certain threshold are considered for clustering purposes.

Once the dimensions of the gene expression data have been reduced using the Cox score, the reduced data is used to identify subgroup. The existing clustering techniques can then be applied to form clusters. Within each identified cluster, patients share small, pairwise distances. As soon as subgroup is available, supervised learning is used to infer a function from training data to match future patients with appropriate subgroups. The supervised learning technique employed in this study is the nearest shrunken centroids procedure of Tibshirani et al. [75], which calculates the mean expression of each gene with each class and then shrinks these centroids toward the overall mean for that gene by a fixed quantity.

2.2.4.3 Risk Index Method

We now discuss another representative semi-supervised method proposed by Beer et al., which involves the prediction of future examples using a risk index inferred function [5]. The objective of the method is to assign a diagnosis to a future patient as described by the cumulative effects of the significant genes of the patient. The state or quality of significance is numerically measured from the risk index and the selected percentile cut-off point determines the assignment of subgroup.

First, a risk index must be created according to the cumulative effects of the selected genes of the patients in the training set. The selected genes are correlated with the primary phenotype of interest, and the selection is based on the FSCS [43]. The selected genes are individually fit with a univariate Cox proportional hazard model, and then a linear combination is constructed from the subset of genes by multiplying the estimated regression coefficients by their corresponding gene expression values and adding the results. This linear combination defines the risk index and evaluates the association of the identified genes with the phenotype.

The next step is to examine the proper cut-off point based on the distribution of risk index values calculated in the training set. Percentile, a common way of reporting scores from norm-reference tests, is then used to categorize patients into different groups. Note that it is difficult to estimate or judge the number of identified genes and the related percentile that have the best overall association with survival. Therefore, a continuum of cut-off points for different numbers of selected genes is examined in order to produce an optimal inferred function. Using the risk index function in the training set, future patients in the testing set are placed in the appropriate subgroups.

Both semi-supervised clustering and the risk index based-method make use of X and S for the identification of cancer subtypes and by doing so guarantee the identified subtypes are biologically meaningful and will strongly predict survival time. These two methods yield satisfactory results in many datasets [85, 11, 5].

Chapter 3

Using the Machine Learning Approach for High-Dimensional Survival Data

This chapter outlines the machine learning approach to feature selection and the survival prediction of patients from high-dimensional survival data. Two new methods designed to cope with censoring and to discover latent class memberships are first introduced, followed by the details of the proposed approach.

3.1 Coping with Censoring

Classical statistical methods account for censoring with the Cox model, which keeps censoring individuals in the risk set along with other individuals who have not yet experienced the event of interest [43]. In preference to standard statistical methods,

we use k-nearest neighbor based on clinical parameters that are truly associated with survival time to cope with patients' censoring. These clinical parameters are selected using penalized logistic regression and the penalized proportional hazards model with the Expectation Maximization (EM) algorithm. They are then used to estimate censored survival time.

3.1.1 Selecting Clinical Parameters that Associate with Phenotype of Interest

A new survival time associated for each censored individual is computed according to the proximity of clinical data to observations. Variable selection in the proportional hazards mixture cure model based on penalized likelihood is used to select important clinical covariates that are associated with the phenotype of interest (i.e., patient survival time) [50].

Let $p(t|z)$ be the probability of being cured given a covariate vector $z = (z_1, ..., z_q)'$, $S(t|x)$ be the survival function for uncured patients, conditional on $x = (x_1, ..., x_m)'$, and $\odot = p(t_i, \delta_i, x_i, z_i, y_i)$ represent the complete data for the ith individual, $i = 1, ..., n$. As specified in Chapter 2, the observed survival time that is possibly censored is given by t_i; δ_i is the censoring indicator; z_i and x_i are covariates in the incidence and latency parts of the function, respectively; and y_i is an indicator of cure status, where $y_i = 0$ if the patient is uncured and $y_i = 1$ if the patient is cured. Here, there is missing information because if $\delta_i = 1$, $y_i = 0$, but y_i can either be 1 or 0 when $\delta_i = 0$. Therefore, y is partially missing information and the EM algorithm is appropriate for estimating the parameter of interest $\Theta = (x_i, z_i, S_0(t))$, where

$S_0(t)$ is the corresponding baseline survival function with respect to $h(t)$. Given $\odot$, the complete likelihood function of proportional hazards mixture cure model is

$$\prod_{i=1}^{n} p(z_i)^{1-y_i}[1-p(z_i)]^{y_i} h(t_i|x_i)^{\delta_i y_i} S(t_i|x_i)^{y_i} \tag{3.1}$$

The EM algorithm can be implemented based on the complete likelihood function Equation 3.1. To select significant clinical parameters, we add a Lasso penalty to the likelihood function to form the penalized likelihood function. The new penalized complete log likelihood function can be written as

$$\begin{aligned} l(\gamma,\ \beta;\ \odot) \ &= \ \sum_{i=1}^{n}(1-y_i)\log[p(z_i)] + y_i\log[1-p(z_i)] \\ &+ \ \sum_{i=1}^{n} y_i\delta_i\log[h(t_i|Y=1,\ x_i)] + y_i\log[S(t_i|Y=1,\ x_i)] \\ &+ n\lambda_{1j}\sum_{j=1}^{q}|\gamma_j| \ + \ n\lambda_{2k}\sum_{k=1}^{m}|\beta_k| \end{aligned} \tag{3.2}$$

where $\lambda|\cdot|$ is the Lasso penalty function and $\lambda = (\lambda_{11},\ldots\lambda_{1q},\ \lambda_{21},\ldots\lambda_{2m})$ is the tuning parameter, which can be chosen via GCV as discussed in Chapter 2.

For a selected value of λ, the EM first calculates the expected value of the penalized log likelihood function with respect to the conditional distribution of y_i, given the observed data and current estimates of parameters $(\gamma^r,\ \beta^r,\ S_0^r(t))$. In order to accelerate the estimation process, we assign values to the first estimates that maximize the un-penalized log-likelihood. The logarithm of the penalized complete likelihood function can be expressed as $l_c(\gamma,\ \beta,\ S_0(t);\ \odot) = \ l_{c1}(\gamma;\ \odot) \ + \ l_{c2}(\beta,\ S_0(t);\ \odot)$, where

$$l_{c1}\left(\gamma;\ \odot\right) \ = \ \sum_{i=1}^{n}(1-y_i)\log[p\left(z_i\right)\,] + y_i\log[\,1-p\left(z_i\right)] \ + \ \mathrm{n}\lambda_{1j}\sum_{j=1}^{q}|\gamma_j| \tag{3.3}$$

$$l_{c2}\left(\beta,\ S_0\left(t\right);\ \odot\right) \ = \ \sum_{i=1}^{n} y_i\delta_i\log[h(t_i|Y=1,\ x_i)\,] + y_i\log[\,S_0(t_i|Y=1,\ x_i)] + \mathrm{n}\lambda_{2k}\sum_{k=1}^{m}|\beta_k| \tag{3.4}$$

The conditional expectation of y_i will be enough to complete this step because both (3.3) and (3.4) are linear functions of y_i. The expectation of $\mathrm{E}\left(y_i|\ \gamma^r,\ \beta^r,\ S_0^r\left(t\right)\right)$ can be written as follows:

$$w_i^{(r)} = \mathrm{E}\left(y_i|\ \gamma^r,\beta^r,\ S_0^r\left(t\right)\right) = \ \delta_i + (1-\delta_i)\,\frac{\left[1-p\left(z_i\right)\right]S_0\left(t_i|Y=1,\ x_i\right)}{p\left(z_i\right)\ +\ \left[1-p\left(z_i\right)\right]S_0\left(t_i|Y=1,\ x_i\right)}\Big|_{\gamma^r,\,\beta^r,\ \ S_0^r(t)} \tag{3.5}$$

It is clear that $w_i^{(r)}=1$ if $\delta_i=1$ and $w_i^{(r)}$ is the probability of uncured patients if $\delta_i=0$. The second part of $w_i^{(r)}$ actually represents the conditional probability of the ith individual remaining uncured.

The M-step in the *(r+1)*th iteration is to maximize (3.3) and (3.4) with respect to Θ. The parameters in function (3.3) can be maximized using a penalized logistic program to obtain γ^{r+1}. Likewise, the function (3.4) is the penalized log-likelihood function of the PH model with the additional offset variable $logw_i^{(r)}$ with fixed coefficient 1. The Lasso-penalized clinical parameter selection is made through the quadratic approximation procedure [36]. In addition, if prior biological knowledge shows that a certain variable has a known involvement in the cancer process, we can

remit the penalty on the variable by setting the corresponding tuning parameter in λ to zero. For example, we do not place any penalty on the intercept because it is always in the logistic regression part. Note that the estimation of γ and β do not depend on the assumption of the distribution [48].

3.1.2 Estimating Censored Survival Time

After a list of significant clinical parameters has been compiled, we approach this problem with the intuitive idea that things of a kind come together. Specifically, we compute the new survival time with the selected clinical covariates with regards to the proximities among them. The definition of "proximity" we employ here is a variant of the Euclidian distance such that it is applicable to numerical clinical variables as well as nominal clinical variables [54]. The expression for proximity is

$$d\left(x,\ y\right) = \sqrt{\sum_{i=1}^{p} \phi_i\left(x_i,\ y_i\right)} \tag{3.6}$$

where $\phi_i()$ is the distance between two variables defined as follows:

$$\phi_i\left(v_1,\ v_2\right) = \begin{cases} 1 & \text{if } i \text{ is a nominal variable and } v_1 \neq v_2 \\ 0 & \text{if } i \text{ is a nominal variable and } v_1 = v_2 \\ \left(v_1 - v_2\right)^2 & \text{if } i \text{ is a numeric variable} \end{cases} \tag{3.7}$$

The ten uncensored neighbors with the smallest proximities are selected to compute the event time of interest associated with the censoring time. In addition, weights are assigned to the contributions of the neighbors, such that the nearer

neighbors contribute more to the average than the more distant ones. The Gaussian function is used to obtain the weights [62]. If the neighbor is positioned at a distance d away from this censored observation, then the weight of this uncensored survival time is

$$w(d) = e^{-d} \tag{3.8}$$

3.2 Identify Latent Class Membership

As stated in Chapter 2, in order to identify biologically meaningful subtype membership that accurately determines the subtypes of future patients, features need to be chosen respective of the phenotype of interest. Current statistical approaches exclusively rely on the Cox score category to identify subgroup [47, 22, 11]. We instead propose to discretize the survival time of patients; thus, sets of patients with measured similarities will share an identical class label, while this value is minimized between patients with different class labels.

A prerequisite for applying the partitioning technique to discretize the continuous phenotype of interest in clinical data, which applies to the probabilistic naturalized PAM algorithm [63] used here, is the optimal choice of the number of splitting bins from the data. "The best k" of clusters to be formed should allow an appreciation of the relative quality of the clusters and overall structure of the data [14]. There are methods for choosing "the best k" for different discretization algorithms, such as the heuristic method from Dougherty [26], in which $k = \max\{1,\ 2 * \log l\}$, where l is

the number of distinct observed values for the attribute being partitioned. Another method, proposed by Crescenzi [21], relies on the knowledge of data properties to determine the "natural" number of clusters. However, the degree to which these apriori clustering schemes reflect a specific dataset is still under-investigated in the literature [7].

While the aforementioned methods specify the number of clusters to be formed ahead of the clustering process, we apply another approach, silhouettes [66], to select the optimal number of clusters in a partitioning by evaluating the validity of the produced clustering. More specifically, the silhouettes method interprets the clustering results and selects the best clustering scheme with the most compact and clearly separated clusters regardless of the clustering algorithm used.

3.2.1 Construction of Silhouettes

Suppose that there are n objects to be clustered. The clustering technique assigns these n objects into k clusters such that objects within the same cluster bear more resemblance to each other than to those in other clusters. Although it is easy to construct clusters of data based on the clustering algorithm, little is known about the number of "natural" clusters that are actually present. The silhouettes method is designed to provide additional guidance and deeper insight into an optimal choice of partition.

Silhouettes are constructed from the partitions obtained through a selected clustering algorithm, as well as the metric of collected proximities between each individual object. Assuming the data have been clustered via a certain technique,

silhouettes are constructed such that the value *s(i)*, which measures how well an object i has been classified, is defined for each object i. These numbers are then combined to form silhouettes representing each cluster to provide an overview of data configuration. The computation of *s(i)* concerning a specific datum i in the dataset consists of calculating the average dissimilarities of i to the objects in the same group as i and those in a different group. These two dissimilarities, denoted *a(i)* and *c(i)*, can then be compared to give a quantitative measure of how well i is assigned to its home group relative to other groups. Let A be the cluster to which i has been assigned, where A contains other objects apart from i (not a singleton); cluster C is any distinct partition from A; and B is the cluster in which i is not a member (i.e., the neighbor cluster of object i) and holds the lowest average dissimilarities *c(i)*. These dissimilarities are formulated as follows:

$$\begin{aligned} a\,(i) &= \text{ average dissimilarities of } i \text{ to all other objects of } A \\ c\,(i) &= \text{ average dissimilarities of } i \text{ to all other objects of } B \\ b\,(i) &= \text{ minimun } d\,(i,\ C)\,,\ C \neq A \end{aligned} \tag{3.9}$$

The neighbor of i can be viewed as the second-best choice for object i, that is, if cluster A is discarded, then cluster B is the next best fit. It is worth mentioning that the construction of silhouettes depends on the availability of clusters distinct from A. Therefore, the number of attained clusters, k, is at least greater than one. The value *s(i)* is obtained by combining *a(i)* and *b(i)* as follows:

$$s(i) = \begin{cases} 1 - a(i)/b(i) & if\ a(i) < b(i) \\ 0 & if\ a(i) = b(i) \\ b(i)/a(i) - 1 & if\ a(i) > b(i) \end{cases} \tag{3.10}$$

As we can see from this formula, the value of *s(i)* for each object i ranges from -1 to 1, and remains invariant when all the original dissimilarities are multiplied by a positive constant. Consequently, a dissimilarity of 4 is considered twice as large as a dissimilarity of 2.

A high *a(i)* value indicates a strong dissimilarity between datum i and its own cluster, whereas a small *a(i)* value suggests that it is well classified within its cluster. Moreover, a large *b(i)* value implies that the next best fit is poorly matched to datum i. Thus, for *s(i)* to be close to 1, we require the "within" similarity *a(i)* to be much smaller than the minimum "between" similarity, *b(i)*. This suggests that object i resembles more objects within the same cluster than those in other clusters, giving strong evidence that i has been assigned to the most suitable cluster. On the other hand, if *s(i)* is close to -1, then, by the same logic, datum i generally lies much closer to the neighboring cluster than its home cluster. In this circumstance, assigning object i to the home cluster instead of a neighboring cluster would be questionable. The intermediate is observed when object i lies equally far away from cluster A and B, giving two roughly identical values of *a(i)* and *b(i)*, and resulting in an *s(i)* close to zero. Therefore, it remains uncertain which cluster is the best choice for object i. In the special case where A is a singleton, the value of *s(i)* is neutrally set to 0. When the data consist of similarities, the silhouettes method can be used with a

slight modification:

$$s(i) = \begin{cases} 1 - {}^{b'(i)}/_{a'(i)} & if\ a'(i) > b'(i) \\ 0 & if\ a'(i) = b'(i) \\ {}^{a'(i)}/_{b'(i)} - 1 & if\ a'(i) < b'(i) \end{cases} \tag{3.11}$$

where $b'(i) = maximum\, d'(i,\, C),\ C \neq A$.

After computing the quantities *s(i)* from either similarities or dissimilarities, the construction of silhouettes is possible. The silhouette of a certain cluster plot contains *s(i)* for all objects belonging to this cluster ranked in decreasing order. Its height equals the number of objects contained in this cluster. The length of each line printer is proportional to the corresponding *s(i)*. Therefore, the silhouette provides a succinct graphical representation of how well each object lies within its cluster, that is, the wider a silhouette, the larger the *s(i)* value, and the more pronounced the cluster. The final step of the construction is to incorporate the silhouettes of different clusters into a single plot. The entire clustering is displayed one after another, which enables us to distinguish appropriate clusters from unnatural ones and provides an evaluation of clustering validity.

3.2.2 Selecting the Appropriate Number of Class Label

We illustrate the silhouettes method using the dataset of Beer et al. [5], which is described in the description of datasets section. In Figure 3.1, the silhouettes representing $k = 2$ for the clustering of lung cancer data is shown. Below the plot, there is a vertical scale ranging from 0.0 to 1.0 with steps of size 0.2. The computed

quantity *s(i)* of each patient corresponds to the markings on the scale. The second silhouette is higher than the first because the second cluster contains 61 patients compared to 25 in the first cluster. The rightmost section of the plot reflects the average *s(i)* for all objects i in a cluster (i.e., the average silhouette width of that cluster) and measures how tightly grouped the data is in the cluster. Both silhouettes in Figure 3.1 are rather wide, contributing to high average silhouette widths that imply a relatively strong clustering structure. If we compare these two cluster results from the same partition, the second cluster is tighter and better separated due to its wider average silhouette width. In particular, the patients in the second cluster possess the largest *s(i)*, which means that it is classified with the least amount of doubt.

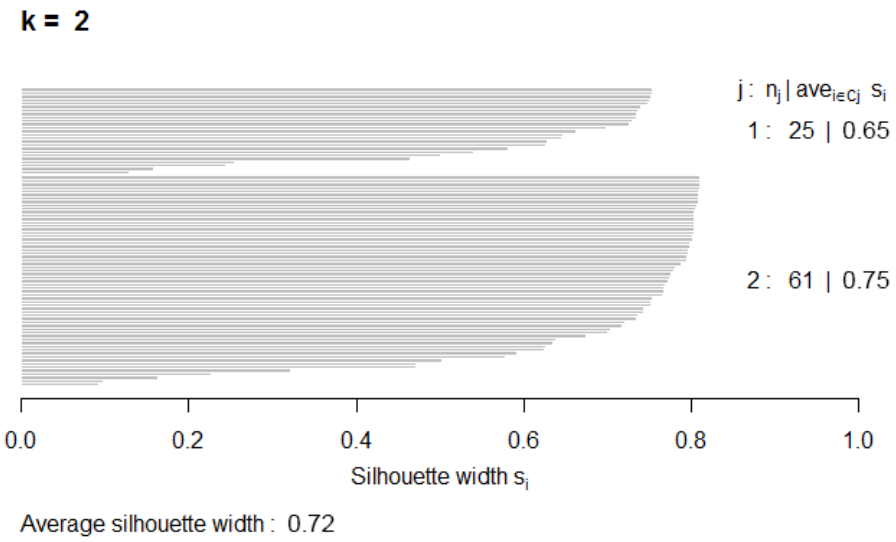

Figure 3.1: Silhouettes of clustering with $k = 2$ of the lung cancer data.

For each individual patient, the silhouettes also reveal the index of the cluster to which it belongs, the index of its neighboring cluster, and the exact numerical value of *s(i)*; however, this information is omitted due to the limited space on the plot. If the computed *s(i)* value is close to zero, the length that corresponds to the marking on the scale becomes zero as well. Thus, the object lies on the border of two natural clusters. In the case of negative values, the length of the line printer is the absolute value of *s(i)*. We can identify the "misclassified" object by looking at the sign of *s(i)*. Alternatively, because the silhouette is plotted in decreasing order, any objects ranked below an object with a *s(i)* value of 0 are "misclassified" cases. For these objects, the longer its silhouette length, the more naturally it fits into its neighboring clustering. We can improve the clustering results by moving these objects to their neighbors.

The very last number, listed below the scale, is the average silhouette width for the entire dataset. This number, denoted by $\bar{s}\,(k)$, plays a key role in selecting the natural value of k, as it measures how appropriately the data has been clustered. Figure 3.2 shows silhouette plots of the lung cancer data for k with the four highest average silhouette widths after computing average silhouette widths for PAM partitions corresponding to all possible values of k. In general, different k varies in average silhouette width. The silhouettes should appear as wide as possible for a natural value of k, thus one way to choose k appropriately is to select the value that yields the maximum average silhouette width. We see that the computation of average silhouette widths for all possible k returns a highest average silhouette width value of 0.72 when $k = 2$, so we select $k = 2$ to discretize the continuous time space specific to this lung cancer dataset.

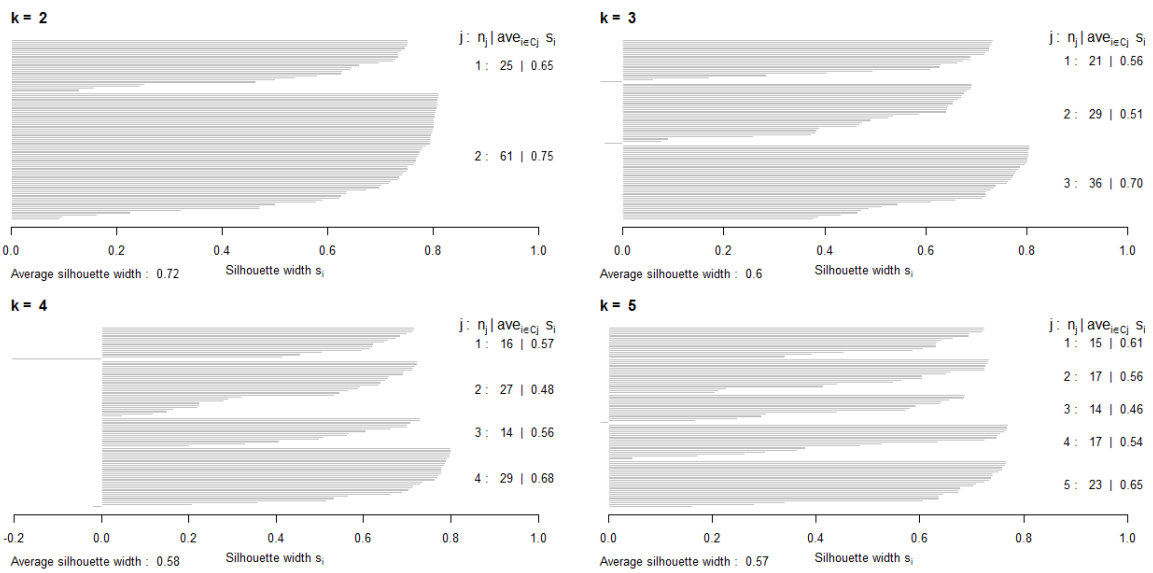

Figure 3.2: Silhouettes plots of the lung cancer data, for k with 4 top average silhouette widths

The aforementioned procedure can be extended to other datasets by computing the maximal average silhouette width for the entire dataset, defined as the silhouette coefficient (SC),

$$\mathrm{SC} \;=\; \max \bar{s}\,(k) \tag{3.12}$$

where the maximum is selected over all k for which silhouettes can be constructed.

In the context of our proposed approach, which seeks to identify meaningful survival subtypes that accurately predict outcome, the value range of k is set to less than or equal to 10 because there rarely exists a cancer with more than 10 subtypes [6]. The SC offers a useful measure of the relative quality of the clusters that have

been discovered by the classification algorithm. Table 3.1 gives an interpretation of the SC [42].

Table 3.1: Interpretation of the silhouette coefficient (SC)

Silhouette coefficient	Interpretation
0.70 - 1.00	A very clear structure has been found
0.50 - 0.69	A reasonable structure has been found
0.26 - 0.49	The structure is weak and could be artificial; extra improvement or alternative method is needed
<= 0.25	No substantial structure has been found

Indeed, an SC close to 1 means the data is officially clustered, while a low SC suggests that an alternative method of data analysis is more appropriate. For the lung cancer dataset, $\bar{s}(k) = 0.72$, which is indicative of a strong structure. One scenario that requires special attention is when certain far-enough clusters contain only a single object or relatively few objects. Such singleton clusters are regarded as outliers. Depending on the context and the task at hand, we can set aside the outliers for further investigation and rerun the clustering algorithm on the remaining data. The overall average silhouette width as well as the graphical output itself assist us in determining the natural number of clusters within a dataset.

3.3 Feature Selection for High-Dimensional Survival Data

As mentioned previously, current methods of survival analysis with high-dimensional data use the Cox score to quantify how well each feature predicts survival, and the selected features are candidates for follow-up experiments [43]. This selection process is based on the association between each individual feature and the survival outcome. The computation also requires iteratively fitting a Cox model for each feature, which can be computationally inefficient when the number of features is very large. Also, that model does not have the ability to reduce relevancy; thus it can potentially select highly redundant features. It is well known that when the dimension is high, redundant features can cause over-fitting, which jeopardizes the generalization capability. Many feature selection methods in machine learning, on the other hand, can identify relevant features as well as redundancy among relevant features in an efficient manner.

Selecting an informative subset of features from an existing feature space plays an important role in any machine learning applications. This is even more so in our context where the class label is itself one of the learning targets. As a result, uninformative features can produce overfitting with more unpredictable results in our context than in applications where the class labels are given a priori. Therefore, the appropriateness of the employed feature selection method is crucial in success of the approach and should be adopted in extreme care.

Feature selection in machine learning falls into two broad categories: the wrapper method [44] and the filter method [24]. In the wrapper method, the feature subset

selection algorithm conducts a search for a good subset using the induction algorithm itself as part of the function evaluating feature subsets. It enforces a straight order and may be stopped at a local maximal. A typical example of the wrapper method is Gradient Boosted Feature Selection (GBFS) [83]; GBFS starts with pre-specified parameters, then iterates using embedded greedy CART algorithm to minimize the selected impurity function for learning regression tree until the stopping condition is met. Unlike a wrapper approach, a filter method defines a metric that, independently of any learning algorithm, measures the relevance of each feature with the outcome. Many sophisticated filer methods also incorporate redundancy reduction mechanisms. All these proceed without involving any learning algorithm. An example of a filter method is Fast Correlation-Based Filter (FCBF) [84]; FCBF first removes and ranks features according to predominant correlation. It then selects a subset of features according to a user-specified threshold. Different from GBFS which assumes that one can pre-process the data with limited-depth trees, FCBF is free of data assumption. In our approach, since it plays a pivotal rule, feature selection should serve as an independent component, without being influenced by the bias of any specific learning algorithm. Furthermore, the feature selection methods used by many existing works in predicting patient survival are based on Cox scores, which is essentially a filter method. Therefore, using filter also in our context can make comparison with these works more meaningful. In summary, we prefer filter in our work, and will use FCBF for feature selection due to its effectiveness in handling both relevance and redundancy, as well as its pronounced efficiency.

The FCBF selects significant features for prediction based on correlation analysis. Two information-theoretical concepts, entropy and information gain, are chosen to

measure the correlations. The first concept, entropy, is a measure of the uncertainty of a random variable [10], where the entropy of a variable X is defined as

$$\mathrm{H}\,(X) = \; - \sum_{i} P\,(x_i) \log_2 (P\,(x_i)) \tag{3.13}$$

and the entropy of X after observing values of another variable, Y, is defined as

$$\mathrm{H}\;\;(X \mid Y) \;\; = \;\; - \sum_{j} P\,(y_i) \sum_{i} P(x_i |\; y_i) \log_2 (P\,(x_i \mid y_i)) \tag{3.14}$$

where $P\,(x_i)$ are the prior probabilities for all values of X, and $P(x_i |\; y_i)$ are the posterior probabilities of X given the values of Y. The amount by which the entropy of X decreases provides additional information about X from Y and is called information gain [61].

$$\mathrm{IG}\;\;(X \mid Y) \;\; = \;\; \mathrm{H}\,(X) \;\; - \;\; \mathrm{H}\,(X \mid Y) \tag{3.15}$$

If we have $\mathrm{IG}\,(X \mid Y) \;\; > \;\; \mathrm{IG}\,(Z \mid Y)$, it means a feature Y is more correlated to feature X than to feature Z. In reference to this work, X and Z are genes to be selected and Y can either be a gene or the class label estimated from the discretization of time space. When Y is the feature, IG measures the redundancy between these two genes. Alternatively, IG quantifies the predictive ability of the gene in class memberships.

However, information gain is biased in favor of features with more values. The values also have to be normalized to ensure they are comparable and have the same

effect. Therefore, symmetrical uncertainty (SU) [60] is introduced:

$$\mathrm{SU}\left(X,\ Y\right) = 2\left[\frac{\mathrm{IG}\left(X \mid Y\right)}{\mathrm{H}\left(X\right)\ -\ \mathrm{H}\left(Y\right)}\right] \tag{3.16}$$

The theory of the FCBF algorithm is based on the aforementioned concepts and is executed as follows:

1. Calculate the SU value for each feature related to the classification;
2. Eliminate irrelevant features according to a predefined threshold SU value in order to build a list of relevant features, S_{list}, based on their SU values and appearing in descending order;
3. Calculate the SU value for each feature in S_{list} (except for F_r itself) related to F_r, which is the first feature of S_{list};
4. Eliminate redundant features in S_{list} according to a predefined threshold SU value and take the remaining feature beside F_r as the new reference;
5. Repeat steps 3 and 4 until a target number of features is selected or there are no more features to be removed from S_{list}.

In general, FCBF first decides whether a gene is relevant to the survival outcome, and then determines whether the gene is redundant when considering it in relation to other relevant genes. The computation of SU is also much more efficient than iteratively fitting a Cox model for each feature.

3.4 Classification

After generating a list of features that correlate with the phenotype of interest, a classifier can be applied to this subset of genes to determine which subtype is present in a future patient. We use Naïve Bayes and Decision Tree for their known effectiveness and applicability.

3.4.1 The Naïve Bayes Classifier

In machine learning, the Naïve Bayes classifier (NB) is a probabilistic classifier using Bayes' theorem [68]. Instead of assigning an instance to a certain category, it calculates the probability of this instance belonging to each category and chooses the largest one. Suppose there are m classes, $\mathbf{C} = \{C_1, C_2, \ \ldots, \ C_m\}$, and a given instance with feature $\mathbf{X} = \{X_1, X_2, \ \ldots, \ X_m\}$, then the posterior probability that this randomly given instance belongs to a class C_i is $P(C_i \mid X) = {}^{P(C_i, \ X)}/_{P(X)}$. Using Bayes' theorem, the Bayes' classifier can be represented as follows:

$$P(C_i \mid X) = \frac{P(X \mid C_i)\, P(C_i)}{P(X)} \tag{3.17}$$

That is, the classification function assigns each instance to the class that has the highest probability of containing this instance. In practice, there is interest only in the numerator of the fraction since $P(X)$ is the same for all C_i, and the Bayes' criterion is equivalent to classifying X in the class that maximizes $P(X \mid C_i)\, P(C_i)$.

We choose the NB algorithm mainly for its simplicity and effectiveness. In addition, the NB classifier is general-purpose and good enough for most applications,

even if the NB assumption does not hold [70, 51]. Therefore, it is one of the first methods to try in a classification problem.

3.4.2 The Decision Tree Classifier

Another type of commonly used classifier in the field of machine learning is the decision tree (DT) classifier [68]. A decision tree represents classification rules and is a flowchart-like structure in which each internal node is a test on some attribute values; each branch is one of the outcomes of the test, and each leaf node is a class or a group of classes. The algorithm begins with the original set S as the root node and iterates through every unselected attribute of S. The "best" attribute is chosen at each iteration to split the set S into subsets of data. The "best" commonly means the homogeneity of the target attribute within the subset and can be measured with a myriad of metrics, such as the information gain of that attribute. The iteration stops when all objects at a node have an identical class label, or when there are no more attributes to be selected nor examples in the subset. The pseudocode for building the decision tree is as follows:

input : $\mathrm{S}\,(R_1,\ R_2,\ \ldots\ ,\ R_N,\ X)$

// a training data set contains N attributes and all objects X

output: decision tree

begin

while $R = \emptyset\ or X = \emptyset, or\ all\ objectives\ in\ X\ have\ an\ identical\ class\ label$

do

for $i \leftarrow 1$ **to** N **do**

calculate the metric value for R_i;

end

select the attribute with the highest metric value and remove it from R;

create a decision node that splits on the selected attribute;

end

end

Compared to other classifiers, a decision tree can effectively cope with outliers and nonlinearly separable data due to its non-parametric characteristics and its ability to recognize feature interactions [65]. Moreover, it is easy to interpret and explain.

In summary, our proposed approach uses both gene expression data and clinical data to identify latent class membership and to determine which subtype is present in a future patient. We first discretize the time-space to discover the hidden structure based on the estimated class label. A list of genes is then selected based on their association with survival time by using feature selection combined with identified class label. The selected genes are used to classify patients, and the resulting classification solution is then applied as the scheme for identifying cancer subtypes. Subtype is predicted for future patients using the trained classifier.

Chapter 4

Results and Discussion

This chapter presents an empirical study designed to evaluate the performance of our proposed method on high-dimensional survival data and a comparison to related methods reported in the literature. The evaluation is based on feature selection and patient prediction using two real datasets in addition to simulation data. The "leave one out" approach is used to further validate experimental results and to maximize predictive accuracy.

4.1 Experiments on Real-World Datasets

4.1.1 Description of Datasets

Compared with other applications, survival prediction has a unique requirement for its datasets: a sufficient long-term follow-up of patients. This renders many existing datasets not being applicable. We hereby selected two benchmark datasets, which are publicly available and have been used by many existing works for survival analysis

[22, 11, 3]. The first dataset, the lung cancer dataset, is described in detail by Beer et al. [5]. Briefly, the gene-expression profiles of 86 primary lung adenocarcinomas were obtained using the oligonucleotide arrays; there are 7,129 genes. Several clinical variables, including patient outcome, are available for each patient. The second dataset, renal cell dataset, consists of conventional renal cell carcinoma (cRCC) data on 177 patients who underwent radical nephrectomy for cRCC; there are 5,559 genes [85]. For each of the subjects in the renal cell dataset, gene expression was assessed in cRCC tissue samples using the oligonucleotide arrays. In addition to the clinical data, survival outcome was also included. Because the survival data consist of both clinical data and gene expression data, we discuss the preprocessing of each. This process is illustrated with the example of lung cancer data, and the process for renal cell data is in an identical manner.

4.1.2 Data Preprocessing

4.1.2.1 Clinical Data Preprocessing

Clinical parameters that are truly associated with survival time are selected to cope with censored survival time, as discussed in Chapter 3. As one of the preprocessing steps, continuous parameters have to be discretized and a numerical representation is needed for non-numerical variables. Except for survival time, there are ten clinical covariates available. Sex is a binary covariate indicating whether a patient is male or female, coded as 1 or 0, respectively. Age is a binary covariate with a value of 1 if a subject's age is less than or equal to 50 years old, and 0 otherwise. Nodal status is a binary covariate for the presence or absence of lymph nodes and is labeled 1 if

present and 0 if not. Differentiation is a measure of categorical levels of tumor grade and is coded as 3, 2, and 1 corresponding to well, moderate, and poor, respectively. The p53 nuclear protein plays an important role in directing the protein into the nuclear compartment; its label is p53 nucl.accum. and it is assigned a value of 1 if the status is positive and 0 if negative. Similar to the p53 nuclear protein, KRAS mutations, which define a distinct molecular subset of the disease, have a status equal to 1 if positive or 0 if negative. Pack year measures the frequency of smoking: a non-smoker is coded as 0; fewer than 20 packs per year is regarded as slight smoking and is coded as 1; moderate smoking of 20 to 60 packs is coded as 2; and deep smokers who smoke more than 60 packs per year are coded as 3. A list of 86 patients with these baseline characteristics is given in Table 4.1. Other covariates (e.g., tumor size, estrogen receptor status and disease stage) are already numerically represented and well categorized. Therefore, preprocessing of these covariates is not necessary.

4.1.2.2 Gene Expression Data Preprocessing

The representation and quality of the gene expression data affect the success of the machine learning approach on a given task. However, raw gene expression data normally presents genes that are not expressed, genes that do not code for any protein, or genes that are expressed but show minimal variation across the sample. Moreover, excluding biological variations in gene expression levels that we are actually interested in, variations due to the measurement process should be eliminated or minimized if possible.

We employ a two-step method to address these two problems that accompany gene expression data. First, a variation filter is used to determine transcript abundance

Table 4.1: Baseline characteristics of the 86 patients in lung cancer study

Factor		**Number of patients**
Sex	Male	35
	Female	51
Age	<= 50 years	8
	> 50 years	78
Nodal status	Presence	69
	Absence	17
Differentiation	Poor	21
	Moderate	42
	Well	23
p53 nucl.accum.	-	69
	+	17
K-ras mutation	-	46
	+	40
Smoking (pack years)	Non- smoker	9
	Slight smoker	13
	Moderate smoker	38
	Deep smoker	26

and to exclude uninformative genes. Specifically, the dataset as a whole is trimmed of genes expressed at extremely low levels and with minimal variation across the sample, that is, genes were excluded if their 75th percentile value was less than 100 or the variance was less than one-fifth of the interquartile range of the whole dataset. Although potentially resulting in the loss of some information, trimming in this manner facilitates the success of the machine learning algorithms by reducing the impact of genes with little or no expression in these samples, leading to groups of data that can be used to assign biological meaning to the expression profiles. This step yields 4,804 genes from the initial 7,129 genes in the lung cancer data. Next, the resulting gene expressions are centralized as the computation of normalization in order to compensate for technical differences and enable informative comparisons between different genes. In addition, as discussed in Section 2.1.2.6, continuous gene expression features are discretized by using the recursive minimal entropy partitioning method [31] to fulfill the nominal features requirement of the entropy-based methods [84].

4.1.2.3 Data Splitting

After the previous two steps, the processed clinical data along with gene expression data are ready to use experimentally. The data is then split for the purpose of calibration and validation. The half-and-half allocation is employed as discussed in Section 2.2.4. Specifically, the lung cancer dataset consists of 86 primary lung adenocarcinomas and 4,804 genes from the initial 7,129 genes, which were randomly assigned to an equal number of patients, that is, 43 patients in the training set and 43 patients in the testing set. The renal cell dataset examined is comprised of 177

patients and 4,438 genes after variation filter. This dataset was partitioned into a training set of 89 patients and a test set of 88 patients.

4.1.3 Experimental Setup

All programs were developed using RStudio as the integrated development environment for *R* programming language, with functions taken from DMwR, discretization, survival, cluster, rpart, glm, coxph, caret, e1071, optim, constrOptim, cvTools, and infotheo libraries for existing algorithms. Otherwise, data generation, data preprocessing, variable selection in semiparametric cure models based on penalized likelihood, and FCBF were implemented from scratch due to a lack of available functions in the *R* library. All experiments were carried out on the compute nodes "defiant" and "aalen" at Memorial University of Newfoundland.

4.1.4 Empirical Study and Comparison with Statistical Methods

The objective of this section is to evaluate our proposed machine learning approach for high-dimensional survival data and to compare its performance to state-of-the-art statistical methods on real-world datasets.

4.1.4.1 Selecting Significant Clinical Parameters

Except for survival time, there are ten clinical parameters available from the lung cancer dataset. As discussed in Chapter 3, the penalized logistic regression and the penalized proportional hazard model are used to select significant clinical parameters.

The full model, including all parameters, is also estimated as the reference. The estimated coefficients in the full model and the final model selected by penalized function are listed in Table 4.2.

Table 4.2: Selecting significant clinical parameters

Clinical	**Logistic Part**		**Survival Part**	
Paramters	Full Model	Lasso	Full Model	Lasso
Age	-0.117		0.148	
Differentiation	0.172		-0.183	-0.129
Disease stage	1.178	1.320	1.531	1.681
Estrogen receptor	0.259		-0.196	
K-ras mutation	0.338	0.321	0.465	0.481
Nodal status	-0.408	-0.348	-0.250	-0.232
p53 nucl.accum.	0.291	0.217	0.371	0.154
Sex	0.065		-0.073	
Smoking (pack years)	-0.203		-0.328	
Tumor size	0.565	0.547	0.407	0.421

As we can see from Table 4.2, the Lasso-penalized method suggests that clinical parameters – disease stage, k-ras mutation, nodal status, p53 nucl.accum. and tumor size – play a key role in predicting the patient's probability of being cured. Regarding the survival probability of the uncured patients, Lasso highlights one more clinical parameter: differentiation. The optimal values of λ_1 in logistic regression and λ_2 in survival regression that were selected by GCV are 0.381 and 0.012, respectively. Note

that the intercept is always found in the logistic regression part, but considering that our purpose is to select important clinical parameters instead of identifying the exact model, this is not shown in Table 4.2. We therefore use these six important clinical parameters to estimate censored survival time as discussed in Section 3.1.2.

4.1.4.2 Comparison of Feature Selections between Machine Learning and the Statistical Perspective

Using FCBF, we generate a list of significant features. We compare FCBF with a typical statistical approach in terms of the co-relations of the identified features to the phenotypes of interest and the time they take to identify such a list of significant genes.

The idea of Bair's semi-supervised approach [3], covered in Chapter 2, is chosen as a good fit for the comparison. First, FCBF and FSCS are used as representative feature selection in machine learning and statistical feature selection, respectively, to identify pairwise subsets of genes with the same target numbers ranging from 10 to 70. The identical classifier, that is, the nearest shrunken centroids procedure of Tibshirani et al. [75], is then applied to these subsets of genes to classify patients from the testing set into subgroups. We then use the log-rank test to obtain the significance level at which the subgroups differ in survival time for each method. The results are compared for both methods on the two aforementioned survival datasets. Our motivation for using an identical classifier is that, since our main interest at this stage is to compare the effectiveness and efficiency of the feature selection method used by each approach, we would like to minimize the influences on the results from the bias of the classifiers. The results are shown in Table 4.3 and 4.4, respectively.

Table 4.3: Comparison of p-values from FSCS and FCBF applied to two datasets when the nearest shrunken centroids is used as the identical classifier

Feature Selection Method	p-values of different methods on two datasets	
	Lung Cancer Data	Renal Cell Data
FSCS	0.0476	0.0374
FCBF	0.0107	0.0292

As shown in Table 4.3, both FCBF and FSCS identified lists of genes related to patient survival and made significant statistical survival prediction. Genes such as $S100P$, *crk oncogene* and *VEGF*, to name a few, known to be significantly survival-associated in lung cancer that proven by clinical trials [78], were selected as correlated with patient survival in our study. The best predictions obtained by FCBF are 0.0107 for the lung cancer data with 30 selected genes and 0.0292 for the renal cell data with 25 selected genes. FSCS yields its best p-value (0.0476) for the lung cancer data when 55 genes are identified, and for renal cell data ($p = 0.0374$) when 50 genes are identified. From the collection of p-values over these two datasets, we observe that both FCBF and FSCS require a certain number of selected genes in order to arrive at a significant survival prediction. One possible reason for this is if the number of selected genes is limited, the information they contain is insufficient in return. Also, FCBF method generally performs better than FSCS in cases where the number of selected genes is limited and is able to make significant survival prediction with fewer selected genes, which are consistent with the theoretical analysis demonstrating FCBF's ability to identify redundant features. The fact that selected significant genes are not highly correlated with each other ensures that the information contained is

maximized.

Table 4.4: Time taken (CPU units) by the FSCS and FCBF for a single trial on each dataset

Feature Selection Method	Dataset	
	Lung Cancer Data	Renal Cell Data
FSCS	26.0	37.0
FCBF	3.0	8.0

Table 4.4 shows the execution times reported in CPU units on "aalen" for FSCS and FCBF, and it is clear that FCBF runs significantly faster than FSCS. The FSCS method consistently identifies significant genes by iteratively fitting a Cox model for each gene. This can incur a relatively high cost in high-dimensional settings. On the contrary, FCBF only calculates SU values to identify significant and redundant genes, which are a linear time complexity and $O(NMlogN)$ in terms of the number of instances M and genes N, respectively. By identifying and removing the remaining genes that are redundant peers to already identified genes in each iteration, the whole process is also greatly accelerated. The results given in Table 4.4 verify FCBF's superior computational efficiency, which allows it to scale to larger datasets.

These experimental results suggest that feature selection in machine learning is feasible for the analysis of high-dimensional survival data. It can effectively and efficiently identify features truly associated with the phenotype of interest and can enhance prediction with significant features.

4.1.4.3 Comparison When Different Classifiers are Used

In order to compare feature selection in machine learning with feature selection from a statistical perspective, we follow the idea of Bair's semi-supervised approach to cope with high-dimensional survival data and harness the identical classifier (i.e., the nearest shrunken centroids method), thus making feature selection methods the only distinction between the two. Indeed, other than the nearest shrunken centroids method, alternative learning algorithms from the machine learning community can be paired with FCBF to improve the survival prediction power. For each dataset, we run the four previously discussed statistical approaches (see Chapter 2) along with our proposed procedure with the identical feature selection method FCBF and two typical classifiers, NB and DT. The log-rank statistics comparing the survival time of different subgroups obtained from different methods are recorded for comparison.

The Clustering-Cox method uses FSCS to compile a list of genes that correlate with survival time and applies the nearest shrunken centroids to classify testing data into the identified subgroups that fit with the training data. The list of top 55 genes has the best overall association with survival, with a p-value held at 0.0476. The Risk Index of the top 60 genes with the 60th percentile as a cutoff point identifies future patients with a p-value of 0.0402. Our proposed method (FCBF in conjunction with NB) yields a p-value of 0.0071 by selecting the 40 most significant genes. The FCBF combined with DT has a p-value of 0.017 using the top 25 genes. In addition, Median-Cut and Hierarchical Clustering are also employed, as representative supervised and unsupervised approaches, respectively, to distinguish patients into subgroups. The results of these methods applied to the lung cancer data and renal cell data are shown

in Table 4.5.

Table 4.5: Comparison of different methods applied to two datasets. Median-Cut, using median survival time to assign patients into cancer subtypes; Hierarchical Clustering, using a clustering dendrogram to assign subtypes of patients based on all genes; Clustering-Cox, using clustering based on the genes with the largest Cox scores; Risk Index, using the cumulative effects of the significant genes selected with the largest Cox scores; Naïve Bayes, using FCBF in conjunction with Naïve Bayes classifier; and Decision Tree, using FCBF in conjunction with Decision Tree classifier.

Method	p-values of different methods on two datasets	
	Lung Cancer Data	Renal Cell Data
Median-Cut	0.0287	0.0523
Hierarchical Clustering	0.078	0.0479
Clustering-Cox	0.0476	0.0374
Risk Index	0.0402	0.0489
Naïve Bayes	0.0071	0.0106
Decision Tree	0.017	0.0247

Table 4.5 shows that our proposed approach for high-dimensional survival data is predictive of survival and performs better than established methods. However, the predictive results might differ in a single training-testing set due to random sampling issues. Therefore, the leave one out cross-validation method is used to further validate and generalize the results. Specifically, we first identify underlying subtypes and train the classifier in each training set, which includes all patients except for the one that

is held out from the full set as a test case. We then apply the trained classifier to the single one test case and determine which subtype is present in the hold out patient. Log-rank statistics comparing the survival times of different subgroups in the test cases are computed to compare the effectiveness of different methods. The efficiencies are indicated by the time needed to complete the whole leave one out trial on each dataset with respect to a number of selected genes in each method. These two results are shown in Table 4.6 and 4.7, respectively.

Table 4.6: Comparison of the leave one out approach of different methods applied to two datasets. Median-Cut, using median survival time to assign patients into cancer subtypes; Hierarchical Clustering, using clustering dendrogram to assign subtypes of patients based on all genes; Clustering-Cox, using clustering based on the genes with the largest Cox scores; Risk Index, using the cumulative effects of the significant genes selected with the largest Cox scores; Naïve Bayes, using FCBF in conjunction with Naïve Bayes classifier; and Decision Tree, using FCBF in conjunction with Decision Tree classifier.

Method	Dataset	
	Lung Cancer Data	Renal Cell Data
Median-Cut	0.0285	0.054
Hierarchical Clustering	0.069	0.0461
Clustering-Cox	0.0113	0.0098
Risk Index	0.0072	0.0121
Naïve Bayes	0.0031	0.0057
Decision Tree	0.0085	0.0096

The evaluation of Table 4.6 verifies that the proposed approach is effective for high-dimensional survival data. It is capable of identifying subtypes of cancer and can enhance the use of this knowledge to diagnose future patients. From Table 4.6, we can see that our method made a much more significant statistical survival prediction than other methods. In particular, Naïve Bayes in conjunction with FCBF gives the best results when 33 of the most significant genes are identified among lung cancer data and 38 genes among renal cell data, with p-values of 0.0031 and 0.0057, respectively. These results are significant predictors of survival. All methods in general undergo a major improvement in their ability to predict patient survival expect for Median-Cut; this is not surprising because the robustness of Median-Cut is mainly affected by the degree of overlap between different groups rather than random sampling issues. Therefore, although the leave one out approach provides more valuable information with minimized random sampling issues, the performance of the Median-Cut method is not necessarily enhanced. On the contrary, the performance improvement of other methods agrees with the much more detailed sampling information provided by the designed leave one out approach and the theory of how these methods work. Overall, the enhanced leave one out approach depicts the robustness of our proposed method and is a significant predictor of survival.

The results of Table 4.7 provide evidence regarding the efficiencies of different methods. Median-Cut is extremely efficient because it only requires the calculation of median time and can be trained at little expense. Therefore, although the leave one out method is very expensive to compute, Median-Cut still runs in very little time compared with other methods. The high cost of Hierarchical Clustering is primarily attributed to calculating inter-group distance, as it does not involve feature selection,

Table 4.7: Time taken (CPU units) by different methods for completing a leave one out trial on each dataset with specific to a certain number of selected genes.

Method	Dataset	
	Lung Cancer Data	Renal Cell Data
Median-Cut	< 1.0	< 1.0
Hierarchical Clustering	82.0	297.0
Clustering-Cox	2241.0	7594.1
Risk Index	2417.0	8016.0
Naïve Bayes	263.0	1476.0
Decision Tree	371.0	1730.0

and the classification step is also highly efficient. For the other four methods, different classifiers do make a difference in the time needed to complete a leave one out trial on each dataset for a certain number of selected genes, but it is the feature selection that dictates the time variances. The running times of these four methods over the datasets are in accordance with our previous time analysis shown in Table 4.4. However, since each iteration of the leave one out method uses feature selection once, the running time differences are additive. The larger the size of the sample and the greater number of features contained in the sample (especially redundant ones), the greater the difference between methods that employ FCBF and FSCS. Modern biomedical technology has caused an explosion of data, which severs the small n large p problem. Although predictive of survival, the Clustering-Cox and Risk Index methods are time consuming. Therefore, improvements in the computational efficiencies of these methods are warranted. Our proposed method merits with good scalability and is an

efficient predictor of survival.

4.2 Experiments on Simulation Data

In this section, we carry out Monte Carlo simulations to evaluate our approach and compare its performance to statistical methods.

4.2.1 Simulation Data Generation

In addition to the two real-world datasets, simulation data was also used to test the effectiveness of the proposed method. Simulation imitates a real-world process without the need to carry out a pilot test while it permitting a sufficient understanding of the process and maximizing the benefits of limited resources. We designed a data generation algorithm to specifically address survival time data. In particular, we used a logistic distribution and Weibull distribution as examples to model covariates' effect on the patient's probability of being cured and the survival probability of uncured patients, respectively. The algorithm can be extended to generate other generic data by employing the desired distributions and using any other proper survivor function for which the inverse function is well defined. The data generation algorithm is outlined as follows:

For predefined values of γ, β and p, as well as an identified covariate dimension, d, we generate features X and Z for incidence and latency segments (without loss of generality, we let $Z = X$), survival time t, censoring indicator δ, and the indicator of cured status, y, for each patient.

1. Randomly generate a realization Z or X with the desired dimensions, where

each dimension is from a Bernoulli distribution with success parameter, p.

2. Randomly generate a realization censoring time $t_{censoring}$ from a desired distribution.

3. Compute the probability of being cured $p(z_i)$, where $p(z_i) = {\exp(\gamma * \mathrm{z})}/{1 + \exp(\gamma * \mathrm{z})}$. Randomly generate a realization, y , from a Bernoulli distribution with success parameter $p(z_i)$. If $y = 1$, then set $\delta = 0$, $t_{event} = +\infty$ and skip to step 7; otherwise proceed in a stepwise fashion.

4. Compute $\lambda = \exp(\beta * \mathrm{x})$.

5. Randomly generate a realization s_0 from a desired distribution.

6. Compute $t_{event} = {(-\log(s_0))}/{(-\lambda)}$

7. Obtain the survival time $t = \min(t_{event},\ t_{censoring})$

8. When $y = 0$, if $t_{event} = \min(t_{event},\ t_{censoring})$, then set $\delta = 1$; otherwise δ =0.

Repetition n times of steps 1-8 results in a sample size of n. By selecting different predefined values, we generate clinical data with expected settings to conduct Monte Carlo simulations in the following section.

4.2.2 Results and Comparison with Statistical Methods

We first follow the data generation algorithm, proposed in the previous section, to generate clinical data. The simulated clinical dataset consists of 10 covariates by setting d =10 with γ = (0.5, 0, -0.6, 0.7, 0, 0, 0.3, 0, 0, 0.75, -0.75) and β = (-0.5, 0, 0.75, -0.15, 0, 0, -0.5, 0, 0.3, 0.3, -0.1), and 100 observations. Although we pay

little attention to the intercept, it is part of the logistic regression and has the ability to dominate the cure rate in simulation study. Therefore, the dimensions of both parameters γ and β are 11 instead of 10, with the first dimension being the intercept. Also, the number of nonzero coefficients, cure rate and censoring rate correspond to the setting of lung cancer data. In terms of the setting of survival time, the censoring time of each sample is generated as a uniform random number with a minimum value of 2 and a maximum value of 16. The event time is generated with a value ranging from 8 to 16 for samples 1-50, and 2 to 10 as event times for samples 51-100. Both sets of event times are computed from s_0, which follows the uniform distribution as well.

Next, we generate gene expression data. By adjusting to the characteristics of the gene expression data, different distributions are employed. The number of genes is set as 5,000, which is close to the number of genes after the preprocessing step in evaluating the lung cancer data. Rather than a Bernoulli distribution, all gene expression values are generated as standard normal random numbers with a few exceptions: a mean of 1.0 in genes 1-50 is generated for 30% randomly selected samples, a mean of 2.0 in genes 51-200 is generated for 50% randomly selected samples, and a mean of 0.5 in genes 200-400 is generated for 70% randomly selected samples.

We have now generated the clinical data and gene expression data for training. We define samples 1-50 and 51-100 as belonging to "cancer subtype 1" and "cancer subtype 2", respectively. Finally, the program runs again with exactly the same parameter settings to generate testing data. We evaluate the performance of the penalized selection method by calculating the nonzero coefficients of the generated clinical data. Performances of the other methods discussed are compared by applying

them in the identification and determination of underlying subtypes in the training and testing sets. Initial cluster errors are the number of misclassified samples when the training data is originally divided into two subgroups, and prediction errors are how many samples are mistakenly assigned subtypes in the testing set. Table 4.8 and 4.9 show the simulation results based on 100 replications.

Table 4.8: Simulation result of selecting significant parameters

Clinical	**Logistic Part**		**Survival Part**	
Paramters	True Value	Estimated Coefficient	True Value	Estimated Coefficient
Feature1	0	0	0	0
Feature2	-0.6	-0.63	0.75	0.7371
Feature3	0.7	0.71	-0.15	-0.1415
Feature4	0	0	0	0
Feature5	0	0	0	0
Feature6	0.3	0.3605	-0.5	-0.4712
Feature7	0	0	0	0
Feature8	0	0	0.3	0.3257
Feature9	-0.75	-0.7212	0.3	0.2975
Feature10	-0.75	-0.7487	-0.1	-0.147

In Table 4.8, we see that the Lasso-penalized method successfully selects the nonzero coefficients of γ in the logistic regression part as well as correctly points out the nonzero coefficients of β in the survival regression part. Table 4.9 shows that the

Table 4.9: Comparison of different methods applied to simulation data. Median-Cut, using median survival time to assign patients into cancer subtypes; Hierarchical Clustering, using clustering dendrogram to assign subtypes of patients based on all genes; Clustering-Cox, using clustering based on the genes with the largest Cox scores; Risk Index, using the cumulative effects of the significant genes; Naïve Bayes, using FCBF in conjunction with Naïve Bayes classifier; Decision Tree, using FCBF in conjunction with Decision Tree classifier.

Method	Initial Cluster Errors	Prediction Errors
Median-Cut	36.6	53.8
Hierarchical Clustering	37.8	37.2
Clustering-Cox	20.3	16.4
Risk Index	15.5	12.5
Naïve Bayes	7.2	4.1
Decision Tree	10.7	12.6

fully supervised and fully unsupervised methods performed much more poorly than other methods in this simulation study. The NB method gave the best results, with both the fewest initial cluster errors and prediction errors. The DT method was the second best in terms of initial cluster errors, and it performed slightly worse than the Risk Index method with respect to prediction errors. Overall, both NB and DT made good predictions. The simulation results are consistent with the results from the real dataset and show the feasibility of the machine learning approach.

Along with the experiments on real datasets, these studies provide evidence regarding the positive impact of the proposed machine learning approach to identify survival-associated features and to predict patient survival from high-dimensional survival data in an effective and efficient manner. More importantly, this work can be considered as a preliminary study towards shaping a new research direction. More detailed studies employing a broader range of various machine learning approaches for the development of more powerful diagnostic tools for cancer are anticipated.

Chapter 5

Conclusions and Future Work

The main objective of this research was to incorporate and promote the machine learning approach to predict patient survival from high-dimensional survival data. To accomplish this objective, the discretization of patient survival times via silhouettes clustering validity was used as a strategy to overcome the obstacle of hidden class information in high-dimensional survival data. Class discovery allowed feature selection in machine learning to identify features truly associated with survival. Classifiers were then applied to a subset of selected features to predict subtypes for future groups of patients; therefore, the prediction was based on survival-associated information contained in the high-dimensional survival data. The leave one out method validated and enhanced the identification and predication performance (Chapter 3). Finally, an empirical study was conducted on real datasets, as well as simulation datasets, to test the ability of the proposed method to predict patient survival from high-dimensional survival data (Chapter 4). The primary contributions of this research as well as the potential future directions are outlined in the reminder

of this chapter.

5.1 Research Contributions

Although many studies have focused on developing statistical methods to select survival-associated features and to predict cancer subtypes from the genetic profile of a tumor, research in this area from a machine learning perspective has been underexplored [4, 22, 11]. Therefore, the main contribution of this research was the design and development of a novel machine learning approach that utilizes both gene expression data and clinical data to discover and predict cancer subtypes. Our procedure employed the delicate discretization method on event times to enable the use of feature selection in machine learning. A strategy was found that makes it possible to choose from the rich repository of feature selection and classification methods proposed by computer scientists in the machine learning community to select survival-associated features and to predict cancer subtypes from high-dimensional survival data.

A fundamental research question in applying such a procedure was, *how can machine learning approaches be applied to predict patient survival from high-dimensional survival data*? Although feature selection in machine learning can identify relevant features as well as redundancy among them in an efficient manner, the application of machine learning methods in the analysis of high-dimensional survival data is underexplored. The main obstacle encountered in using these methods for high-dimensional survival data is the lack of explicit class labels in the training set. Unlike statistical approaches that use the Cox score to quantify how well each feature predicts

survival, which is independent of the availability of a class label, feature selection in machine learning is dependent on class label. Therefore, we chose to apply the delicate discretization method on the survival time of patients to find the hidden class label.

In contrast to classical statistical methods that keep censoring individuals in the risk set along with other individuals who have not yet experienced the event, we use the k-nearest neighbor to address censoring, which is based on clinical parameters that are truly associated with survival time. In order to validate the chosen design and to enhance identification and prediction performance, a leave one out method was employed. This research opens a range of possibilities for future work on selecting survival-associated genes and identifying cancer subtypes from a different research direction.

After designing and developing this approach, a second research question emerged: *What is the feasibility of using the machine learning approach for high-dimensional survival data?* To answer this question, an empirical study was conducted to compare the performance of our proposed method on high-dimensional survival data with related methods in the literature using publicly available datasets and simulated datasets. The results of the real datasets and the simulation datasets conclusively match, which suggests that our proposed method is an effective and efficient predictor of survival. Our proposed approach is capable of selecting features truly associated with survival and enhancing prediction with significant features. The subgroups determined in the test cases differed significantly in their overall survival, with p-values of 0.0031 and 0.0057 for lung cancer data and renal cell data, respectively. Our method proved to be superior to the others with both the fewest initial cluster errors and prediction errors in the simulation study. With respect to efficiency, our

method demonstrated promising capability in dealing with high-dimensional survival data. It is significantly faster in selecting a certain number of significant genes, especially when the leave one out method was used; its computational efficiency had a fuller exposure that is 10 and 5 times faster for lung cancer data and renal cell data, respectively.

As discussed in Chapter 4, these findings indicate that the machine learning approach is feasible for the analysis of high-dimensional survival data. To the best of our knowledge, no other work has explored this approach in a systematic way. As such, our work can serve as a complementary paradigm for classical survival prediction in high-dimensional space. More research along this direction is warranted. Some potential topics for future work are discussed in the next section.

5.2 Future Work

The results of the empirical study reveal the ability of the proposed strategy to select survival-associated features and to predict patient survival from high-dimensional survival data. This indicates that there is an opportunity to discover the hidden class label based on other clinical information, such as the stage of the tumor, or whether it has metastasized. For example, information about the risk of metastasis for a given patient is essential for the design of more targeted treatment strategies [11]. If the risk of metastasis is high, the cancer must be treated aggressively, even if serious toxic effects are likely; on the other hand, a milder regimen can be administered to patients with a low risk of metastasis. Therefore, the clinical information of metastasis can be regarded as the phenotype of interest and used to find the hidden class label.

Similarly, genes can be selected based on their association with the risk of metastasis, and this knowledge can be used to identify subgroups by their differences in this respect. Tusher et al. described methods for selecting a phenotype of interest from a variety of possible clinical variables [78]. An appealing research direction is to find the hidden class label based on multiple phenotypes of interest and the interactions between them [71]. Furthermore, in this research, we adopt the methods largely based on their performance in classical machine learning applications, which are not necessarily the best in the context of survival prediction. What criterion we should use in selecting an approach for survival prediction from the very rich repository of the machine learning methods is an interesting topic to explore [7].

There is value in choosing features with the help of purported biological knowledge, despite the fact that this approach is irrespective of this knowledge. For instance, genes involved in specific biological pathways or those that have an established involvement in the disease process under study can be selected in advance to aid in the selection of phenotype-associated genes [78, 71], which may lead to more accurate subtype predictions. In order to expand the analysis of high-dimensional survival data, there must be a significant attempt to include interdisciplinary collaborations [55]. Moreover, even with ways of selecting phenotype-associated features, how to identify environmental factors that increase the risk of cancer is one of the greatest challenges of the research agenda [53].

In this research, the proposed method was applied and evaluated on two real-world datasets that were obtained from completed studies. In order to analyze high-dimensional survival data and assess the performance of our proposed method on datasets obtained from ongoing studies in the future (data become available only

at certain points in time, as in clinical decision making), a sequential classification scheme could be applied [77]. This way, patients are classified only upon sufficient evidence and the repeated use of additional data.

Bibliography

[1] A. Alizadeh et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[2] D. G. Altman and J. M. Bland. Interaction revisited: the difference between two estimates. *British Medical Journal*, 326(7382):219, 2003.

[3] E. Bair and R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLOS Biology*, 2(4):511–522, 2004.

[4] D. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, pages 781–791, 2006.

[5] D. Beer, S. Kardia, C. Huang, T. Giordano, A. Levin, et al. Gene expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 8(8):816–824, 2002.

[6] C. Begg. A strategy for distinguishing optimal cancer subtypes. *International Journal of Cancer*, 129(4):931–937, 2011.

[7] R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, 77:81–97, 2008.

[8] F. M. Blows, K. E. Driver, M. K. Schmidt, et al. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: A collaborative analysis of data for 10,159 cases from 12 studies. *PLoS Medicine*, 7(5):e1000279, 2010.

[9] J. W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society*, 11(1):15–53, 1949.

[10] M. Borda. *Fundamentals in Information Theory and Coding*. Springer-Verlag Berlin Heidelberg, second edition, 2011.

[11] H. Bovelstad, S. Nygard, H. Storvold, et al. Predicting survival from microarray data - a comparative study. *Bioinformatics*, pages 2080–2087, 2007.

[12] Canadian Cancer Society. Cancer statistics at a glance. Retrieved from: `http://www.cancer.ca/en/cancer-information/cancer-101/cancer-statistics-at-a-glance/?region=on`, 2005.

[13] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2008.

[14] J. Catlett. On changing continuous attributes into ordered discrete attributes. In *Proceedings of the European Working Session on Learning*, pages 164–178, 1991.

[15] I. Choi et al. An empirical approach to model selection through validation for censored survival data. *Journal of Biomedical Informatics*, 44(4):595–606, 2011.

[16] W. Choi et al. Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. *Cancer Cell*, 25(2):152–165, 2014.

[17] E. Christensen. Multivariate survival analysis using Cox's regression model. *Journal of Hepatology*, 7:1346–1358, 1987.

[18] H. Clifford, F. Wessely, S. Pendurthi, and R. D. Emes. Comparison of clustering methods for investigation of genome-wide methylation array data. *Frontiers in Genetics*, 2:88, 2011.

[19] D. Collett. *Modelling survival data in medical research*. Chapman and Hall /CRC, second edition, 2003.

[20] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34:187–220, 1972.

[21] M. Crescenzi and A. Giuliani. The main biological determinants of tumor line taxonomy elucidated by of principal component analysis of microarray data. *FEBS Letters*, 507:114–118, 2001.

[22] X. Cui and G. Churchill. Statistical test for differential expression in cDNA microarray experiments. *Genome Biology*, 4:210, 2003.

[23] C. Curtis, S. P. Shah, C. S. Feung, et al. The genomic and transcriptomic architecture of 2000 breast tumors reveals novel subgroups. *Nature*, pages 346–352, 2012.

[24] M. Dash, K. Choi, P. Scheuermann, and H. Liu. Feature selection for clustering - a filter solution. In *Proceedings of the 2nd IEEE International Conference on Data Mining*, pages 115–122, 2002.

[25] S. Datta and S. Datta. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19.

[26] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous ceatures. In *Proceedings of the 12th international conference on machine learning*, pages 194–202, 1995.

[27] S. Dudoit. *Selected Works of Terry Speed*. Springer New York, 2002.

[28] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. In *Proceedings of the National Academy of Science (USA)*, volume 95, pages 14863–14868, 1998.

[29] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[30] V. T. Farewell. Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14:257–262, 1986.

[31] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous valued attributes for classication learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1027, 1993.

[32] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[33] M. George. Cancer: 100 different diseases. *American Journal of Nursing*, 66(4):749–756, 1966.

[34] A. D. Gordon. *Classification*. Chapman and Hall/ CRC, 1999.

[35] J. W. Grzymala-Busse. Three strategies to rule induction from data with numerical attributes. *Lecture Notes in Computer Science*, 3135:54–62, 2004.

[36] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[37] R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–90, 1993.

[38] D. W. Hosmer and S. Lemeshow. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: Wiley, 1999.

[39] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics, second edition, 2002.

[40] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, pages 457–481, 1958.

[41] L. Kaufman and P. J. Rousseeuw. Clustering by means of medoids. *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pages 405–416, 1987.

[42] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: An introduction to cluster analysis*. Wiley Interscience, 2008.

[43] G. D. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text*. Springer-Verlag New York, third edition, 2012.

[44] R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

[45] S. Kotsiantis and D. Kanellopoulos. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1):47–58, 2006.

[46] P. C. Lambert, P. W. Dickman, C. L. Weston, et al. Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Journal of Applied Statistics*, 59:35–55, 2010.

[47] J. Lapointe, C. Li, E. Bair, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 811–816, Jan. 2004.

[48] J. F. Lawless. *Statistical Models and Methods for Lifetime Data*. John Wiley and Sons, second edition, 2003.

[49] H. Liu, F. Hussain, C. Lim, and M. Dash. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6(4):393–423, 2002.

[50] X. Liu et al. Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. *Statistics in Medicine*, 31:2882–2891, 2012.

[51] P. Lucas. Bayesian analysis, pattern analysis, and data mining in health care. *Current Opinion in Critical Care*, 10:399–403, 2004.

[52] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[53] G. C. Montserrat, N. B. Gunsoy, and N. Chatterjee. Combined associations of genetic and environmental risk factors: Implications for prevention of breast cancer. *Journal of the National Cancer Institute*, 106(11):dju305, 2014.

[54] F. Mortiera, S. Robinb, S. Lassalvy, C. P. Barilc, and A. Bar-Hend. Prediction of euclidean distances with discrete and continuous outcomes. *Journal of Multivariate Analysis*, 97:1799–1814, 2006.

[55] S. Ogino, A. T. Chan, C. S. Fuchs, and E. Giovannucci. Molecular pathological epidemiology of colorectal neoplasia: An emerging transdisciplinary and interdisciplinary field. *Gut*, 60(3):397–411, 2011.

[56] S. Ogino et al. Cancer immunology-analysis of host and tumor factors for personalized medicine. *Nature Reviews Clinical Oncology*, 8(12):711–719, 2011.

[57] S. Ogino, C. S. Fuchs, and E. Giovannucci. How many molecular subtypes? Implications of the unique tumor principle in personalized medicine. *Expert Review of Molecular Diagnostics*, 12(6):621–628, 2012.

[58] U. Pfeffer. *Cancer Genomics: Molecular classification, prognosis and response prediction*. Springer Netherlands, 2012.

[59] S. Pocock, T. C. Clayton, and D. G. Altman. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *Lancet*, pages 1686–1689, 2002.

[60] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical recipes in C*. Cambridge University Press, 1988.

[61] J. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, second edition, 1993.

[62] C. E. Rasmussen and K. I. W. Christopher. *Gaussian processes for machine learning*. The MIT Press, 2006.

[63] A. Reynolds et al. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modeling and Algorithms*, 5:475–504, 1992.

[64] J. Rice and M. Rosenblatt. Estimation of the log survivor function and hazard function. *Sankhya: The Indian Journal of Statistics*, 38(1):60–78, 1976.

[65] L. Rokach and O. Maimon. *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing Company, 2008.

[66] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[67] C. Safran, M. Bloomrosen, W. E. Hammond, et al. Toward a national framework for the secondary use of health data: an american medical informatics association white paper. *Journal of the American Medical Informat5ics Association*, 14:1–9, 2007.

[68] G. I. Salama, M. B. Abdelhalim, and M. A. Zeid. Experimental comparison of classifiers for breast cancer diagnosis. In *Proceedings of the Seventh International Conference on Computer Engineering & Systems*, pages 180–185, 2012.

[69] T. Sørlie, C. M. Perou, R. Tibshirani, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. In *Proceedings of the National Academy of Sciences (USA)*, pages 10869–10874, Sep. 2001.

[70] M. Stephens and D. J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–691, 2009.

[71] Y. Suehiro, C. W. Wong, L. R. Chirieac, et al. Epigenetic-genetic interactions in the apc/wnt, ras/raf, and p53 pathways in colorectal carcinoma. *Clinical Cancer Research*, 14(9):2560–2569, 2008.

[72] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3):281–285, 1999.

[73] R. Tibshirani. The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16:385–395, 1997.

[74] R. Tibshirani. Univariate shrinkage in the cox model for high-dimensional data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):Article 21, 2009.

[75] R. Tibshirani, T. Hastie, B. Narashimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 18:104–117, 2003.

[76] S. Tu. *Origin of cancers: Clinical perspectives and implications of a stem-cell theory of cancer*. Springer US, 2010.

[77] G. Tusch. An optimization model for sequential decision-making applied to risk prediction after liver resection and transplantation. In *Proceedings of AMIA Symposium*, pages 425–429, 1999.

[78] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 5116–5121, 2001.

[79] N. R. C. (US). *Toward Precision Medicine: Building a knowledge network for biomedical research and a new xaxonomy of disease*. National Academies Press (US), 2011.

[80] M. J. Van, Y. D. He, H. Dai, et al. A gene expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 8(347):1999–2009, 2002.

[81] L. van't Veer, H. Dai, M. J. Vijver, Y. D. He, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[82] P. Verweij and V. Houwelingen. Penalized likelihood in cox regression. *Statistics in Medicine*, 13:2427–2436, 1994.

[83] Z. Xu, G. Huang, K. Q. Weinberger, and A. X. Zheng. Gradient boosted feature selection. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 522–531, 2014.

[84] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 12th International Conference on Machine Learning*, pages 856–863, 2003.

[85] H. Zhao, B. Ljungberg, et al. Gene expression profiling predicts survival in conventional renal cell carcinoma. *PLOS Medicine*, 3:e13, 2006.