# REGRESSION MODELS FOR READMISSION PREDICTION USING ELECTRONIC MEDICAL RECORDS

by

**YUDI NIU**

**THESIS**

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

**MASTER OF SCIENCE**

2013

MAJOR: COMPUTER SCIENCE

Approved by:

____________________________________

Advisor Date

# DEDICATION

*to*

*my sincere love Rong Cao*

# ACKNOWLEDGMENTS

I wish to give my special thanks to my father Haijun Niu, and my mother Hua Guo for their continuous support and encouragement throughout my life, and without whom I would be unable to pursue my higher studies.

There are several people without whom this thesis would not be possible. I would like to express my gratitude to Dr. Chandan Reddy, my advisor, for his valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously is very much appreciated.

I would like to express my deep gratitude to Dr. Zaki Malik and Dr. David Lanfear for serving on my thesis defense committee, and for their careful reading of the dissertation and for providing useful comments. I would also extend my thanks to other Professors who taught me at Wayne State: Dr. Ming Dong, Dr.Narendra Goel, Dr. Jing Hua and Dr.Nathan Fisher. All of them provided me with the right exposure and the perfect training for me to complete my graduate courses.

I owe so much thanks to my friends Bhanu and Yang who supported me all the time. I also want to convey my thanks to Mr and Mrs Smith, Joshua, Matthew, Daniel and Katherine for continuously guiding me in all my endeavours.

And last but not the least, I am grateful for all the help of my labmates: Yan, Rajul, Vineeth, Rajiur and Zeyad. My thanks to everyone who has provided support and advice in any aspect in accomplishing this thesis.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

The rate of 30-day readmission after discharge from the hospital is very high. According to *Medicare* in 2004, the estimated cost for readmission is $17.5 billion dollars. Nearly 20% of *Medicare* beneficiaries discharged from a hospital are readmitted within 30 days, which not only increases billions of expenses each year but also threatens people's lives. Readmission could be prevented to some extent if a patient of high risk of readmission is well predicted and taken good care of.

Many patients are readmitted very quickly after discharge for the reason that the seriousness of a patient's illness is not fully realized. 30-day readmission is a significant factor to determine the health care performance and it has become a widely accepted quality performance metric for the patient [34]. It indicates that either the hospital fails to adequately address the patients' health issue or the patients are discharged from the hospital prematurely. Either of these conditions could be avoided if the patients are examined regularly using some risk prediction techniques.

Heart Failure (HF) is one of the most common syndromes that results in 30-day readmission. There are more than 5 million people who are suffering from a heart disease in the United States, and that number is increasing by 10% each year. Furthermore, the death rate is rather high. These patients are suffering an estimated one-year mortality from 5% to 75%.

In order to reduce the 30-day readmission rate, several techniques are introduced to predict the patient's readmission probability. Regression and statistical analysis are commonly used for classification of clinical data sets. These two methods are widely accepted by different institutions to predict the 30-day readmission. However, these traditional methods have several shortcomings that limit their predicting ability. When analysing a large number of attributes, the effect of certain important attributes could be ignored due to the lack of domain knowledge or large amount of attributes. On the other hand, these methods do not perform well on special cases due to insufficient statistical information.

A better analysis on the factors contributing to 30-day readmission is needed. To this end, we use data mining techniques to build the classifier. There are four steps in building predictive classification models, which are data understanding, data preprocessing, modeling and evaluation. In our experiment, we use the electronic medical records to build the classifier.

The electronic medical records are obtained from the Henry Ford hospital. When a patient is admitted to the hospital, his information will be recorded, including the basic information of the patient, the procedures the patient has received, the medicines he took and the overall health condition. For each patient, the basic information is recorded and the patient is assigned an ID. Figure 1.1 gives the process on the data is organized in the Henry Ford hospital .

From Figure 1.1 we can see that the process of generating the information of each patient is independent. As a result, the records are different from each other. Some of the patients may have errors in their information due to the examination mistake. All these situations lead to the fact that the clinical data is tough to preprocess, and more methods have to be used for completing the preprocessing. In our study, we use the datasets where numeric and categorical features are present in a same dataset. The distribution of the patient information seldom follows any common distribution, and noise is common to see in the datasets. As a consequence, the clinical data is more complicated, and before applying a data mining model to the data, it is necessary to pre-process the dataset, which is to fill the missing values and take some other procedures to make the

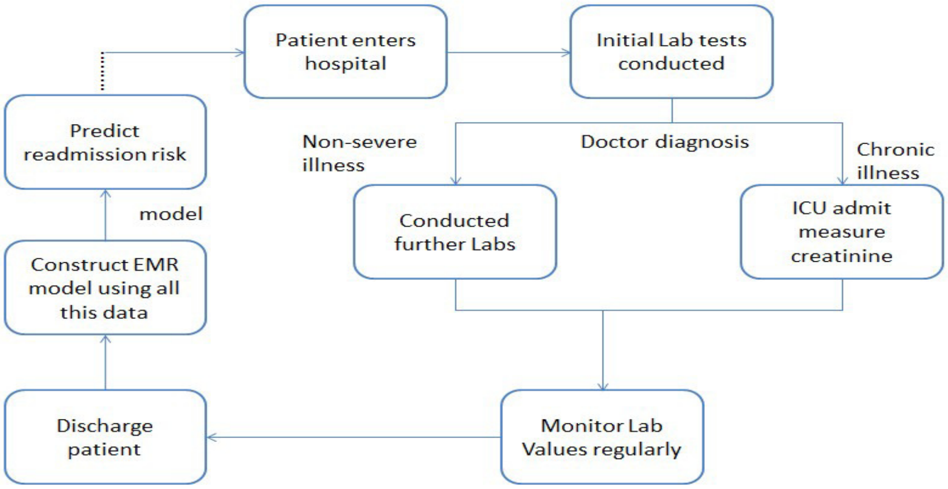


Figure 1.1: Process of Obtaining Real-time Data.

dataset easier to use.

### 1.1.1 Motivation

Readmission very soon after discharge not only brings a huge burden on public finance, but also is a great risk to human health. In order to release the social finance burden and improve the quality of life, it is of great importance to reduce the readmission rate. Our goal is to reduce the 30-day readmission rate. To this end, we need to look into the clinical data and use data mining methods to analyse the factors which contribute most to the 30-day readmission.

By applying proper data mining techniques, the patients who are under a greater risk of readmission can be taken better care of. As a result, the readmission rate can be reduced. The expenditure on 30-day readmission can be largely saved. However, the traditional data mining methods perform bad on predicting 30-day readmission due to the reason that the composition of the clinical data is complex.

### 1.1.2 Overview of our methods

In this thesis, we present two popular methods, SVM (Support Vector Machine) and Cox PH (Proportional Hazard) model, and use them in the clinical scenario. The data for this study is the clinical data which is generated into several different data sources. In each data source, a very diverse patient population with different features are provided. These two methods have good ability in analysing big datasets and provide a more accurate way to analyse the clinical data besides the traditional methods.

The regression model is a statistical technique for estimating the relationship among variables [46]. In general, the variables can be divided in two groups, the independent variable (predictor) and the dependent variable (response), and the response is determined by predictors and their coefficients. The linear regression is widely used in classification. The linear regression provides a single slope or trend so that it is easy to interpret, and it can be fit to unbiased data.

SVM can be used to make prediction using a learning algorithm [15]. The SVM can be applied to different data sources. Before applying SVM to the dataset, the dataset has to be divided into the training dataset and the testing dataset. The training data is used to train the predictive model, while the testing data is used to test the performance of the model. The SVM produces a prediction result with high accuracy, and it also has the ability of analysing the non-linear data by introducing kernel to it. The availability of SVM toolbox makes it simple and convenient to be extended to many fields, such as economy, sociology, medication and psychology.

Survival analysis is a reliable statistic analysis method. In many fields, the survival analysis is applied to obtain the survival results of a certain group of people during a period. In addition to the prediction of 30-day readmission, survival analysis is also widely used in health care problems. In order to do the analysis, it is required to get a patient's censoring time and censoring status.

In this thesis, our major applications of the Cox PH model are the following

1. Compare the survival distribution of different data sources

2. Feature selection using sparse Cox PH models

3. Implement the Cox-LASSO, COCKTAIL (Cox Elastic Net) algorithms.

### 1.1.3 Contribution of this Thesis

The major contribution of this thesis are

- We explore the applicability of using different survival and machine learning algorithms to clinical data.

- We provide a comparison among these methods with respect to different standard metrics such as Area under ROC curve (AUC)

- We explain the importance of clinical variables obtained using shrinkage methods in survival analysis

### 1.1.4 Related Work

In this section, we survey different kinds of important 30 day readmission models. We highlight their importance and also explain their performance on real clinical data.

CMS Medicare conducts a survey on 30-day readmission, which shows the general readmission conditions from 2007 to 2010. The data sources contains 100% Medicare claimed data from 2007 to 2010. This survey is focused on all the inpatient hospital admission that occurs within 30 days of discharging from a previous inpatient hospital admission. The result shows that during 2010, there are 10 million admissions and approximately 1.9 million readmissions. The survey is focused on the factors which are used to divide the population who has readmission.

Table 1.1 shows the readmission rate of each age period. From Table 1.1, we can see

| Age | Readmission Rate | with Readmission |
|---|---|---|
| Age $< 65$ (17%) | 23.1% | 1.8% |
| Age $65 - 74$ (39%) | 17.9% | 3.5% |
| Age $75 - 84$ (29%) | 18.6% | 8.5% |
| Age $> 85$ (15%) | 18.4% | 5.3% |

Table 1.1: Readmission Rate of Different Ages (CMS model)

that people under 65 has the highest risk of readmission. And for all the patients who have readmission, the African-American has a readmission rate of 24.1%. Many patients have even more than one readmissions. These patients are in a great risk of health issue, and should be paid more attention during their next readmission. This statistical results indicates that a specific group of people have a greater possibility of readmission within 30 days. And for those who have more than one readmissions, more careful examine has to be done to prevent the next readmission.

There are several studies [32] sharing some common variables with our datasets. *Amarasingham et al* developed a model in 2010, called ADHERE mortality model [2]. This model uses the features from automated electronic medical records such as blood pressure and sociodemographic factors like age, sex, and drug taking conditions of patients. It is a real-time electronic predictive model that identifies the heart failure patients who are at high risk of readmission. The 1373 heart failure patients are from major urban hospitals between 2007 and 2008. The dataset is composed primarily on patients' electronic medical records. From this dataset we observe that the 30-day readmission rate is 24.1%, compared to the overall 30-day readmission rate of nearly 20%. This group of patients has a higher readmission rate than the average. The ADHERE model reports the outcome with the AUC value between 0.56 and 0.68. The different AUC values suggest that the medical information incorporated with social factors will increase the model's accuracy. In further studies, more social factors should be considered.

Tabak mortality model [45] is a predictive model. The dataset used in this model is the laboratory data of 194,903 admissions, from 2003 to 2009 across 71 hospitals. Demographics, admission-based labs, $ICD-9$ code-based data, vital signs, and altered mental status are all included in this dataset to predict the mortality of heart failure. The distribution of each variable which is in relation to death is examined. This approach allows the model to use all the patients' information when making prediction. This model tries to use laboratory data more and minimally to use the ICD-9 code-based data. The prediction result shows that laboratory data contributes more to the prediction accuracy of mortality rate compared to the other data. Mortality can be well predicted using the proper model and dataset. As a consequence, the 30-day readmission can be well predicted with a proper dataset and model.

Logistic regression is considered as an important method in predicting the 30-day readmission. This regression has a good ability in making prediction with categorical data. A study has been conducted using the data obtained from Centers for Medicare and Medicaid Services (CMS) hospital [17]. The survey is conducted to evaluate the impact of race on 30-day readmission. There are three possibilities in race, which are black, white and others. The dataset used in this survey contains the information of heart disease and social demography information of a patient. This survey compares the result of the dataset with race and without race. The result shows that when race is taken into consideration, the prediction accuracy will be improved, and the black people under competing has the lowest rate of 30-day readmission. This survey shows some good results. However, the limitation of this survey is that the dataset of this survey is unique and only from one hospital, so it is difficult to spread this method.

Cox regression can also be applied to analyse the hospital readmission. A study is conducted to qualify the association between having a clinical culture positive for 1 of 3 prevalent hospital-associated organisms and time to hospital readmission. All data are extracted from the UMMC Clinical Data Repository, which are generated by University of Maryland Medical System Information Technology Group. More than 130,000 patients are included in this study. The data sources contain the patients' administrative, pharmacy, and laboratory information. The result shows that patients with healthcare-associated infections may be at increasing risk of readmission. This result indicates that the readmission rate can be reduced if the hospital has a better infection predicting method.

Meanwhile, Cox regression is also applied to detect the relationship between 30-day readmission and one-year mortality of congestive heart failure patients. The data is collected from the Saint Marys Hospital in Rochester, and more than 15,000 patients are generated in the data sources. Data elements include demographic, clinical, angiographic, procedural, and follow-up variables. The in-hospital performances of the patients are also recorded and used as variables for prediction. These data sources are used in the Cox PH model to detect the effect of 30-day readmission on one-year mortality [34]. In order to get a better result, logistic regression is applied to estimate the association between demographic, clinical, and procedural variables and 30-day readmission. The result shows that nearly 1 in 10 patients have been readmitted within 30 days, and these people have a higher risk of one-year mortality. Also, the study find that the patients who are female and have a higher education level are at the highest risk of one year-mortality among all the patients.

However, the problem with these approaches is that there is no global data that suit for other predictive models. The existing models either perform badly or can only perform well on specific attributes. Several models and ways of generalizing datasets are proven suitable for mortality prediction, and whether they will be suit for 30-day readmission prediction still requires further experiments. The traditional regression model does not perform well on clinical data. In our approach, we introduce new techniques to generalize datasets and make predictions on 30-day readmission and survival analysis.

## 1.2 Uncertain Patient Mining

In order to improve the accuracy of the prediction results, we set up an algorithm called UPAD, which uses SVM to remove the uncertain patients from the data sources so that the prediction accuracy could be improved.

Noise appears in the dataset all the time. It leads to the consequence that even the ideal SVM will make mistake in predicting. The patient whose readmission risk is hard to predict. These patients should be picked up and removed from the dataset, and generalized to a new dataset to do further analysis. To this end, SVM is applied to pick up those patients and find the characteristics of them.

### 1.2.1 Uncertain Patient

In clinical datasets, the patients can be divided into two groups: certain patient (CP) and uncertain patients (UP). Certain patients [26] are the patients whose prediction result has a strong connection with their support vector. Support vector is built by the training model based on the characteristics of the patient. They lie on each side of margin according to the different characteristics of the patients from each class. Hence, it is easy to define CPs as those who can be easily identified and labelled in the right class. This is due to their characteristics being highly identical to those who are labelled the same in the training data.

On the contrary, uncertain patients can be seen as noise [21]. Noise is known as the irrelevant and erroneous data which affects the performance of the classifier negatively. In terms of the support vector notion, uncertain patients can be defined as those who are local outliers with respect to the support vectors near the hyperplane. Our goal is to pick up the UPs who affect the performance of the classifier negatively.

### 1.2.2 UPAD Algorithm

In this section we introduce the Uncertain Patient Detection (UPAD) Algorithm and provide a flowchart for describing it.

In order to build a better training model, the identification of the uncertain patient precisely is needed. A good knowledge about the component of the training model will contribute primarily to picking up the uncertain patients. Hence, before picking up the uncertain patients, we have to look at the training model first. Nowadays there is no existing method that can pick up the uncertain patients automatically, and how to define the uncertain patients largely depends on the model performance. That is to say, different models may pick up different uncertain patients according to their methods of making prediction. As a consequence, before detecting the uncertain patients, a training model has to be built up.

In our approach, the lab data is applied to build the training model. The support vector obtained from the training model can be used to pick up the uncertain patients.

Figure 1.2 shows the process of detecting uncertain patients. In the case that a patient is mislabelled, and the support vector suggests that the patient should be put in the class which he does not belong to, this case is not what we are concerned with. Otherwise, if the support vector suggests the patient to the right class, but he is still mislabelled, that patient is the uncertain patient in the dataset. In the next stage, we will pick up those samples and generate them as a new dataset, and compare the uncertain patient dataset with the lab data. The comparison result shows the internal relationship between the values of each variable and the prediction outcome. Next time when a patient's information is known, it will help to decide if he can be used as a sample to train the model.

The rest of the thesis is organized as follows: In chapter 2, we introduce the SVM algorithm and the toolbox of the LS-SVM in matlab. Chapter 3 explains the Cox PH model

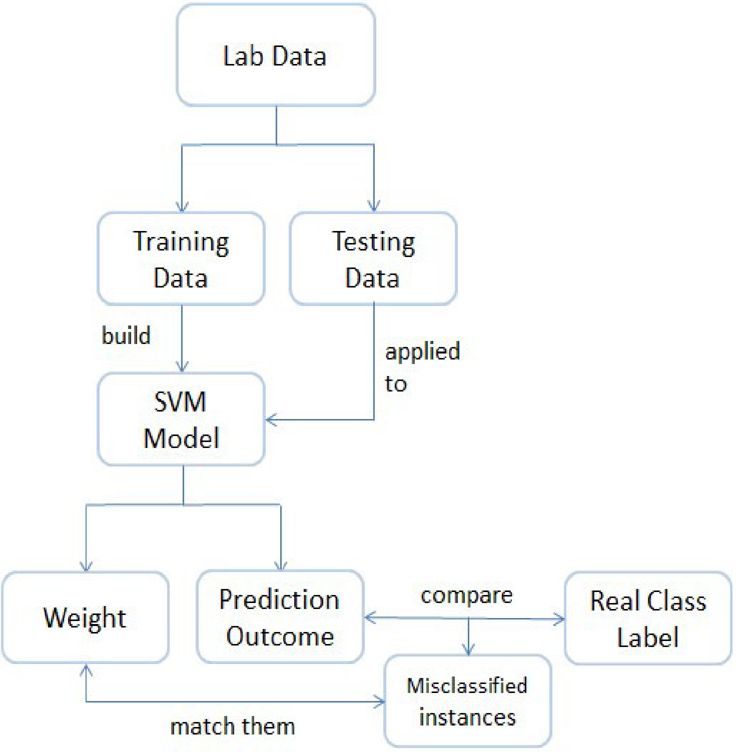


Figure 1.2: Workflow for UPAD algorithm.

and provides several algorithms to optimize the coefficients of input variables. Chapter 4 demonstrates the experimental results and shows the performance comparison of different models. Finally, Chapter 5 concludes the discussion with future research directions.

# CHAPTER 2

# KERNEL BASED METHODS

In this chapter, we will describe the support vector machine and its applications to detecting uncertain patients. The contents of this chapter are as follows. In Section 2.1.1, we give the definition associated with SVM. This is followed by a discussion of the least squares SVM in Section 2.1.2. Later, we give more details about kernels and in Section 2.2 we introduce several popular choices of kernels. Finally in Section 2.3, we give an example of the SVM toolbox and explain how to apply the SVM.

## 2.1 Support Vector Machine

In machine learning, Support Vector Machine (SVM) [41] is a prominent and widely used supervised learning model. This model can perform well on complex data due to its inherent nature of obtaining maximum margin classifiers. Usually, SVMs are used for classification and regression analysis. The basic SVM takes a set of input features together with their associated possible class labels {1,-1} as the training examples. Given some training data D, $D = \{(x_i, y_i) | x_i \in R^{p*1}, y_i \in \{1, -1\}\}_{i=1}^{n}$, where $y_i$ is the class and $x_i$ is the p-dimensional vector.

Once the training model is built, new examples can be classified after they are mapped into the high dimensional space and predicted to belong to a category. The advantages of applying the SVM are:

1. It has a strong ability to manage large input datasets with kernel methods [28].

2. The SVM is robust when dealing with noisy samples.

### 2.1.1 Mathematical Formulation of SVM

The SVM classifier is modeled using a hyperplane [31]. The equation of a hyperplane is given below

$$w \bullet x - b = 0 \tag{2.1}$$

where $x$ is the input dataset, $w$ is the weight vector and $b$ is the bias of the linear model. The hyperplane will divide the space into two parts. These two spaces are defined as

$$w \bullet x - b > 0$$

$$w \bullet x - b < 0$$

A hyperplane is constructed by a SVM in a high dimensional space [1], which is used for classification, regression and other tasks. A good separation can be achieved when the points have the largest distance to the nearest training data point of the other class. In other words, the larger the margin is, the lower will be the generalization error of the classifier. For instance, in Figure 2.1, $H_1$ does not separate the classes. However, one can observe that $H_2$ does separate the classes but with a small margin. Finally,$H_3$ separates the space with the maximum margin.

### 2.1.2 Least Square SVM

Least Square SVM (LS-SVM) [44] is one of the most popular models of SVM. Compared to the basic SVM, LS-SVM provides a better prediction result by applying the least square method, a standard approach to obtain an approximate solution, to determine the Euclidean distance of the vectors and margin [5, 16].

Least Squares SVM (LS-SVM) is formulated in the following way:

$$f(x) = (x, w) + b \tag{2.2}$$

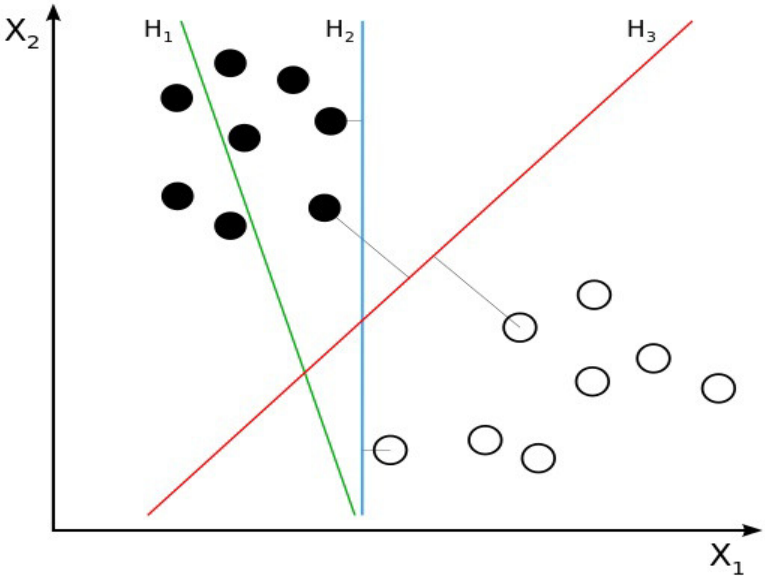


Figure 2.1: Different Hyperplanes Separating the Feature Space.

where $w$ is the weight vector, and $b$ is the bias of the linear model. To set up the LS-SVM, two parameters $w$ and $b$ have to be chosen. We can estimate $w$ and $b$ by minimizing the Lagrange function $L(w, b)$. The Lagrange function can be expressed as

$$L(w, b) = \sum_{i=1}^{n} ||f(x_i) - y_i||_2^2 + C||w||_2^2 \tag{2.3}$$

where in the Lagrange function, $y$ is the vector of class labels, and $C > 0$ is the regularization parameter. The minimization of $L(w, b)$ leads to the optimized value of $w$ and $b$.

LS-SVM works similar to SVM, if $f(x) = (x, w) + b{>}0$, in which the test points, $x$, are either assigned to the positive, or the negative classes. The advantage of using LS-SVM instead of basic SVM are the following:

1. LS-SVM approach makes SVM more generally applicable.

2. LS-SVM reduces the quadratic programming problem of SVM to a linear equation, which largely improves the computational efficiency of the model [52].

3. LS-SVM involves fewer tuning parameters compared to the basic SVM.

### 2.1.3 Advantages of Using SVM

The dataset obtained from distinguished real-world applications might have different types of data. In our study, the clinical data is used, and the dataset contains both numeric and categorical data and the SVM is used to make prediction. The advantage of applying SVMs are given here [3]

1. By introducing kernels to the SVM, it gains flexibility in handling the non-linear dependencies in the data.

2. SVM can provide a good prediction with high accuracy. By choosing the right parameter for the kernels, SVMs can be robust to noisy data.

## 2.2 Kernel Learning

Kernel method is an important data mining technique to deal with the large real-world datasets efficiently [28]. The kernel method is introduced to deal with the databases input to SVM, which will increase the efficiency of the computation and also improve the accuracy of prediction outcome. In this section, we introduce the basic kernel algorithm with several popular kernels.

### 2.2.1 Kernel-based Algorithm

For binary classification problems, given some new input data points $x \in X$, we want to predict the corresponding class $y \in \{\pm 1\}$. The goal here is to find the $y$ for the testing data $(x,y)$. To achieve the goal of finding $y$, the most critical part is measuring the similarity among the input training dataset and the mapping between $X$ and the class label $\{\pm 1\}$. The targets have only two possible categories, thus making it reasonable to measure the similarity of class labels. Nevertheless, to measure the similarity among the input dataset $X$ requires more computation.

For all the $x$, $x^{'} \in X$, the kernel method can be expressed as

$$k(x, x^{'}) = <\Phi(x), \Phi(x^{'})> \tag{2.4}$$

where $\Phi$ maps $x$ into dot product space $H$, which is also known as feature space or Hilbert space [22]. The similarity measure $k$ is called a *kernel*, and $\Phi$ is called its feature map.

The advantage of using kernel to measure the similarity is given here [43]. The data is mapped to the feature space with the implementation of the kernel methods, which simplifies the complex n-dimensional problem. With the help of kernel methods, the complexity of computation can be reduced. For instance, suppose there is a simple classification problem which aims to classify the patients into two groups. A group of patients never readmitted to the hospital is labelled as $\{+1\}$ while the other group which was readmitted, is labelled as $\{-1\}$. Hence, the label should be $Y = \{\pm 1\}$.

The key to separating the different vectors into correct classes is to compute the mean of each class in the feature space, where $c_+ = \frac{1}{n_+} \sum_{\{i:y_i=+1\}} \Phi(x_i)$, and $c_- = \frac{1}{n_-} \sum_{\{i:y_i=-1\}} \Phi(x_i)$ where $n_+$ and $n_-$ are the number of examples with positive and negative target values respectively. The Euclidean distance is calculated between the new patient and the two centroids. The new patient $\Phi(x)$ should be assigned to the class which has a smaller distance between them. That means, if the new patient is closer to $c_+$, then he will be readmitted. The above rule can be written as follows

$$y = sgn(<\Phi(x), c_+> - <\Phi(x), c_-> + b) \tag{2.5}$$

where $b = \frac{1}{2}(||c_-)||^2 - ||c_+||^2$.

Kernel methods are proven to be efficient and robust. We can see that the choice of kernel will play a significant role in building the accurate classification models.

### 2.2.2 Popular Choices of Kernels

In this section, we look at some popular kernels which are widely applied in many fields. These kernels inherit the basic algorithm and develop their own properties. We will now describe several kernels such as *linear kernel, polynomial kernel, Gaussian kernel and RBF kernel.* In the following, $x$ is the input dataset, $y$ stands for the label matrix, and $\alpha$ and $d$ are the parameters of the kernel.

#### Linear Kernel

Linear kernel [29] is the simplest kernel among all the kernel functions. It can be applied to any linear case. However, if the dataset follows a more complex distribution, the linear kernel may not work on it well. The key idea of the linear kernel is that it is given by the inner product $<x, y>$, and in some situations an optional constant $c$ can be added as the bias of the linear model. The linear kernel can be written as

$$k(x, y) = x^T y + c \tag{2.6}$$

Usually the linear kernel is applied to solve simple linear classification problems.

#### Polynomial Kernel

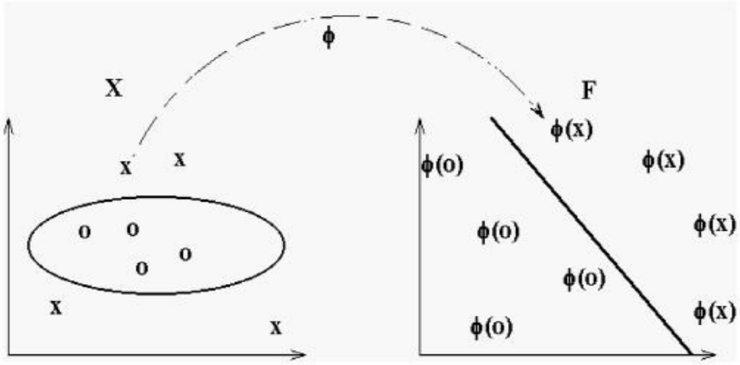
Polynomial kernel is a kernel which has been used in many machine learning applications [18]. In most cases, for a $d$-degree polynomial, the polynomial kernel is defined as follows:

$$K(x, y) = (x^T y + \alpha)^d \tag{2.7}$$

where $x$ and $y$ are the vectors in the input space. This kernel can prevent some types of noise and improve the prediction results.

Figure 2.2 shows an example of how a polynomial kernel is applied to map the

Figure 2.2: Demonstrations of the Kernel Mapping.



data from the original feature space to a high-dimensional space. On the left are the input samples in the original feature space and on the right are the same samples in the high-dimensional kernel space.

### Gaussian Kernel

Gaussian kernel is a widely used kernel [33], which follows Gaussian distribution to measure the similarity between points. The following equation describes the Gaussian kernel

$$k(x_i, x_j) = e^{\frac{-||x_i - x_j||^2}{2\sigma^2}} \tag{2.8}$$

where $\sigma$ is the variance and will determine the width of the Gaussian kernel. To have the optimal performance of the kernel, the value of $\sigma$ needs to be carefully decided.

The Gaussian kernel is a non-linear kernel. It can deal with more complicated non-linear problems. The Gaussian kernel provides an efficient way when an approximate result is required.

### RBF Kernel

RBF kernel [10] is a very common kernel used inside SVM. RBF stands for Radial Basis Function, and is based on the Gaussian distribution. The expression of RBF kernel can be described as

$$K(x_i, x_j) = e^{-\frac{||x||^2}{2\sigma^2}} \tag{2.9}$$

where $x$ is the input for testing data points and $\sigma$ is the parameter to determine the area under influence, which can be scaled to adapt to the distribution. This kernel computes the Euclidean distance between the two data points, and the center of the influenced area will be the support vector.

The performance, or the influence area of RBF kernel highly depends on the value of $\sigma$. In 3-dimensional space, if a RBF kernel has a small $\sigma$, there will be some breakout and peaks on the surface which will be affect the performance negatively. A RBF kernel has an appropriate $\sigma$, the decision surface will be smooth, and there will be no sudden breakout on the surface. Then the performance of the kernel will be improved.

## 2.3 SVM Toolbox

In previous sections we described SVM and kernel method, which have proven to be efficient methods to deal with high-dimensional data. In this section, we will look into the platform that implements the SVM in a real-world project.

SVM is very efficient for classification because of its robustness to noise and overfitting. The application of the SVM is in different fields such as machine learning, data mining [9]. Many toolboxes for SVM are designed for different programming languages, such as Java, C++, Python, Matlab [40] and so on. We use the LS-SVM toolbox which has different functions for using SVM kernels in applications, such as classification and multiple kernel learning.

The LS-SVM toolbox [8] is one of the most popular toolboxes for SVM in Matlab.

It contains different functions of SVMs and kernels, such as linear kernel, polynomial kernel and RBF kernel. For the different input data, users can fix the parameter to get a better training model and well separated support vectors, which will make the new input testing data separated to each side of the margin as accurate as possible according to the support vector. The design of the interface for the toolbox allows one to manually decide the parameter, hence the model possesses great flexibility.

Our kernel experiments are performed using the LS-SVM toolbox. We apply 10-fold cross validation, which is a technique for estimating the performance of a predictive model, for the task of prediction. The toolbox accepts the data and gives a promising prediction outcome. Different datasets with varied parameters have been applied to several kernel and SVM models in this toolbox. The experimental results are given in Chapter 4.

# CHAPTER 3

# COX PROPORTIONAL HAZARDS MODEL

In this chapter, we introduce the basic Cox proportional Hazards (PH) regression model and explain its formulation. In Section 3.1, we explain the notions of survival times and censored status along with providing the equations for Cox Proportional Hazard model. In Section 3.2, we explain the properties of Cox PH model and give some examples. In Section 3.3, we delve into the corresponding regularization methods that can be used to obtain the optimized coefficients, and provide examples of implementation of these methods.

## 3.1 The Cox Proportional Hazard Model

The Cox Proportional Hazard model is a prediction model that is widely used by biostatistic researchers for modeling survival analysis [14]. It can predict the survival time of each individual according to their predictors. Suppose that a dataset consists of $n$ individuals, say $x_1$, $x_2$,..., $x_n$. The time to event is represented as the survival time which measures the time for the event to occur. The censored status $\delta$ is the set for each patient which is 1 if the event of interest has occurred in the given time frame, and it is 0 if the patient drops out. For each individual, the task is to assess the relationship between the time to event and the predictor variables. To this end, we introduce the Cox Proportional Hazard model (Cox PH model), which is given by

$$h(t, X) = h_0(t) \bullet exp(\sum_{j=1}^{p} x_{ij}\beta_j) \tag{3.1}$$

where $h_0(t)$ is called the baseline function, the set $X$=($x_1$, $x_2$, ..., $x_n$) is the set of $n$ feature vectors, and $\beta_j$ is the coefficient for the $j^{th}$ predictor variable. The baseline function is an unspecified function, so that the Cox PH model is a semi-parametric model.

The advantage of being a semi-parametric model such as the exponential model over other models can be understood through the formulation. The advantage compared to the fully parametric model is that it does not ask for user opinion on developing the baseline function, so that it makes no assumption about the shape of the baseline function. It means that the baseline function only depends on the survival time and the status from a dataset. This will make the Cox PH model more robust than other parametric regression models. While at the same time, the Cox PH model also enables the user to control several covariates. For example, the coefficient of each predictor could be decided before applying the model. This is an advantage compared to other non-parametric methods.

The Cox PH model has several properties:

1. The baseline hazard, $h_0(t)$, does not depend on $X$ but only on the time to event. When the baseline model is determined, it only changes with time, and is independent of the predictor variable entered. Consequently, when it works with the clinical data, no assumption about the distribution of the baseline is needed once the time distribution is determined.

2. The predictor set $X$ is time independent. That is, the predictor variables have nothing to do with the time distribution.

In order to have a better understanding of the Cox PH model, we will now discuss some of the related concepts of the model and more details about applying the model to the clinical data. Hazard Ratio (HR) is the ratio which is used to describe the corresponding hazard rates between two levels of a predictor variable [42]. It is widely used to obtain the results in clinical trials involving survival data. When used in survival analysis, it should be noted that hazard ratio does not provide guidance to avoid the death rate or prolong living time, but only reflects the corresponding time between the censored event and survival time. A hazard is the rate indicating the probability of an event happening in a short time period. The hazard could be varied with time.

The PH assumption is a method to measure whether a problem can be solved with the PH model. It requires the hazard ratio for each individual to be proportional to the hazard for other individuals. Intuitively, when we apply a dataset to the PH model and draw the graph of the hazards for each patient, the hazards for different individuals will not cross. If the hazards cross at some point, that means that the PH assumption can not be met, indicating that the Cox PH model is inappropriate for this dataset. In Table 3.1 we present an example to show how the time dependent model works to estimate the survival time and censored status.

## 3.2 Cox Regression

The log partial likelihood is defined as

$$l(\beta) = -\sum_{i=1}^{n} \delta_i x_i^T \beta - \delta_i log(\sum_{j \in R_i} exp(x_j^T \beta)) \tag{3.2}$$

where $R_i$ is the set of all the individuals in the dataset whose survival time is greater than or equal to a survival time of the $i^{th}$ patient [12]. The utilization of partial log-likelihood is valid and when no ties between the survival times exist in the given dataset. If the ties exist in the dataset, two extra approximation methods are needed to optimize the partial log-likelihood [11]. These methods make the partial log-likelihood robust under different conditions.

The Cox PH model is efficient at incorporating the effect of any factor on a time-to-event outcome. The survival prediction outcome of the Cox PH model could be affected by the input covariates, and the average effects on the prediction outcome can be quantified by the hazard ratio.

We now provide an example where we apply the Cox PH regression model on one of the famous datasets in Cox hazards literature, which is called the *Acute Myelogenous Leukemia* (AML) dataset, to do the survival analysis. This dataset has the information of whether a patient keeps undergoing chemotherapy, the information of one's censoring time and one's censoring status.

The Cox PH model is given as $h_0(t)exp(\beta x_i)$, where $h_0(t)$ is the baseline hazard. The task is to find the optimized $\hat{\beta_c}$, which is the notation of the Cox partial likelihood estimator for $\beta$. We apply the Cox regression to the AML dataset. This dataset has the information of whether a patient keeps undergoing chemotherapy, the information of one's censoring time, and one's censoring status. The dataset has 23 patients with 6 features in total. Here is a description of each feature **Time**: survival or censoring time; **Status**: censoring status, **X**: maintenance chemotherapy given, **age**: the age of the patients; **race**: the race of the patients; **prior**: number of prior convictions. The Cox regression gives the coefficient, z-value and p-value of each variable, and it produces the value of partial likelihood. Table 3.1 shows the output

The maximum likelihood ratio obtained from this dataset is 8.48. From Table 3.2 it is obvious to see that each variable has a different coefficient, based on the different $z - value$ of each variable. This indicates each variable has a different weight in the Cox regression model. This dataset focus on the effect of whether a patient is given maintenance chemotherapy on survival situation, so that $X$ is given the highest coefficient, and it should carry on the most weight in predicting the survival situations. The variables are assigned with the coefficients according to the correlation between them and the prediction result. However, in some datasets, some variables contribute little or even bring negative effect to the prediction result. Those variables should be removed or give

Table 3.1: Example of Cox regression

| Variable | coef | exp(coef) | se(coef) | z | p |
|---|---|---|---|---|---|
| X | 0.916 | 2.5 | 0.512 | 1.79 | 0.074 |
| age | -0.064 | 0.833 | 0.036 | -3.13 | 0.011 |
| race | 0.412 | 1.432 | 0.216 | 1.75 | 0.281 |
| prio | 0.133 | 1.287 | 0.324 | 4.299 | 0.001 |

a small coefficient to reduce their influence to the prediction results.

The Cox regression can be applied in many different fields. We explain the different kinds of scenarios where Cox regression has been used effectively: (1)Earthquake study [7]: the hazard function can be used in earthquake forecasting. The information of earthquake is provided to the Cox regression model, and it gives the probability if an earthquake will happen in the coming future; (2)Clinical data analysis: the most famous application of Cox regression. The patients are censored in this study, and several medical information are provided so that the Cox regression can predict the death rate of these patients; (3)Economic study [38] the hazard function can be applied to detect the unemployment duration. The economic factors are applied as the variables and the unemployment is used as the censored status. This study will report the unemployment in the coming future for certain places;(4)Sociology [13]: the Cox regression can be applied to determine the marital instability. The age difference between husband and wife, how old the couple is when they got married and the education level of the husband, are all considered significant in making prediction of marital instability.

### 3.2.1 Evaluation Metrics

The standard evaluation metrics for classifiers are not used for evaluating survival models. Since the data on which Cox models are applied typically include a notion of censored time, so we also need different measures to actually evaluate these models. There are several widely accepted indicators that measures a specific method's performance, such as c-index [49], concordance probability [23] and AUC (Area under ROC curve).

The concordance probability is used to evaluate the discriminatory power and the predictive accuracy of non-linear statistical models. In Cox PH model, the proposed estimator is a function of the regression parameters, and the covariate distribution does not use the observed event and censoring time. This makes the concordance probability asymptotically unbiased in comparison with the c-index when dealing with the Cox regression. As mentioned before, AUC is widely accepted as the standard measure for describing and comparing the prediction accuracy for many different models. Unlike the concordance probability which is designed for Cox regression, AUC can be used as the measure for different models and suitable for many situations.

The concordance probability is designed to evaluate the predictive accuracy of non-linear models. Suppose there are two patients in the Cox regression, $(X_1, t_1)$ and $(X_2, T_2)$, the concordance probability is defined as follows

$$K_{X,T} = pr(T_2 {>} T_1 | X_2 \geq X_1) \tag{3.3}$$

If $X$ is binary and $T$ is ordinal, then the concordance probability can be treated as AUC. A concordance probability of 1 shows that a model has a perfect performance on certain data; On the contrary, a value around 0.5 indicates that the model performs poorly on the data.

In the meanwhile, c-index is calculated as the sum of concordance value divided by all possible pairs

$$c - index = \frac{(concordant + \frac{1}{2}ties)}{allpairs} \tag{3.4}$$

where *concordant* stands for two individuals being compared.

In this study, we use AUC as the way to describe the prediction model, since the AUC is a widely accepted measure for different models. The results of the Cox regression in our study is used to compare with other models, and the AUC is the best metrics to measure the performance of the model. The AUC of the Cox PH model is different from the traditional models. In traditional cases, the AUC is the accuracy of a prediction result. However, in Cox regression, AUC shows the prediction accuracy during a particular period of time. Different AUC values forms a curve and each point on the curve represent the AUC at that time point. The AUC at each point is similar to traditional cases, showing the prediction accuracy of that moment. Hence, when dealing with the Cox regression, a good result requires both an overall good AUC curve and a high AUC at some specific points.

## 3.3 Regularized Cox Models

In this section, we look at the usage of regularization methods in the Cox model. The regularization methods can yield models that are more efficient and robust. Due to the recent advancements in the medical field, high-dimensional data appears in many applications. The traditional way of maximizing partial likelihood estimator does not work well in the high-dimensional data where the number of predictors are large. In such cases, we need a robust method for feature selection which can effectively identify important features. In order to solve this problem, we introduce several shrinkage methods to combine with the Cox PH model in order to get better optimization results [47]. These shrinkage methods are used to solve the optimization problem in high-dimensional spaces instead of using the traditional method to maximize the partial likelihood.

### 3.3.1 Regularization Methods

- LASSO: LASSO is an application of the shrinkage method, and is standard for least absolute shrinkage and selection operator. LASSO achieves a better prediction outcome by applying shrinkage method, and meanwhile, it gives a sparse solution, which means that some coefficients for the predictor are set to zero. Hence, LASSO can also be used to perform feature selection. The main idea of LASSO is to use the $L_1$ norm constraint to obtain the estimator by minimizing the empirical risk in partial log-likelihood. A positive constant, $S$, is used as the threshold for the $L_1$ norm. To generalize LASSO, some special optimization techniques are required. Tibshirani [20] used a method called *quadratic programming* (QP) method for regression.

Though LASSO provides a good method to deal with the high-dimensional constraint optimization, there are several limitations of LASSO.

1. In the $p > n$ case, the LASSO can only select $n$ features during the process of feature selection.
2. If there is a group of variables with a high correlation, the LASSO tends to select only one of them and leaves the rest.
3. In standard cases, if the variables have a high correlation, the prediction performance of the LASSO will be very close to ridge regression.

- Ridge: Ridge regression uses the $L_2$ norm penalty. Ridge regression cannot zero out any coefficients. That means it cannot achieve feature selection. As a result, after applying ridge regression to the dataset, one could only keep all the variables or discard all of them.
- Elastic Net: This uses a convex combination of the $L_1$ and $L_2$ norms in the penalty.

### 3.3.2 Cox-LASSO

Cox-LASSO is a popular survival model that determines the coefficient of different variables [48]. Instead of using the least squares method to maximize the partial log-likelihood in Cox PH model, the LASSO method subjects the partial log-likelihood to a certain constant using shrinkage method. The shrinkage method is an estimation and prediction method in regression problems. In general terms, the shrinkage method improves the estimate by combining the original estimate with other information. This means that the value of the improved estimate will be approximately closer to the 'other information' offered by the input dataset.

In order to get the estimator $\hat{\beta}$ that maximizes the partial log-likelihood, suppose $X$ denotes the input variables and $\eta = X\beta$, then we can define $u = \frac{\partial l}{\partial \eta}$, $A = \frac{\partial^2 l}{\partial \beta \beta^2}$ and

$z = \eta + A^{-1}u$. Then a one-term Taylor expansion for the partial log-likelihood can be expressed as

$$l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.5}$$

Hence the procedure of using the shrinkage method to maximum the partial log-likelihood can be summarized as shown in Algorithm 1.

The last value of $\hat{\beta}$ is the coefficient derived from the Cox-LASSO method. A proper value for the boundary $s$ can often make some of the solution coefficients to zero, so that the best subset of the variables are selected. This makes the final model more interpretable and stable. There are two advantages of this method compared to the traditional methods.

**Algorithm 1** Cox-LASSO Method

```
1. Initialize $\hat{\beta}_0 = 0$ and find a proper value for the boundary constant $s$.
2. Find out the value of $\eta$, $u$, $A$ and $z$ based on the current $\hat{\beta}$.
3. Minimize the Equation 3.5 according to the constraint and record the value of new $\hat{\beta}$.
4. Repeat steps 2 and 3 until the value of $\hat{\beta}$ stops changing.
```

1. **Better Prediction Accuracy**: the traditional estimates often have low bias but large variance, in this case, shrinking or setting some coefficients to 0 can improve the prediction accuracy.

2. **Ease of Interpretation**: the shrinkage method can determine a smaller subset from the large number of the variables. By doing this the variables are displayed in a better manner.

### 3.3.3 Cox-Elastic Net

Elastic net is a regularized regression method that combines the LASSO's $L_1$ and Ridge's $L_2$ penalties [53]. This method is suitable for the linear regression models, especially when there are several highly correlated variables. Similar to the LASSO, the elastic net can be set up like this, assuming that the predictors are standardized, the estimates from the elastic net method can be obtained as $\alpha \parallel \beta \parallel_1 +(1-\alpha) \parallel \beta \parallel_2$, where we can see that the elastic net method includes both LASSO and Ridge regression. When $\alpha$ is set to 1, the elastic net will be same as LASSO. On the other hand, if $\alpha$ is 0, the elastic net behaves as ridge regression. Hence, the elastic net has the following advantages:

1. It results in a sparse model of selected variables.
2. It can eliminate the bad effect of high correlations between several variables in a group, which encourages grouping effects.
3. The $l_1$ path will be robust by applying the elastic net.

Figure 3.1 shows a 2-dimensional illustration of ridge, LASSO, and elastic net. In Figure 3.1, $\alpha = 0.5$. In this Figure, it is obvious that the path in the middle is the elastic net regularization path. Experimental results on different kinds of datasets also prove the effectiveness of the elastic net penalty over the LASSO and Ridge [27]. The Cox Proportional Hazard model can also use the elastic net penalty. We provide a formulation of this approach, which is called the Cocktail algorithm [51].

### Cocktail Algorithm

The shrinkage method is an efficient method to deal with multiple linear regression, while the Cox PH model is a linear regression model. Instead of maximizing the partial log-likelihood, a faster and more reliable algorithm called Cocktail is introduced to compute the coefficients of the Cox PH model. This algorithm is the application of the elastic net penalized Cox PH model. Compared with other algorithms, the Cocktail Algorithm is faster and gives better solutions in terms of its robust and high efficient.

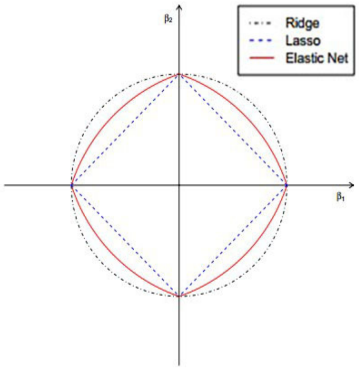


Figure 3.1: Comparison of LASSO, Ridge and Elastic Net.

The Cocktail Algorithm combines the strengths of coordinate descent and the majorization-minimization principle [19]. This combination can be generated into a new coordinate-majorization-descent (CMD) algorithm, which is used to minimize the following function

$$Cocktail formulation = l(\beta) + \sum_{j=1}^{p} \lambda[\alpha w_j|\beta_j| + \frac{1}{2}(1-\alpha)\beta_j^2] \qquad (3.6)$$

where $l(\beta)$ is the likelihood (In Equation 3.5) and $w_j$s are the non-negative weights making estimate more flexible. Cocktail uses the weighted $L_1$ norm in the elastic net penalty. The parameters $\lambda$, $\alpha$ are provided by the user. It has been proved that the Coordinate descent is a promising method of solving LASSO penalized models. Coordinate descent can be used to obtain a solution for the formulation in Equation 3.6.

The Cocktail Algorithm is shown in Algorithm 2

```
Algorithm 2 cocktail algorithm
Require: Features $Feat$, Censored Status $Stat$, Survival Times $Time$, $\lambda$ regularization
    parameter, $\alpha$ elastic net parameter
    1. Use $Feat$, $Stat$, $Time$ to build the cox likelihood using Equation 3.5
    2. Formulate the convex optimization problem as given in Equation 3.6
    3. Use unconstrained iterative optimization methods such as coordinate descent for
    solving Equation 3.6
    4. Output the coefficient vector $\hat{\beta}$ to obtain the AUC and important features
```

### 3.3.4 FASTCOX

A package called '*fastcox*' is used to implement Cocktail algorithm for Cox elastic net method in R [24]. The package is available on the *CRAN* fastcox site.

Here we have an example on utilizing fastcox. *Dat* is a sample dataset to demonstrate the commands of this package.

```
0. Load fastcox library and Dat.
     library(fastcox)
     data(dat)
1. fit the solution paths
     m1 $\leftarrow cocktail(x = Dat\$x, y = Dat\$y, d = Dat\$status, alpha = 0.5)$
2. make a plot for solution paths.
     plot(m1)
3. compute the coeffcient of each variable.
     predict(m1,type="coefficients")
```

# CHAPTER 4

# EXPERIMENTAL RESULTS

In this chapter, we demonstrate the performance of the proposed models using several performance metrics such as classification accuracy, AUC value and efficiency. Initially, in section 4.1, we describe the dataset used along with the pre-processing methods. In section 4.2, we compare the result of different kernels and the performance of SVM with other models applied on our data. Then, in section 4.3, we show the results of uncertain patient detection, and finally, in section 4.4, we show the results of Cox Proportional Hazard model and model comparison.

## 4.1 Experimental Setup

For our experiment, we obtained the EMR data from the Henry Ford Health System in Detroit, MI. This dataset contains the information of 8692 heart failure patients from the year 2001 – 2010. For each patient, more than 100 different features are selected for further analysis. The data is used for the prediction of 30-day readmission for heart failure. Table 4.1 summarizes the six different data sources used in our experiments with the 8692 patients.

Table 4.1: Summary Statistics of the EMR Data Used in our Experiments

| | No. of attributes | Min Value | Max Value | Average Value |
|---|---|---|---|---|
| Demographics | 86 | 0 | 1096 | 276.72 |
| Admissions | 2 | 0 | 72 | 3.01 |
| Medications | 6 | 0 | 246 | 3.75 |
| Procedures | 10 | 0 | 94 | 1.92 |
| Labs | 48 | 0 | 8261 | 302.65 |
| Furosemide | 4 | 0 | 113 | 4.54 |

### 4.1.1 Data Description

The data is collected from six different data sources. In general, the dataset contains the patients' information: *Demographics, Admission, Medications, Procedures, Labs, and Furosemide.*

1. **Demographics:** this shows the demographic information of a patient, such as gender, age, and race.

2. **Admission:** this data shows the number of times and exact date of readmission for each patient after the discharge date.

3. **Medications:** it shows the drugs each patient received during their hospitalization.

4. **Procedures:** this data shows the procedures each patient had received during hospitalization.

5. **Labs:** this data gives the basic health information about different labs for each patient and their results.

6. **Furosemide:** this data covers the drugs each patient received during their hospitalization.

Among the 8692 patients, 5890 of them were readmitted to the hospital. Those patients are the ones that will be considered for the further analysis. The rest of the patients are not recorded because they were discharged from the hospital after their first hospitalization. The features of each patient come in three different forms of data types: numerical data, categorical data or mixed value data. The composition of each data source is different, and it directly influences the performance of the results. The demographics is composed of mixed value data; procedure, medication and furosemide datasets are composed of categorical data, while the lab file is composed of numerical data. As a result, different data mining methods are applied to different data sources.

### 4.1.2 Performance Metrics

Area under ROC curve is a method to evaluate the performance of a model [36, 6]. Receiver Operating Characteristic curve, famous for diagnostic studies in Clinical data processing, has been accepted widely as the standard metric for describing and comparing the prediction accuracy of many different models. If the curve rises from the left-corner rapidly towards the right-corner, or in other words, the area under the curve (AUC) is large, we can concluded that the model performs well on this dataset. On the contrary, if the curve goes directly towards the right-corner, it will be obvious that the model performs poorly on this dataset. Theoretically, the ideal model will have an area close to 1, indicating a model's perfect performance. On the other hand, if the area is close to 0.5, it shows that the model performs badly. When the curve goes below the diagonal line, the model performs worse than the random classifier.

ROC can be measured by some parameters. For every point in the dataset, there are four different possibilities.In some cases, when the data points are classified, the patients with diseases are correctly classified (TP), while other patients are classified into the wrong class, the class without any diseases (FN). On the other hand, some healthy cases are classified correctly to the group without the disease (TN), while yet again in some cases the healthy people are classified into the group with diseases (FP). Those four notations denote the four possibilities of a prediction model. An ideal model will have $TP + TN$ close to 1. That is, the AUC will approach 1. Table 4.2 demonstrates the relationship among the four categories.

The confusion matrix acts as the basic element to form the ROC curve. The two

Table 4.2: Confusion Matrix

| | Predicted | |
|---|---|---|
| Actual | TP | FP |
| | TN | FN |

axis of the ROC curve depend on this matrix.

The ROC curve is determined by sensitivity and specificity. Sensitivity is an index indicating the positive identifying ability. In the field of clinical information, it is used to show the proportion of a group of people who are diagnosed with disease and classified positive. It can be expressed as

$$sensitivity = \frac{(TP)}{(TP) + (FN)} \tag{4.1}$$

It is obvious when a model has high sensitivity, it has a better ability to pick up the right people and place them in the correct class. It is important, however, to notice that when calculating the sensitivity, the indeterminable results are not taken into account. All the indeterminable data points have to be excluded or treated as false or negative.

Specificity refers to the model's ability to identify the negative results in a study. In the field of clinical information, it is used to present the proportion of people who are healthy and also classified them as such. It can be expressed as

$$specificity = \frac{(TN)}{(TN) + (FP)} \tag{4.2}$$

If a model has a result with high specificity, then this model can perform well on this dataset.

The ROC curve is widely accepted as a tool to judge how good the prediction results are. In many studies, varied variables are applied to raise the AUC value. The most common variable is the socio-demographic factors. Almost all the studies include this as a variable. However, the overall health and function is considered the least significant so that only a few studies take it into consideration. In this thesis, AUC will be used as the method to measure and compare the performance of different models.

### 4.1.3 Data Preprocessing

Since the original datasets contain many missing values of patients that failed to return to the hospital in a certain period of time, some data were not collected. As a result, we used *Weka* to replace all the missing value and get statical results of traditional machine learning algorithms. *Weka* is an open source data mining tool [25], that collects most of the popular algorithms for data mining tasks. We use Weka's inbuilt function 'ReplaceMissingValue' to replace all the missing value in the datasets. If a column has more than 60% missing value, then that column was removed.

The next step is to generate the labels for each patient. In most studies, the label is built according to the 30-days readmission. The fact that 30-day readmission rate is the most significant performance quality measure method is publicly accepted. For the SVM model, we create the label class depending on the 30-day readmission rates as well as with the survival status. The patients are divided into two classes $\{1, -1\}$, which means that if a patient is readmitted within 30 days from discharge for heart failure, whether dead or alive, this patient will be put into class $\{1\}$; or if the patient have not been readmitted within 30 days but is still alive, he will be put into class $\{-1\}$; otherwise if one have not been readmitted within 30 days but is dead, this patient will be removed from the dataset. When the missing values are filled, the datasets are integrated into a single large dataset to use the useful attributes for classification.

For the Cox PH model, we label it using a different method. The baseline function of Cox PH model is determined by the censored time and censored status. The censored time is the length of time from the start of the censoring or study, till the end of it. If a patient is censored, in another words, dead or never returned to the hospital during the censoring period, that patient is put into class $\{1\}$, otherwise he will be put into class $\{0\}$. It is important to note that the censored time and censored status are all the indexes during the study, and should be distinguished from the real-world time.

Before further analysis conducted on these labelled datasets, Weka's machine learning algorithm such as *decision tree* (DT), *Adaboost* and *logistic regression* are applied to analyse the datasets. Table 4.3 demonstrates the results of Weka's algorithm. These results will show the original performance of each dataset.

*Demographics* and *Lab* data sources gave the best results among all the datasets, and these two data sources are especially used in the experiments to follow later on. *Medications*, *Procedures* and *Furosemide*, however contain few attributes. These three data sources should be integrated with other data sources during further analysis.

Table 4.3: AUC Values for Different Sources

| Datasets | | ROC | | |
|---|---|---|---|---|
| Source | # of attributes | DT | Adaboost | Logistic |
| Demographics | 86 | 0.57 | 0.58 | 0.64 |
| Medications | 6 | 0.51 | 0.51 | 0.51 |
| Procedures | 10 | 0.5 | 0.53 | 0.43 |
| Labs | 48 | 0.51 | 0.51 | 0.57 |
| Furosemide | 4 | 0.50 | 0.54 | 0.53 |

## 4.2 Detecting Uncertain Patient

In this section, we show the SVM result and the application of SVM in the uncertain patient detection using the proposed UPAD algorithm. SVM is a powerful model for solving non-linear classification problems, and the recent development of kernels makes the performance of SVM even better. LS-SVM uses the least square method for implementing SVM. The results of LS-SVM compared with basic SVM are more accurate. This model is robust and efficient in our clinical datasets.

| Kernel Name | kernel parameter | | | |
|---|---|---|---|---|
| RBF kernel | $\sigma$ =1 | $\sigma$ =10 | $\sigma$ =50 | $\sigma$ =100 |
| Linear kernel | 0 | 0 | 0 | 0 |
| Polynomial kernel | d=1 | d=10 | d=50 | d=100 |

Table 4.4: Kernel Parameters Used in Our Work

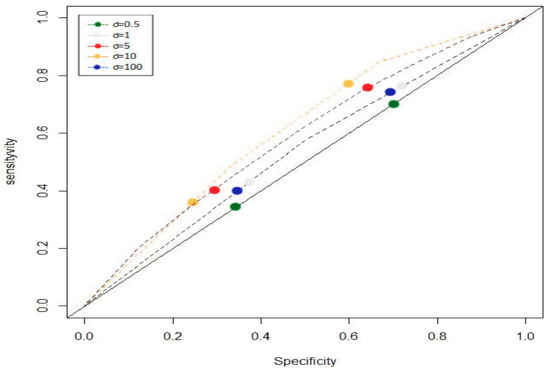


Figure 4.1: AUC of RBF kernel.

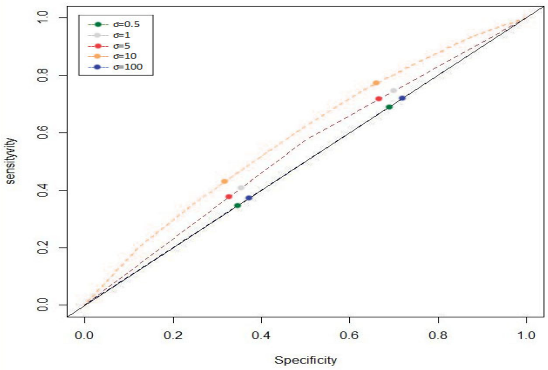


Figure 4.2: AUC of Polynomial kernel.

### 4.2.1 Experimental Setup

Five data sources are used in LS-SVM to make prediction. A patient will be removed from the dataset if 60% of the attributes have missing values. Hence, there are 7788 patients left in each dataset. We use WEKA to fill in the missing values. 10-fold cross validation is applied to obtain the AUC (Area under ROC curve). The training data contains 7008 patients and the rest of the 779 patients are contained in the testing data. For the training data of each data source, we implement *linear kernel, polynomial kernel* and *RBF kernel* to obtain the kernel matrix. The parameter of each kernel can be adjusted to achieve the best performance of kernel. Table 4.4 demonstrate the parameter we tested on each kernel.

In order to get the best parameter for each kernel, we integrate the five data sources and apply different kernels on it. Figure 4.1 and Figure 4.2 demonstrate the AUC of each kernel's performance on the integrated dataset.

From the experiment we discover that when $\sigma$= 10, the RBF kernel gives a result with the highest AUC value, while the result of linear kernel will not change. Table 4.5 shows the results of the performance of kernels on each dataset.

The datasets for this study are the non-linear datasets, hence the linear kernel can not produce a good result using linear method, while the RBF kernel gives a good result. Due to this, when we select the RBF kernel and $\sigma$= 10, the highest accuracy results are obtained.

From Table 4.5, we observe that LS-SVM results have been improved compared to the Weka results. For LS-SVM results, the RBF kernel produces the best results and the *Lab* data source as well as *Demographics* data source, have the highest AUC value. The attributes in these two data sources are considered significant in diagnosing a patient. *Lab* data source have several significant indices for a patient which includes variables such as BUN and CRET. *Lab* data source generate the variables associated with the overall health condition, and is considered important since the attributes of lab data directly shows the health condition of a patient, while *Demographics* data source shows the basic information (age, sex and race) of a patient. Figure 4.3 and Figure 4.4 give the ROC curve of the result of RBF kernel on Lab and HF sources.

Table 4.5: LS-SVM Results

| | parameter=10 | | |
|---|---|---|---|
| Source | Linear Kernel | Polynomial Kernel | RBF Kernel |
| Demographics | 0.52 | 0.53 | 0.56 |
| Medication | 0.51 | 0.51 | 0.53 |
| Procedure | 0.52 | 0.51 | 0.52 |
| Lab | 0.53 | 0.53 | 0.57 |
| Furosemide | 0.51 | 0.53 | 0.53 |

Table 4.6 and Table 4.7 gives the confusion matrix of RBF kernel applying on both Demographics and Lab sources.

Kernel methods perform better than standard machine learning algorithms, such

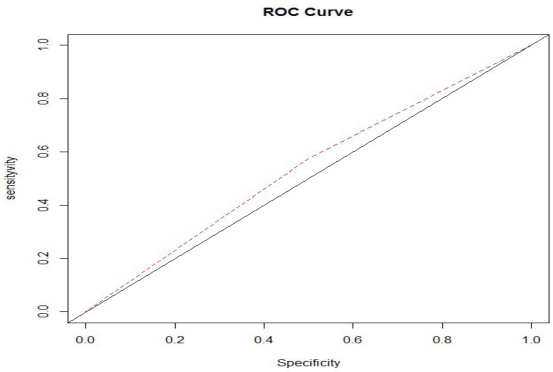


Figure 4.3: ROC for Demographics with RBF kernel.

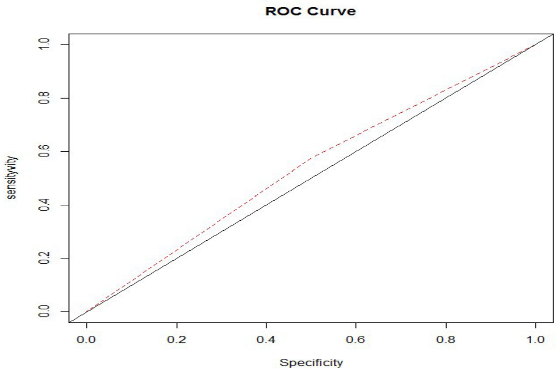


Figure 4.4: ROC for lab with RBF kernel.

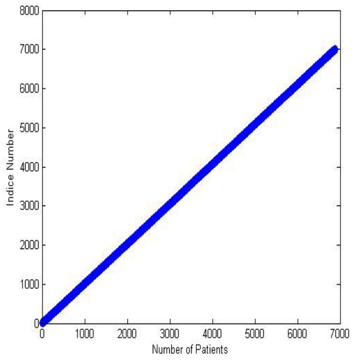
Table 4.6: CM for RBF(HF)

|  |  | Actual |  |
|---|---|---|---|
|  |  | True | False |
| Prediction | Positive | 0.804 | 0.009 |
|  | Negative | 0.183 | 0.004 |

Table 4.7: CM for RBF(Lab)

|  |  | Actual |  |
|---|---|---|---|
|  |  | True | False |
| Prediction | Positive | 0.811 | 0.003 |
|  | Negative | 0.187 | 0 |

as *decision tree, adaboost* and *logistic regression.* SVM classifier performs well on non-linearly separable datasets. LS-SVM is better suited for this problem. *RBF kernel* follows the Gaussian distribution, while the other two kernels are linear kernels. Hence, *RBF kernel* obtains the best results. Kernel methods are designed for SVM, when using kernels in SVM there is a higher probability to get a better result using kernel matrix instead of using the original datasets.



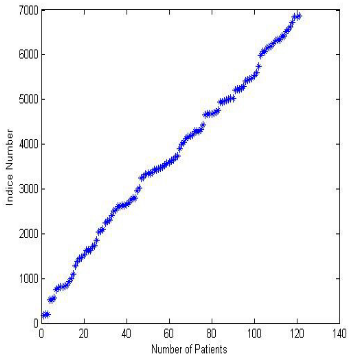


Figure 4.5: Certain Patients Distribution.

Figure 4.6: Uncertain Patients Distribution.

### 4.2.2 Uncertain Patient Detection

Noise is defined as the irrelevant and erroneous data which affects the performance of the classifier negatively, and is unavoidable in many real-world applications. Noise patients are considered to be those that are hard to predict into the right class. These patients are known as uncertain patients. To improve the performance of the model, the noise should be removed from the dataset. We provide the results obtained from our UPAD (Uncertain Patient Detection) algorithm.

Among the training data, 121 patients are identified as uncertain patients (UP) and 7008 are certain patients (CP) in our experiments. Figure 4.5 and Figure 4.6 show the distribution of CP and UP.

Using our UPAD approach we obtain some plots for certain lab attributes of the patients. We compare the lab attributes distributions for the set of uncertain patients (noisy patients) and certain patients.

Figure 4.7-4.12 show the distribution of three attributes of certain patient and uncertain patient. These three attributes are *MG, CRET* and *BUN*. These attributes are highly related to the patient's health condition. *MG* stands for Magnesium, which affects one's muscle behaviour. *CRET* means Creatinine, and is a very important indicator of

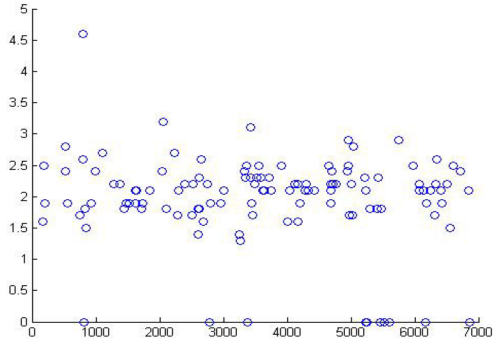


Figure 4.7: MG for uncertain patient.

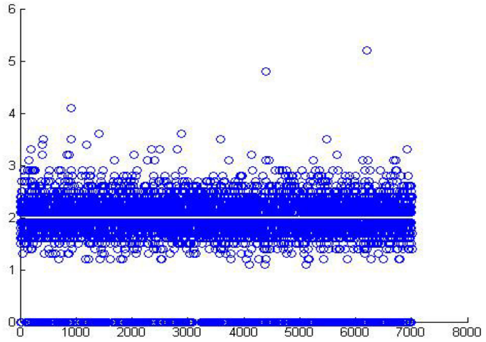


Figure 4.8: MG for certain patient.

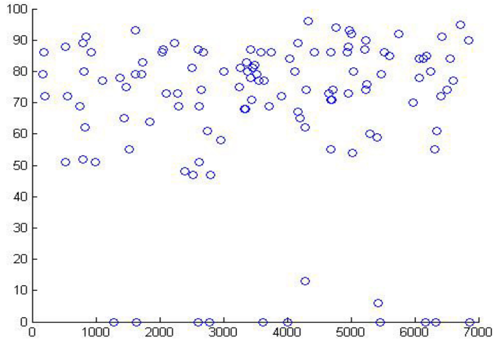


Figure 4.9: CRET for uncertain patient.

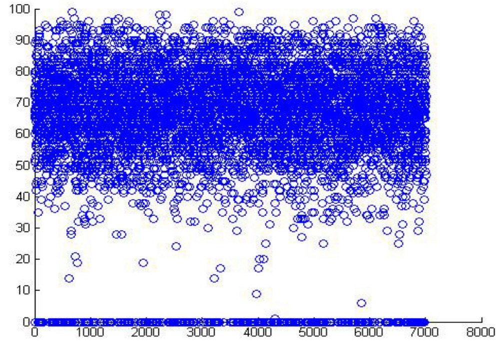


Figure 4.10: CRET for certain patient.

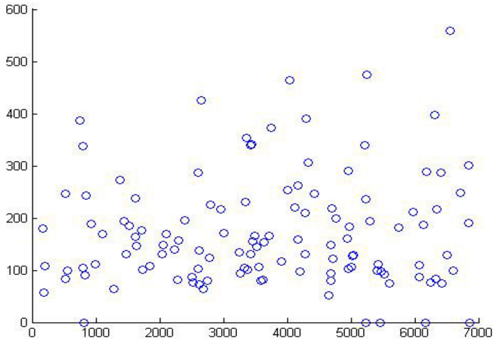


Figure 4.11: BUN for uncertain patient.

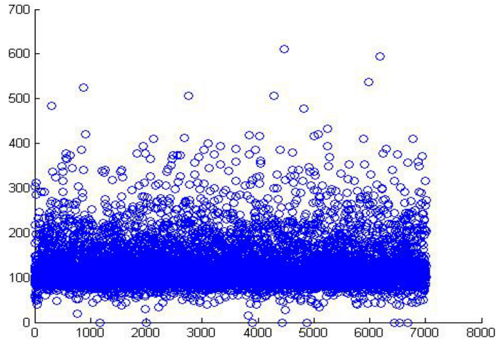


Figure 4.12: BUN for certain patient.

a person's health condition. *BUN* represents Blood Urea Nitrogen. Our test measures the amount of nitrogen in patient's blood that comes from the waste product urea and it directly indicates whether one's kidneys are working normally.

From these plots, we observe that the distributions of the value in a certain feature has a big distinguish in the two group of patients. The UP (uncertain patient) usually has a distribution focusing on a certain area while the distribution of CP (certain patient) is more emanative; and in most cases, comparing to CP, the UP does not have the 'extreme' values, which means the minimum or maximum value of an attribute. In the meanwhile, for all the attributes, the values of the attributes of CP have a certain distribution. However, the values of the attributes of UP may distribute orderless. For instance, the attribute value distribution of an UP who readmits after discharge may be more likely to a patient who never readmits.

## 4.3 Results on Survival Analysis

In this section, we introduce the survival analysis results obtained from using Cox regression, Cox Lasso and Cocktail algorithm.

Survival analysis is a reliable statistics method which deals with physiological characteristics and censored failure time in medical system. Cox PH model is suitable to interpret censored data., and it performs better on a regularized dataset. As a consequence, our datasets will be regularized before applying the model.

In machine learning, regularization is a method to prevent overfitting via introducing additional penalty terms to the model. The additional information is a constraint of the model, such as the boundary of a parameter. Regularization is also applied to do feature selection. It penalizes models based on their parameters. In order to have a better survival analysis result, we regularize the dataset before applying it to the Cox model. The algorithm to regularize the model is called Cocktail. It is a Cox-elastic algorithm which uses shrinkage method to optimize the coefficient instead of maximizing the partial log-likelihood. This algorithm gives us the authority to control the parameter $\alpha$ while keeping the results robust at the same time. When $\alpha = 0$, the model turns to Cox-LASSO model and when $\alpha = 1$, the model becomes Cox-Ridge model.

### 4.3.1 Experimental Setup

We use a package called '*survival*' in 'R' language [30]. After loading the clinical data, the survival library is loaded using the command 'coxph' and survival analysis is performed on the clinical data.

The result gives the coefficient of each variable in the dataset and the maximum likelihood. We apply the Cox PH model to the integrated dataset to make the survival analysis. Table 4.8 shows the results of the Cox PH model. The input dataset is the original dataset and the processed ones which is regularized by different regularization methods such as Cox-LASSO (CL), Cocktail (Algorithm for Cox Elastic Net) and Cox-Ridge (CR).

Table 4.8: AUC of Cox PH model

| Data Source | Cox | CL | COCKTAIL($\alpha = 0.2$) | CR |
|---|---|---|---|---|
| Integrated dataset | 0.58 | 0.65 | 0.65 | 0.56 |

### 4.3.2 Experimental Results

Survival curve is a graph showing the proportion of a population living at a given time after contracting a serious disease or receiving a radiation dose. The survival curve displays the predicting ability of each attribute. Figure 4.13 and Figure 4.14 show the survival curves of Cox PH model and the Cox PH model with regularization.

From Table 4.8 we can observe that Cox-LASSO and Cocktail give the best result. For Cocktail, many experiments have been conducted to test the best value of $\alpha$. The number of attributes is decreasing with the value of $\alpha$ increasing, for that the more coefficients are set to zero as the $\alpha$ is increasing. When $\alpha = 0.2$, the proper number of attributes

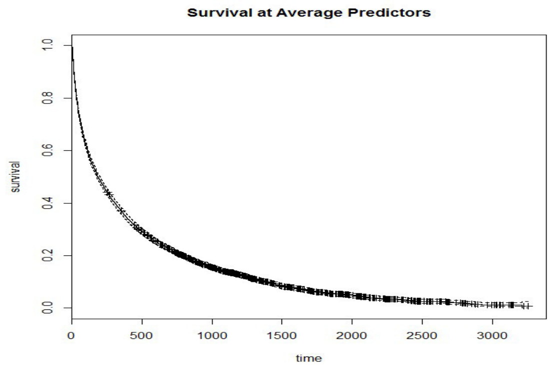


Figure 4.13: Survival Curve for Cox PH Model.

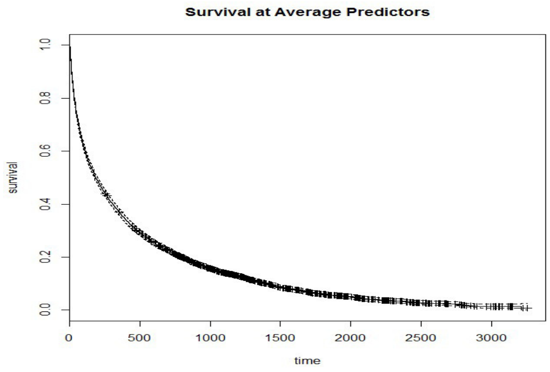


Figure 4.14: Survival Curve for COCKTAIL ($\alpha = 0.2$).

are maintained, and for each attribute, a proper coefficient is given to make the best result. Figure 4.15 and Figure 4.16 demonstrate the ROC curve of Cox regression and COCKTAIL.

The survival curves for both methods are plotted by obtaining the coefficient values from the shrinkage methods, and then the Cox survival function in R package is applied. The X-axis represents the time in days and the Y-axis represents the survival rate. Cox model prediction is based on MLE (Maximum Log-likelihood Estimation), and this model does not have sparsity. By using the sparse methods, we obtain the correct set of predictors, which result in building a more superior model.

The Cocktail algorithm also has the function to do feature selection. Table 4.9 and Table 4.10 report 13 attributes with the highest weight, respectively. From Table 4.9 and Table 4.10, we observe that *ckd, heart failure, Hx_ESRD, FU status* and *cva_tia* appear in both tables. Those attributes are considered more important in survival analysis.

Table 4.9 shows the top ranked features obtained from Cocktail algorithm. It gives the attributes from four data sources. This data consists of a mixed binary and real value attributes. These attributes carry the information of the medical usage, present health condition and socio-demographic information of a patient. The results show that there is a strong correlation between the information of a patient and the censored status.

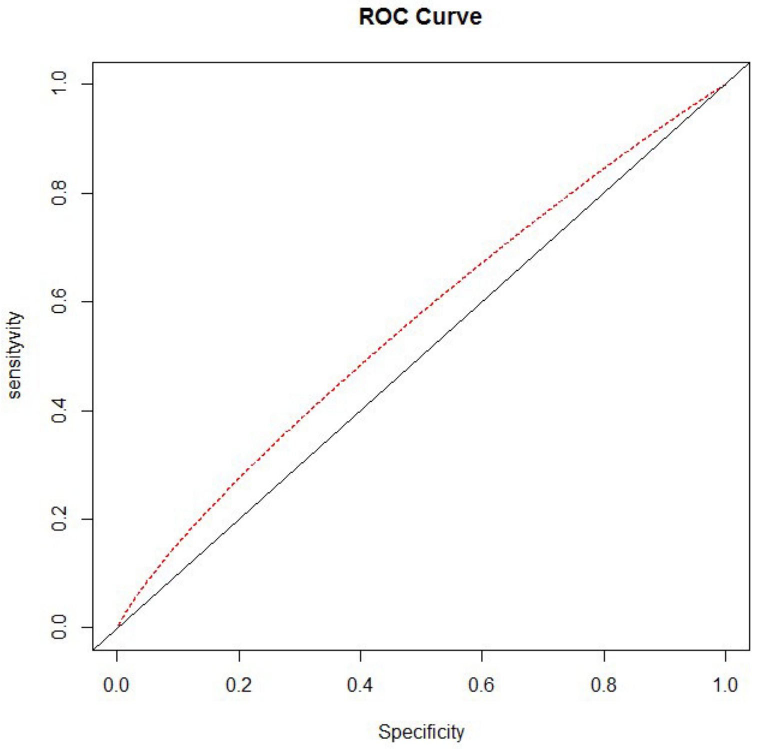


Figure 4.15: ROC of the Standard Cox model.

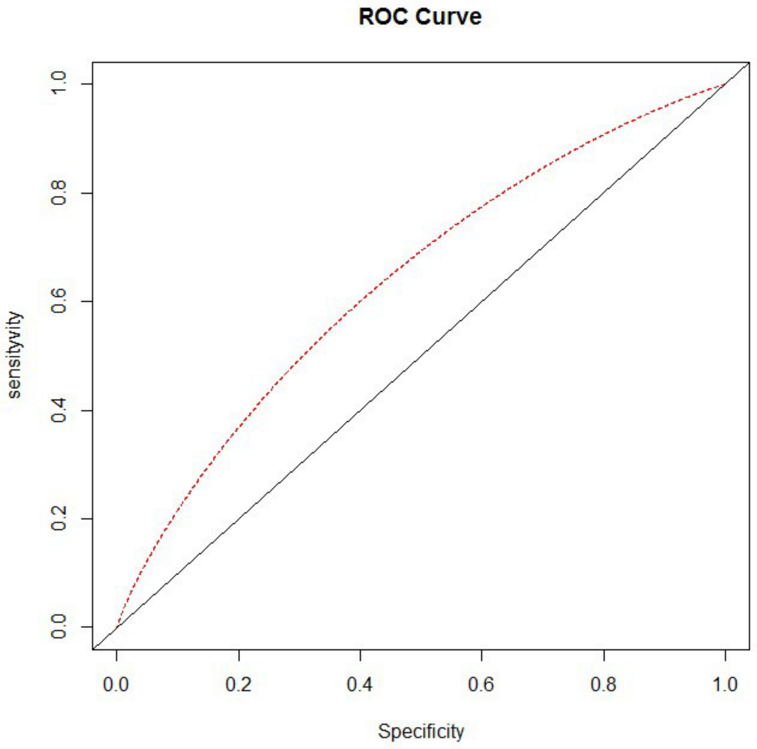


Figure 4.16: ROC of the COCKTAIL Algorithm with $\alpha = 0.2$.

Table 4.9: Top Ranked Attributes Obtained Using the COCKTAIL Algorithm

| Data Source | Attribute | z-value | p-value | Description |
|---|---|---|---|---|
| Demographics | cva_tia | 3.77 | <0.05 | transient ischaemic attack |
| Demographics | heart_failure | 8.30 | <0.05 | heart failure diagnosis |
| Demographics | chd | 2.08 | <0.05 | coronary artery disease |
| Lab | MG | 1.43 | 0.15 | Magnesium |
| Demographics | cardiac_cath | 0.39 | 0.70 | cardiac catheterization |
| Demographics | ckd | 3.47 | <0.05 | chronic kidney disease |
| Procedure | constrast.dye | 0.11 | 0.91 | HF Procedure |
| Medication | nesiritide | 1.25 | 0.21 | HF Medication |
| Lab | CRET | 3.11 | <0.05 | Creatinine |
| Demographics | WRF_GE_25pct_dischg | 1.00 | 0.32 | increased creatinine measurement |
| Demographics | FU_status | 4.84 | <0.05 | patient follow up |
| Demographics | race | -3.47 | <0.05 | Race of patient |
| Demographics | RHC | -0.03 | 0.98 | whether patient received RHC |

Table 4.10 presents the top ranked features of Cox-LASSO algorithm. These selected attributes primarily come from lab and demographics data sources. Those attributes mainly carry the information of a patient's health condition. The average p-value of these attributes is around 0.5, meaning these attributes contribute significantly to the prediction result.

ADHERE and TABAK are two popular models used in the domain of clinical information. They select the predictors which carry the information of a patient's health condition and social background from different hospitals in a large set of patients. The two models are used to give the result of 30-day readmission prediction. We observe that the attributes chosen by these Cox methods fit well with the famous set of predictors used in the heart failure readmission. All comorbidities variables are detected by these models, which are considered to be very important predictors for readmission in heart failure. Labs such as Creatinine and BUN are also established indicators which are ranked by these models.

Table 4.10: Top Ranked Attributes Obtained Using the CL Algorithm

| Data Source | Attribute | z-value | p-value | Description |
|---|---|---|---|---|
| Demographics | ckd | 3.43 | <0.05 | chronic kidney disease |
| Demographics | heart_failure | 7.71 | <0.05 | heart failure |
| Demographics | Hx_ESRD | 1.60 | 0.11 | history of ESRD |
| Demographics | pvd | 3.10 | <0.05 | peripheral vascular disease |
| Demographics | cva_tia | 3.73 | <0.05 | transient ischaemic attack |
| Demographics | FU_status | 4.93 | <0.05 | last known follow up |
| Demographics | htn | 6.29 | <0.05 | hypertension |
| Demographics | hemodialysis | -0.155 | 0.88 | hemodialysis |
| Demographics | diabetes | 2.37 | <0.05 | diabetes |
| Demographics | afib | 1.13 | 0.26 | atrial fibrillation |
| Demographics | WRF_GE_pt5 | 0.04 | 0.96 | increase in creatinine measurement |
| Demographics | min_RF | -1.37 | 0.17 | NKF Renal classification |
| Lab | BUN | 0.41 | 0.68 | Blood Urea Nitrogen |

As mentioned in the related work, several studies build models which perform well on clinical data. We have similar data with two of the models, *ADHERE* model and *TABAK* model. Those two models have the information of a patient's medical condition and the procedures taken during hospitalization. As a consequence, the datasets *procedure* and *furosemide* are integrated to make up the attributes for two models. Table 4.11 demonstrates the performance of each model using our datasets. For CL and Cocktail, we demonstrate the AUC at the best point.

Lab data gives the health condition and drug information of a patient but lacks

Table 4.11: AUC Comparison of Different Individual Sources

| Source | ADHERE | TABAK | CL | Logistic | COCKTAIL |
|---|---|---|---|---|---|
| Demographics | 0.579 | 0.552 | 0.65 | 0.663 | 0.65 |
| Lab | N/A | 0.632 | 0.65 | 0.611 | 0.64 |
| Medication | 0.543 | 0.516 | 0.61 | 0.516 | 0.58 |
| Pro + Fur | 0.551 | 0.542 | 0.58 | 0.519 | 0.57 |

several significant attributes for the ADHERE model. As a consequence, lab data does not have the result for the ADHERE model, and we mark N/A to show it. From Table 4.11, it is obvious that Cox-LASSO has a higher AUC over other models in general. This model uses the shrinkage method to generalize the coefficient of each attribute, which is more suitable for clinical data. The shrinkage method provides a promising way to deal with the different types of data. As a consequence, after several iterations, a better optimization result will be obtained comparing to logistic regression method in clinical data. For the Demographics data, the logistic regression gives a better result. Demographics data is a mixed dataset of both numeric data and binary data. In this case, logistic regression has a better ability to deal with mixed data, thus the logistic regression gives a better result.

# CHAPTER 5

# CONCLUSION AND FUTURE WORK

## 5.1 Conclusions

In this section, we highlight the main inferences we obtained from conducting different kinds of experiments in this thesis. We analyzed clinical data and integrated it across different sources. We modelled this as a regression problem and conducted survival analysis on this dataset. We understand that survival models are extremely useful in clinical data studies and they tend to select important features.

We applied the SVM and Cox PH models to the clinical data. We used machine learning package to fill the missing values in the data sources. Then, we applied logistic regression models, SVM models and Cox Proportional Hazard models to test the performance of different methods on our data.

Sparse methods were used to optimize the performance of Cox PH model. These methods obtain a trade-off between the goodness measure and sparsity of the result. These methods not only focus on the accuracy or any other performance measure, but also aim to obtain more interpretable results. Sparse methods are effective for clinical data, it is useful in the understanding of large data sources. LASSO is applied to optimize the coefficients for Cox PH model, which makes the result more accurate and interpretable.

Cox regression model is applied for the survival analysis. It is a statistical technique for exploring the relationship between the survival of a patient and several explanatory variables. It has a reliable performance on clinical data. A Cox model provides an estimate of the treatment effect on survival after adjustment for other explanatory variables. The sparse methods are utilized to optimized the coefficients of the Cox regression model and achieve feature selection. The selected attributes can be explained by domain knowledge and utilized in further research.

We also propose a unique problem in finding uncertain patients presented in this dataset. We define the notion of uncertainty in patients and explore the difference in behaviour of these patients with respect to certain labs. Such patients can be identified by our UPAD algorithm. The UPAD algorithm uses a SVM framework in its formulation.

SVM is used to build the classifier. We use LS-SVM to build the classifier, and the prediction results are compared with traditional machine learning results. The results show that demographic data source and lab data source perform better than other data sources, and in general, LS-SVM has better results compared to the traditional methods. The results are used to detect the noisy instances in the datasets. The SVM model is robust in identifying noisy instances in the dataset. Noisy (or uncertain) patients are distributed different with CPs (certain patient). Without the negative effect of such noisy patients, the prediction results can be improved.

## 5.2 Future Work

In the future, we plan to use Cox regression models within an ensemble framework. Ensemble forecasting is a method to do better prediction in an adaptive manner. This method will improve the prediction accuracy significantly and can potentially make the Cox regression model suitable for more formats of data sources. We also plan to build transfer learning models over healthcare clinical data which can be used on longitudinal data for further analysis.

In addition, we plan to also study the use of different kernels within SVM. Single kernel has a limited ability in dealing with complexed data. The multiple kernels can largely improve the accuracy and efficiency of SVM by introducing different kernels to different types of variables in a large dataset.

# REFERENCES

[1] N. I. Akhiezer, I. M. Glazman, and M. K. Nestell. *Theory of linear operators in Hilbert space*, volume 1. F. Ungar Publishing Company, 1961.

[2] R. Amarasingham, B. J. Moore, Y. P. Tabak, M. H. Drazner, C. A. Clark, S. Zhang, W. G. Reed, T. S. Swanson, Y. Ma, and E. A. Halm. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical care*, 48(11):981, 2010.

[3] L. Auria and R. A. Moro. Support vector machines (svm) as a technique for solvency analysis. *DIW Berlin Discussion Paper*, 2008.

[4] A. Ben-Hur and W. S. Noble. Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl 1):i38–i46, 2005.

[5] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[6] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[7] D. S. Brookshire, M. A. Thayer, J. Tschirhart, and W. D. Schulze. A test of the expected utility model: evidence from earthquake risks. *The Journal of Political Economy*, pages 369–389, 1985.

[8] F. Chauchard, R. Cogdill, S. Roussel, J. Roger, and V. Bellon-Maurel. Application of ls-svm to non-linear phenomena in nir spectroscopy: development of a robust and portable sensor for acidity prediction in grapes. *Chemometrics and Intelligent Laboratory Systems*, 71(2):141–150, 2004.

[9] W.-H. Chen, S.-H. Hsu, and H.-P. Shen. Application of svm and ann for intrusion detection. *Computers & Operations Research*, 32(10):2617–2634, 2005.

[10] K.-M. Chung, W.-C. Kao, C.-L. Sun, L.-L. Wang, and C.-J. Lin. Radius margin bounds for support vector machines with the rbf kernel. *Neural Computation*, 15(11):2643–2681, 2003.

[11] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[12] D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[13] M. J. Cox, B. Paley, M. Burchinal, and C. C. Payne. Marital perceptions and interactions across the transition to parenthood. *Journal of Marriage and the Family*, pages 611–625, 1999.

[14] N. Crichton. Cox proportional hazards model. *Journal of clinical nursing*, 11(6):723–723, 2002.

[15] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[16] P.-E. Danielsson. Euclidean distance mapping. *Computer Graphics and image processing*, 14(3):227–248, 1980.

[17] S. H. David Foster, Hanet Young. Risk-standardized rates for 30-day mortality and readmissions vary significantly by race, 2011.

[18] J. Fan, N. E. Heckman, and M. P. Wand. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90(429):141–150, 1995.

[19] M. A. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak. Majorization–minimization algorithms for wavelet-based image restoration. *Image Processing, IEEE Transactions on*, 16(12):2980–2991, 2007.

[20] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[21] B. R. Glasberg and B. C. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1):103–138, 1990.

[22] A. M. Gleason. Measures on the closed subspaces of a hilbert space. *The Logico-algebraic Approach to Quantum Mechanics: Historical evolution*, 5:123, 1975.

[23] M. Gönen and G. Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005.

[24] J. Gui and H. Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008, 2005.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[26] I. Hellström, K. E. Hellström, and G. A. Warner. Increase of lymphocyte-mediated tumor-cell destruction by certain patient sera. *International Journal of Cancer*, 12(2):348–353, 1973.

[27] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[28] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

[29] C. Hsu, C. Chang, C. Lin, et al. A practical guide to support vector classification, 2003.

[30] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.

[31] F. Jurie and M. Dhome. Hyperplane approximation for template matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):996–1000, 2002.

[32] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani. Risk prediction models for hospital readmission. *JAMA: The Journal of the American Medical Association*, 306(15):1688–1698, 2011.

[33] S. S. Keerthi and C.-J. Lin. Asymptotic behaviors of support vector machines with gaussian kernel. *Neural computation*, 15(7):1667–1689, 2003.

[34] F. J. Khawaja, N. D. Shah, R. J. Lennon, J. P. Slusser, A. A. Alkatib, C. S. Rihal, B. J. Gersh, V. M. Montori, D. R. Holmes, M. R. Bell, et al. Factors associated with 30-day readmission rates after percutaneous coronary intervention. *Archives of internal medicine*, 2011.

[35] K. I. Kim, K. Jung, and H. J. Kim. Face recognition using kernel principal component analysis. *Signal Processing Letters, IEEE*, 9(2):40–42, 2002.

[36] C. X. Ling, J. Huang, and H. Zhang. Auc: a statistically consistent and more discriminating measure than accuracy. In *International Joint Conference on Artificial Intelligence*, volume 18, pages 519–526. Lawrance Erlbaum Associates LTD, 2003.

[37] M. D. McHugh and C. Ma. Hospital nursing and 30-day readmissions among medicare patients with heart failure, acute myocardial infarction, and pneumonia. *Medical care*, 51(1):52–59, 2013.

[38] R. Moffitt. Unemployment insurance and the distribution of unemployment spells. *Journal of Econometrics*, 28(1):85–101, 1985.

[39] J. Rao and A. Singh. A ridge shrinkage method for range restricted weight calibration in survey sampling. In *Proceedings of the section on survey research methods*, pages 57–64, 1997.

[40] A. Schwaighofer. Svm toolbox for matlab. *website: http://www. igi. tugraz. at/aschwaig/software. html*, 2002.

[41] J. Shawe-Taylor and N. Cristianini. An introduction to support vector machines: and other kernel-based learning methods. *Cambridge University Press*, 29:136, 2000.

[42] T. Shumway. Forecasting bankruptcy more accurately: A simple hazard model*. *The Journal of Business*, 74(1):101–124, 2001.

[43] B. W. Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 97–99, 1981.

[44] J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[45] Y. P. Tabak, R. S. Johannes, and J. H. Silber. Using automated clinical data for risk adjustment: development and validation of six disease-specific mortality predictive models for pay-for-performance. *Medical care*, 45(8):789–805, 2007.

[46] H. Tanaka, S. Uejima, and K. Asai. Linear regression analysis with fuzzy model. *IEEE Trans.sys.man and Cyber*, 12(6):903–907, 1982.

[47] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[48] R. Tibshirani et al. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[49] G. Tripepi, K. J. Jager, F. W. Dekker, and C. Zoccali. Statistical methods for the assessment of prognostic biomarkers (part i): Discrimination. *Nephrology dialysis transplantation*, 25(5):1399–1401, 2010.

[50] J. Ye and T. Xiong. Svm versus least squares svm. In *Proc. international conference on artificial intelligence and statistics*, pages 640–647, 2007.

[51] Y. Yu. D-optimal designs via a cocktail algorithm. *Statistics and Computing*, 21(4):475–481, 2011.

[52] S. Zheng, J. Liu, and J. W. Tian. A new efficient svm-based edge detection method. *Pattern Recognition Letters*, 25(10):1143–1154, 2004.

[53] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

# ABSTRACT

## REGRESSION MODELS FOR READMISSION PREDICTION USING ELECTRONIC MEDICAL RECORDS

by

**YUDI NIU**

April 2013

**Advisor:** Dr. Chandan Reddy

**Major:** Computer Science

**Degree:** Master of Science

Hospital readmissions are not only expensive but are also potentially harmful, and most importantly, they are often preventable. Providing special care for a targeted group of patients who are at a high risk of readmission can significantly improve the chances of avoiding rehospitalization. Despite the significance of this problem, not many researchers have thoroughly investigated it due to the inherent complexities involved in analysing and estimating the inherent predictive power of such complex hospitalization records. In this thesis, we propose using support vector machines and survival analysis methods to analyse data collected from Electronic Medical Records (EMR). We define the notion of abnormal patients and understand how they affect the performance of classifiers. We use sparse methods with survival regression models to build clinical models which are suitable to apply on such complex clinical data. These models are compared with existing readmission models such as ADHERE, TABAK and logistic regression models. Finally, we provide inferences and conclusions on how to extend this work to build better regression models.