Risk Stratification of Cardiovascular Patients using a Novel Classification Tree Induction Algorithm with Non-Symmetric Entropy Measures

by

Anima Singh

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2011

Author ....
Department of Electrical Engineering and Computer Science
December 21, 2010

Certified by ....
John V. Guttag
Professor, Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ....
Terry P. Orlando
Chair, Department Committee on Graduate Students

Risk Stratification of Cardiovascular Patients using a Novel Classification Tree Induction Algorithm with Non-Symmetric Entropy Measures

by

Anima Singh

Submitted to the Department of Electrical Engineering and Computer Science
on Dec 21, 2010, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

Risk stratification allows clinicians to choose treatments consistent with a patient's risk profile. Risk stratification models that integrate information from several risk attributes can aid clinical decision making.

One of the technical challenges in developing risk stratification models from medical data is the class imbalance problem. Typically the number of patients that experience a serious medical event is a small subset of the entire population. The goal of my thesis work is to develop automated tools to build risk stratification models that can handle unbalanced datasets and improve risk stratification.

We propose a novel classification tree induction algorithm that uses non-symmetric entropy measures to construct classification trees. We apply our methods to the application of identifying patients at high risk of cardiovascular mortality. We tested our approach on a set of 4200 patients who had recently suffered from a non-ST-elevation acute coronary syndrome. When compared to classification tree models generated using other measures proposed in the literature, the tree models constructed using non-symmetric entropy had higher recall and precision. Our models significantly outperformed models generated using logistic regression - a standard method of developing multivariate risk stratification models in the literature.

Thesis Supervisor: John V. Guttag
Title: Professor, Electrical Engineering and Computer Science

Acknowledgments

I express my sincere gratitude to Professor John Guttag for supervising this thesis with great care and for his insightful suggestions. His guidance and support have been invaluable in helping me grow as a researcher during the past two years at MIT.

I am also thankful to Collin Stultz for his clinical insights and his technical inputs that has helped me throughout the course of my work.

Thanks to Jenna Wiens, Gartheeban Ganeshpillai, Ali Shoeb and Zeeshan Syed for being there to share and bounce off ideas.

Thanks to my good friends Anisha Shrestha, Neha Shrestha and Anunaya Pandey for being there during good and bad times, and helping me maintain my sanity.

Finally I dedicate my effort to my family. I am indebted to them for their love, support, encouragement and faith in me.

Contents

1 Introduction 17

1.1 Related work 18

1.2 Proposed Approach 20

1.3 Clinical Application 20

1.4 Contributions 21

1.5 Organization of Thesis 23

2 Background 25

2.1 Acute coronary syndrome 25

2.2 Post-ACS risk stratification 26

2.2.1 TIMI Risk score 26

2.2.2 Heart rate variability 27

2.2.3 Deceleration capacity 28

2.2.4 Morphological variability 29

3 Discretization 31

3.1 Overview 31

3.2 Proposed Entropy-based Supervised Discretization Algorithm 35

3.3 Entropy measures 36

3.3.1 Shannon entropy 37

3.3.2 Asymmetric entropy 37

3.3.3 Warped entropy 38

3.4 Experiments 38

3.4.1 Methodology 39
3.4.2 Effectiveness of Bagging 39
3.4.3 Stability of Cutoffs 41
3.4.4 Performance Evaluation of the Cutoffs 43
3.5 Summary 47

4 An Integrated Approach for Risk Stratification 49

4.1 Classification Trees 49
4.2 Using Classification Tree for Risk Stratification 52
4.2.1 Algorithm for Classification Tree Induction 53
4.3 Summary 60

5 Splitting criterion for classification trees 61

5.1 Notations 62
5.2 Types of splitting criterion 62
5.2.1 Hellinger distance 63
5.2.2 DKM 63
5.2.3 Entropy Measures 64
5.2.4 Class Confidence Proportion 64
5.3 Experiments 65
5.3.1 Methodology 65
5.3.2 Classification Trees with Global Discretization 66
5.3.3 Classification Trees with Local Discretization 69
5.3.4 Comparison of classification trees with global discretization versus local discretization 72
5.3.5 Evaluation of Classification Stability 74
5.4 Summary 77

6 Comparison of different models for post-ACS risk stratification 79

6.1 Methodology 79
6.2 Classification Tree Model versus TIMI 80

6.3 Classification Tree Model versus Linear Multivariate Logistic Regression Model . . . 83
6.4 Summary . . . 84

7 Understanding the rules of a classification tree **85**
7.1 Classification tree model for MERLIN . . . 85
7.2 Model Analysis and Discussion . . . 87
7.2.1 Analysis of Classification Rules . . . 87
7.2.2 Using the classification tree model in a medical setting . . . 88
7.3 Summary . . . 89

8 Conclusions and Future Work **91**
8.1 Summary . . . 91
8.1.1 Goals and Contributions . . . 91
8.2 Future Work . . . 94

List of Figures

3-1 Entropy measures. 39

3-2 Coefficient of Variance (COV) of a single cutoff generated using our discretization algorithm *with* Bagging and *without* Bagging for different continuous variables. 40

3-3 An example to illustrate why Shannon entropy measure may yield unstable cutoffs for some unbalanced distribution of classes. The x-axis represent sorted values for an arbitrary continuous variable and the y-axis correspond to the class label of the examples associated with the continuous values of the variable. Figures 3-3(a) and (b) show two possible instances of training set derived from the sample population. The class distribution is unbalanced with the minority group proportion of 7%. The dotted arrow and the solid arrow represent the position of the cutoff identified by Shannon and warped entropy measure respectively. Since Shannon entropy places equal weights on both minority and majority examples, redistribution of a few examples between instances causes the cut point to change drastically. On the other hand, warped entropy measure yields the same cut point. 42

3-4 The distribution of MERLIN population in different categories for literature cutoffs and warped cutoffs. 44

4-1 An example of a binary classification tree [53]. 50

4-2 Comparison of the class coherence measures for binary classification problems [53]. The x-axis is the proportion of examples that belong to class 1. 55

5-1 Pair-wise fractional Hamming distance between the predicted labels for the holdout patients using the 100 classification tree models generated using the 100 training samples. 76

5-2 Histogram of the fractional Hamming distance for every pairwise comparison of the rows of matrix D. 77

6-1 Comparison of the distribution of recall, precision and odds ratio for three different classification models. It shows the sorted values of each classification performance measures for different models. 82

7-1 The classification tree based risk stratification model generated from the MERLIN dataset. We used our algorithm with warped entropy-based local discretization and asymmetric entropy-based splitting criterion to construct the tree. The numbers above each node show the number of dead patients and the number of alive patients at each node i.e. *number of dead patients/number of alive patients.* The superscripted number next to the 'Low risk' or 'High risk' label is the rule number. For example, 'Low risk$_1$' means *the first low risk rule.* . . . 86

List of Tables

3.1 Coefficient of variance for a single cutoff using different entropy measures 41

3.2 Cutoffs for the continuous risk metrics 43

3.3 The mean recall and precision of the *NB classifier* built using the *same* number of cutoffs as in the literature. The percentage in the parentheses next to each method is the mean percentage of patients that were labelled as high risk in the test sequence. 46

3.4 Results from the paired t-test and the Wilcoxon paired samples signed rank test to compare recall and precision of NB classifier built using cutoffs derived from warped entropy measure with literature cutoffs. A positive mean difference indicates higher performance by the classifier using warped entropy cutoffs. 47

4.1 A 2×2 contingency table for two binary categorical variables. 56

4.2 A example contingency table for two categorical variables. 57

4.3 Contingency tables with with equal or stronger dependence between the variables 'Treatment' and 'Outcome', relative to Table 4.2. The probability of each contingency table is also shown. 57

5.1 Risk variables available to us for the patients in MERLIN. The variables that are highlighted had a significant ($p < 0.05$) univariate association with CVD within 90 days, in the MERLIN population and are used for the induction of decision tree. 66

5.2 The mean recall and precision of the classification trees with global discretization, using different splitting criteria. The mean percentage of patients that were labelled as high risk was about 27% for each of the splitting criteria. 67

5.3 The mean difference in recall and precision along with the p-values of paired samples t-test and the Wilcoxon test for classification trees using global discretization. The results show the comparison for every pair of splitting criteria. For example, the first table (**Hellinger vs.**) compares the recall and precision of classification trees obtained using Hellinger distance with all the other splitting criteria. A positive mean difference in that table suggests that Hellinger performed better than the other criterion. The highlighted rows indicated significantly different recall (or precision). For instance, in the third table (**Shannon vs.**), the two highlighted rows show that the recall of Shannon entropy-based trees is significantly higher than those of Hellinger-based trees and DKM-based trees respectively. 68

5.4 The mean recall and precision of the binary cutoff for different continuous variables using global and subset-specific cutoffs. The highlighted rows indicate the instances where the subset-specific cutoffs had an improved classification performance relative to the global cutoffs for the same variable. 70

5.5 The mean recall and precision of decision trees induced using local discretization and different types of splitting criteria. The mean percentage of patients that were labelled as high risk was 27% for all the criteria. 72

5.6 The mean difference in recall and precision along with the p-values of paired samples t-test and the Wilcoxon test for classification trees using local discretization. The results show the comparison for every pair of splitting criteria. For example, the first table (**Hellinger vs.**) compares the recall and precision of classification trees obtained using Hellinger distance with all the other splitting criteria. A positive mean difference in that table suggests that Hellinger performed better than that criterion. The highlighted rows indicated significantly different recall (or precision). For instance, in the first table (**Hellinger vs.**), the highlighted row shows that the difference in recall of Hellinger-based trees is significantly higher than of warped entropy-based trees respectively. 73

5.7 The best mean recall and precision of classification trees induced using different types of discretization. 74

5.8 Results from the paired t-test and the Wilcoxon paired samples signed rank test to compare recall and precision of the classification trees using asymmetric entropy with *local* discretization versus those of the classification trees using asymmetric entropy with *global* discretization. A positive value implies that trees with local discretization had a higher value. 74

6.1 The mean recall, precision and odds ratio of AsymCT model and TIMI risk score-based risk stratification model. 81

6.2 Results from the paired t-test and the Wilcoxon paired samples signed rank test to compare recall, precision and odds ratio of AsymCT with those of the TIMI risk score on the test sets. A positive value implies that AsymCT had a higher value than the TIMI score. 81

6.3 The mean recall, precision and odds ratio of the AsymCT model and the logistic regression model (LRM). 83

6.4 Results from the paired t-test and the Wilcoxon paired samples signed rank test to compare recall, precision and odds ratio of AsymCT with those of LRM on the test sets. A positive value implies that AsymCT had a higher value than LRM. 83

Chapter 1

Introduction

Risk stratification is an important classification problem in medicine. It refers to the task of predicting whether a patient is likely to develop a condition or experience a medical event in the future. For example, predicting the probability of having a heart attack. Accurate risk stratification allows clinicians to choose treatments consistent with the patient's risk profile. However, the diversity in clinical presentation of patients present a challenge.

At present, there exists a divide between therapy and diagnostics. This divide is evidenced by the example of implantable cardioverter defribrillators (ICDs). An ICD is a device that is surgically implanted in patients believed to be at a high risk of suffering a fatal arrhythmia. An estimated 90% of the patients who receive ICDs never end up needing them [16]. But every year hundreds of cardiovascular patients, without an ICD, die of a fatal arrhythmia. This illustrates a clear clinical need to develop tools and predictive models that can improve risk stratification.

Our work is motivated by an increasing availability of health information about patients in the form of electronic health records (EHR) from medical institutions. An EHR contains information such as patient demographics, progress notes, past medical history, vital signs, medications, immunizations, laboratory results and radiology reports [21]. Several of these individual measures have prognostic capabilities. Moreover, some of the measures are complimentary to each other. Therefore, information from several measures can be exploited to aid clinical decision making. EHRs allow easy accessibility to this vast amount of patient data. Therefore, they present one with the opportunity to utilize the data to create data-driven models that can improve risk stratification.

The goal of this thesis work is to develop automated tools to build risk stratification models that facilitate knowledge discovery and improve risk stratification. In this thesis, we specifically investigate classification tree based models for risk stratification of patients with non-ST elevation acute coronary syndrome (NTSEACS) from the MERLIN trial [40]. This database is a highly unbalanced data set with a 2% death rate.

Using the dataset from the MERLIN trial, we devise and evaluate an algorithm for classification tree. The classification tree model was derived from a training dataset. The model was then used to to risk stratify patients on a disjoint test set. Several instances of training and test sets were used to evaluate our algorithm. We speculate that our algorithm can be successfully applied to generate models for risk stratifying patients for other medical outcomes as well.

1.1 Related work

Researchers have investigated several methods to generate integrated risk stratification models for different medical outcomes.

The most commonly used technique in the literature is multivariate logistic regression [3, 5, 10]. A multivariate logistic regression yields a model characterized by regression coefficients. The values of the regression coefficients that are statistically significantly gives us a measure of association of the variables with the outcomes. One of the disadvantages of logistic regression is that it may not be able to identify complex interaction between variables that exist in the data. For example, the value of one variable (say, age) may substantially change the importance of another variable (say, history of smoking). With the vast amount of data, it is virtually impossible to include all potential interactions in a regression model. Moreover, logistic regression models are rarely used in clinical practice.

In a clinical setting, simple integrated risk scores are used. These simple and easy-to-use risk scores are derived from independent prognostic variables selected using multivariate logistic regression. The TIMI risk score, used for risk stratification of patients with non-ST elevation acute coronary syndrome (NTSEACS) [4], is an example of such a score. Its simplicity and the ease of use have made it appealing for clinical use.

Besides logistic regression, risk stratification models derived using machine learning algorithms are gaining popularity in the medical literature [29, 50, 56]. Complex models generated using machine learning techniques show promise in improving risk stratification. However, clinical applicability of complex 'black box'-type models is limited since they do not give any explanation for classification results predicted by the model. Examples of such machine learning techniques include Support Vector Machines (SVM) and Artificial Neural Networks (ANNs). Advanced models that can discover complex interactions between variables and also provide justifications for classification results are more well-suited for use in a clinical setting. A classification tree is an example of such a model.

Classification tree learning is a supervised data-driven algorithm that generates a predictive model comprised of a set of classification rules in the form of a tree. The rules extracted from the tree are simple to interpret. The main components of classification tree induction process are:

- **Splitting criterion:** The splitting criterion is used to select a variable for generating a split at a node of a tree during the tree induction process.

- **Discretization of continuous variables:** Classification tree can only handle discrete variables. Therefore, continuous variables need to be discretized.

- **Pruning:** Pruning helps to generate a model that is less susceptible to overfitting the training data and easier to interpret.

One of the technical challenges in constructing a classification tree based risk stratification model from medical data is the class imbalance problem. Typically the number of patients that experience a serious medical event is a small subset of the entire population. For example, amongst the patients with high cholesterol, only a small fraction of patients actually suffer heart attacks. Despite class imbalance, a good classification tree model should be able to extract useful information from the data that can be used for risk stratification.

The most popular classification tree induction algorithm uses Shannon entropy or the Gini index based splitting criterion during classification tree induction. These functions do not take into account the class distribution of the data. Therefore, in unbalanced datasets, they are biased towards the majority class.

1.2 Proposed Approach

We propose a tree learning algorithm for unbalanced data that uses non-symmetric entropy measures for both discretization and for the criterion to select variables to split during tree induction. Our algorithm performs local discretization during tree induction to yield subset specific cutoffs. To generate classification rules that are statistically significant, we use Fisher's exact test based method to prune the branches that are not statistically significant.

1.3 Clinical Application

While the methods developed in this thesis are potentially applicable to data from any application, we present and evaluate our methods in the context of risk stratification of patients for cardiovascular death.

Cardiovascular death is an important clinical area because of the high incidence of cardiovascular death in the US and around the world. In 2006, there were about 831,300 cardiovascular deaths in the US [2], which accounted for 36.9% of all deaths. The total costs, both direct and indirect, associated with addressing cardiovascular disease in US alone was about $366.6 billion [2]. These statistics highlight the need of improved tools to identify patients with cardiovascular disease at high risk of death.

1.4 Contributions

We briefly review the primary contributions of this thesis. A more detailed discussion of the contributions are deferred to the subsequent chapters in the thesis.

- **Concept of Warped Entropy:** We propose the concept of a warped entropy, a new non-symmetric entropy measure that takes into account the class distribution of the dataset. The classical entropy measure, Shannon entropy is maximized at the distribution where all the classes are present in equally proportions. The new measure allows the maximum entropy point to be set at a specified distribution.

- **Evaluation of symmetric and non-symmetric entropy measures for discretization of continuous attributes in an unbalanced dataset:** We evaluated Shannon entropy and two non-symmetric measures: warped entropy and asymmetric entropy [33] for discretization of continuous variables. All of the entropy measures were evaluated based on the stability of the cutoffs generated by each of the measures over several training instances. The stability of the cutoffs was measured using the coefficient of variance (COV) of the cutoffs. The cutoffs were generated for four continuous variables: age, morphological variability, heart rate variability and deceleration capacity using data from the MERLIN trial [40].

  We compare the COV for all the four continuous variables for each of the three entropy measures. Our results show that warped entropy measure not only yielded cutoffs with the smallest average COV but also had the best worst case COV in the MERLIN dataset.

- **Identification of the potential usefulness of subset specific cutoffs for risk stratification:** Typically, when a continuous risk attribute is used to risk stratify patients into a low risk or a high risk group, the same cutoff is used for patients, irrespective of their clinical background. For example, when using morphological variability (MV), an ECG based metric for risk stratification of patients with NTSEACS, a cutoff of 50 is suggested, regardless of whether or not a patient has ST-depression [51]. We hypothesized that using subset specific cutoffs - different cutoffs for the same continuous risk attribute based on clinical background of the patients, can improve risk stratification.

We demonstrated that for several continuous variables, subset specific cutoffs can yield improved performance in identifying patients at high risk of CVD. For each of the available continuous risk metrics from the MERLIN trial, we found 'global' cutoffs using the entire training set and subset specific cutoffs based on a clinical attribute such as presence or absence of ST-depression. We evaluated the cutoffs on a disjoint set of test examples. Our results show that subset specific cutoffs can improve both recall and precision in identifying high risk patients.

- **Present a novel approach of building classification trees from highly unbalanced dataset**: We present an algorithm for development of classification trees from unbalanced data. Our algorithm performs binary discretization (*with* Bagging) of continuous variables at each node during the tree induction process. This procedure exploits the usefulness of subset specific cutoffs for risk stratification. Given a class coherence measure, the algorithm selects the variable that has maximizes class coherence (or minimizes class incoherence) after a split.

- **Evaluation of different class coherence measures for induction of classification tree models, with local discretization, from unbalanced datasets:** We evaluated our algorithm using different class coherence measure-based splitting criteria. The classification trees developed using asymmetric entropy or class-confidence proportion (CCP) as the splitting criterion yielded post-ACS risk stratification models with significantly higher recall and precision than other splitting criteria proposed in the literature for unbalanced data. Moreover, our results suggest that using asymmetric entropy to generate classification tree yields models with significantly higher recall than those generated using CCP, with a comparable precision. We also compared the performance of these classification models with those generated using global discretization. Our results show that the recall and precision obtained using asymmetric entropy using local discretization is significantly higher than the best recall and precision obtained using global discretization.

- **Comparison of classification tree based models with other models:** We compared the risk stratification performance of the classification tree based models developed using non-symmetric entropy measure with those of the TIMI risk score. Our results show that the classification tree models yield a significantly higher recall, precision and odds ratio than the TIMI risk score. It is important to note that the TIMI risk score and the classification tree models uses different set of risk attributes with some overlap.

  We also developed multivariable logistic regression models (LRM) using the *same* set of variables used of the classification tree model. Again, the performance of the classification tree based models were significantly higher than LRMs.

1.5 Organization of Thesis

The remainder of the thesis is organized as follows. Chapter 2 provides background and an overview of the clinical application of interest - risk stratification of post-ACS patients. Chapter 3 presents and evaluates different symmetric and non-symmetric entropy measure for discretization of continuous variables. Chapter 4 describes classification trees and presents the algorithm used to construct classification trees. In Chapter 5, we evaluate classification trees obtained using global and local discretization with different splitting criteria. In Chapter 6, we compare the risk stratification performance of classification tree models with the TIMI risk score and models developed using logistic regression. In Chapter 7, we present a classification tree model generated by our proposed algorithm to risk stratify patients with NSTEACS. We also analyze the classification rules derived from the tree. Chapter 8 concludes with a summary and a discussion of future work.

Chapter 2

Background

In this chapter, we provide background information about acute coronary syndrome (ACS) and post-ACS risk stratification techniques. We start with a discussion of ACS and its subclassifications in Section 2.1. In Section 2.2, we review existing risk stratification methods to identify high risk patients. The goal of this chapter is to provide background of the clinical application on which the methods presented in this thesis are evaluated.

This chapter borrows heavily from the discussion of these subjects in [49, 52].

2.1 Acute coronary syndrome

An acute coronary syndrome (ACS) is a cardiac event in which blood supply to a part of the myocardium (heart muscle) is blocked or severely blocked. This leads to ischemia or cell death in the myocardium. An ACS is also subclassified based on the extent to which the coronary artery is occluded, which can often be inferred from ECG recordings. An ECG showing elevation in the ST segment is indicative of complete occlusion of an artery and necrosis (and therefore, myocardial infarction). Such patients are given a diagnosis of ST-elevation MI (STEMI) and are typically higher risk relative to patients with non-ST-elevation ACS. Non-ST-elevation ACS (NSTEACS) is indicative of partial occlusion of an artery and is a less severe condition. NSTEACS may be diagnosed by the presence of certain ECG irregularities (ST depression or T wave inversion).

2.2 Post-ACS risk stratification

Patients who have experienced ACS are at elevated risk of death. Among patients stabilized after ACS, some patients are at higher risk than others. Therefore, post-ACS risk stratification is an important clinical step to manage patients and to guide their treatments. This section provides background information on post-ACS risk stratification methods that are considered in the thesis. In particular, we consider the TIMI risk score (TRS) and long-term ECG-based techniques. There exists other risk stratification measures, such as echocardiography, that are not discussed here.

2.2.1 TIMI Risk score

The TIMI Risk score [4, 38, 39] is a simple risk stratification tool that incorporates several binary clinical risk variables that are easily acquired at the time of admission. It is designed to be simple so that it can easily be evaluated by a medical personnel without the use of a computer. It can therefore be used for immediate decision making with regard to treatment options. The GRACE [19] risk score performs a similar function.

The TRS considers the following 7 binary predictor variables:

- Age 65 years or older,
- At least 3 risk factors for coronary artery disease among the following: hypertension, hypercholesterolemia, diabetes, family history of coronary artery disease, or being a current smoker,
- Prior coronary stenosis (narrowing of an artery) of 50% or more,
- ST-segment deviation on ECG at presentation,
- Severe anginal symptoms (at least 2 anginal events in prior 24 hours),

- Use of aspirin in prior 7 days, and
- Elevated serum cardiac markers (CK-MB or troponins)

The TIMI11B and ESSENCE trials [4] showed that higher TRS is associated with cardiovascular death and other adverse cardiac events. Based on the TRS, patients are further grouped in three categories: low risk (TRS = 1,2), moderate risk (TRS=3,4) and high risk (TRS = 5,6,7) [40].

2.2.2 Heart rate variability

The ECG-based risk stratification techniques that has been most extensively discussed in the literature is based on measurements of heart rate variability (HRV) [1, 25]. The physiological theory behind using heart rate variability is that in a healthy individual, the body is able to adjust the heart rate to compensate for the changes in oxygen demand during different daily activities. A heart rate that changes little, is not able to respond to such demands. This suggests that the heart and its control mechanisms are not actively responding to the stimuli. HRV based measures attempt to quantify the change in a patients instantaneous heart rate over a period of monitoring in order to yield an estimate of risk.

Heart rate is modulated by the autonomic nervous system, which is comprised of the sympathetic and the parasympathetic nervous systems. The sympathetic nervous systems effects are mediated by the release of epinephrine and norepinephrine, which increases the heart rate. The parasympathetic nervous system lowers heart rate by release of of acetylcholine by the vagus nerve. Decreased vagal or parasympathetic modulation (i.e. reduced downregulation of heart rate) is thought to be strongly linked to increased risk of death [9, 48] . One possible explanation is that reduced down-regulation corresponds to an increase in heart rate which imposes stress on heart muscle already affected by ischemia or infarction.

HRV measures are calculated from the sequence of intervals between two consecutive QRS complex of the ECG signal [1] . Abnormal beats are ignored, since the focus of HRV is to study how the nervous system modulates heart rate. While abnormal beats change the heart rate, these changes are the result of a different physiological phenomenon (e.g., the presence of abnormal beat foci) and are ignored so as not to be confused with heart rate changes caused by impulses from the nervous system. Since only heartbeats resulting from normal depolarization of the SA node are considered, the sequence of R-wave to R-wave (RR) intervals studied for HRV analysis is termed the NN (for normal-to-normal) series. One of a number of methods is then used to summarize this series with a single number. HRV measures can be roughly divided into time domain, frequency domain, and nonlinear measures. [1] provides a more complete overview of HRV metrics.

Frequency domain HRV methods rely on the fact that vagal and sympathetic activity are mediated by biochemical pathways associated with different time scales [1]. In particular, acetylcholine (which mediates vagal activity) is faster acting than epinephrine and norepinephrine (which mediate sympathetic activity). As a result, it is believed that changes in heart rate in the high frequency (HF) range (0.15-0.40 Hz) correspond to vagal activity, while changes in heart rate in the low frequency (LF) range (0.04-0.15 Hz) correspond to sympathetic activity.

One of the most commonly used frequency domain metrics, LF/HF, is defined as the ratio of the total power at LF and HF frequencies in the power spectral density (PSD) of the NN series. The PSD of the NN series is usually calculated using the Lomb-Scargle periodogram [31], which is designed to estimate the frequency content of a signal that is sampled at irregular intervals. The LF/HF ratio is computed for 5-minute windows, as in [1], and the median value across windows is used as the LF/HF value for that patient. Patients with low HRV-LF/HF are considered to be at risk. In the literature, HRV-LF/HF has performed better at identifying patients who are at risk of death due to coronary heart disease than the time domain measures [54].

2.2.3 Deceleration capacity

Like HRV, Deceleration capacity (DC) attempts to measure impaired vagal modulation of heart rate, believed to be associated with high risk. The underlying theory behind the DC measure is that the vagal and sympathetic activity of the heart can be distinguished because the former causes deceleration of heart rate while the latter causes heart rate acceleration [6].

To compute DC, we begin with the RR interval sequence RR[n] and search for *anchors*, i.e., RR intervals that are longer than the ones preceding them. Let the index of the i^{th} anchor be n_i and the total number of anchors be N. If we define $X[n]$ as the average RR interval length around each anchor i.e.

$$X[n] = \frac{1}{N}\sum_{i=1}^{N} RR[n_i + n]$$

DC is then computed from this information as:

$$DC = \frac{(X[0] + X([1]) - (X[-1] + X[-2])}{4}$$

2.2.4 Morphological variability

Morphological variability is a measure of subtle changes in the shape of signals over a period of time. The theory behind measuring MV is that in a stationary and homogenous cardiac conducting system, the activated pathways through excitable cardiac cells are usually similar for consecutive cardiac cells. However, in the presence of ischemia, the conducting system may have multiple islands of depressed and relatively unexcitable myocardium [15] that lead to discontinuous electrophysiological characteristics [23]. The presence of several possible adjacent pathways that can invade the nonfunctioning area leads to variations in the spatial direction of the invading vector [8]. The overall effect of such minor conduction inhomogeneities is not well understood but it could potentially correlate with myocardial instability and hence have predictive value for adverse cardiovascular outcomes.

MV quantifies the inhomogeneities and subtle morphological variation using a variant of dynamic-time warping to compute time-aligned morphology changes between consecutive sinus beats [52]. The original ECG signal is then transformed into a sequence of energy differences, the morphological distance (MD) time series, between consecutive beats. The MD time series is smoothed using a median filter. Finally the power spectral density of the MD time series between 0.30-0.55 Hz is obtained for a five minute time window using the Lomb-Scargle periodogram [31]. The power spectral density is computed for every uncorrupted 5 minute windows from 24 hours of continuous ECG signal. The 90^{th} percentile value of the power spectral density is used as the MV value for that patient [49]. A more detailed discussion of MV can be found in [52].

Chapter 3

Discretization

In this chapter, we present our work on discretization of continuous variables using entropy-based measures. We investigate two existing approaches, Shannon entropy and asymmetric entropy, and one new approach, warped entropy. Both asymmetric entropy and warped entropy take into account the distribution of the classes in the dataset during discretization but Shannon entropy does not. Since there is typically a substantial class imbalance in medical datasets, this difference in significant. We evaluate the performance of the cut points generated using each of the measures in terms of stability and their classification performance using a Naive Bayes classifier.

3.1 Overview

In medicine, multiple risk metrics are used to evaluate the risk profile of a patient. These risk metrics consist of both continuous (e.g. age) and categorical (e.g. history of diabetes) variables. From a clinical perspective, categorization of continuous variables into discrete categories (e.g. high risk) is useful since it offers a simple risk stratification tool for both physicians and patients. In addition, categorization, more generally referred to as discretization, of continuous risk variables is also important to generate multivariable models using algorithms, e.g. decision tree induction methods, that can only handle discrete variables. Moreover, many machine learning algorithms are known to generate better models when discretized variables are used [26].

Discretization of continuous risk variables requires partitioning continuous variables into sub-ranges where each sub-range can be treated as a category. The purpose of discretization is to identify cutoffs that partition the range into intervals such that members of each category exhibit good class coherence.

In recent years, because of advances in medical research and extensive use of computational techniques for medical data analysis, several new risk metrics have been developed [51, 43]. These metrics show promise in improving risk stratification of patients for adverse outcomes. However, categorization of these newly developed risk variables is a challenge, since there is little prior expert knowledge. Therefore, there is a need to develop algorithms that can identify informative cut points for a continuous valued variable.

Discretization algorithms are based on assumptions about the distribution of values of the continuous variable and the outcomes associated with them. They assume that there exists a correlation between the variable and the outcome. Equal width technique is commonly used discretization technique. This method hope to find clusters of observations with the same class labels along the values of the variable by defining bins within the range of the variable. The equal width discretization method splits the range of the continuous variable into the user-defined number of equal width discrete intervals. The obvious disadvantage of this algorithm is that in cases when the outcome are not distributed evenly within each bin, a lot of information is lost after discretization. This technique is generally referred as class-blind or unsupervised method. Percentile based approach is often used to find cutoffs for continuous risk variables for binary class values. The algorithm uses the class labels associated with the values to determine whether there exists a positive or a negative correlation between the continuous attribute and the classes. Depending on the direction of correlation, either the top or bottom x percentile is used to find the binary cut point. Although this method uses the class label information to determine the direction of linear correlation, the class labels are not directly used to determine the cut point itself.

Supervised methods, on the other hand, use class label information to identify cutoffs. The discretization cutoffs are derived from training data which consists of examples, each represented by a value of the continuous variable and a class label associated with the example. While unsupervised methods such as equal width and equal frequency methods hope to find clusters by binning, supervised methods use the class label information from the training set to guide the search for such clusters. If any clusters are identified, the values of the variable that define the cluster are identified as cut points. Previous research has indicated that supervised discretization methods are better than unsupervised methods for classification using machine learning algorithms such as decision trees and Naive Bayes classifier [14, 28]. In [28], Kurgan et al. show that decision tree classifiers built from discretized data using supervised methods yielded higher accuracy than the classifiers generated using unsupervised discretization methods. The algorithms were tested on both medical and non-medical datasets.

One of the popular supervised discretization approaches is based on measuring class entropy of a sequence using Shannon entropy [26]. Class entropy of a sequence is a measure of uncertainty of the class labels of the examples that belong to the sequence. It is a measure of information where a lower value of entropy corresponds to higher amount of information. Entropy based discretization algorithms evaluate every candidate cut point based on a joint measure of the entropy of the two resulting subsequences generated by the cut point. The candidate cut point that minimizes the joint entropy of the class in the subsequences is chosen as the 'best' cut point.

According to Shannon entropy, a sequence that has an equal class distribution is considered to be the most entropic. This property makes an implicit assumption about equal prior probabilities of classes. However, in most of the medical applications, prior distribution of classes is highly unbalanced. In case of risk stratification for cardiovascular deaths, deaths (positive outcome) are much less represented in the datasets than non-deaths (negative outcome). If we use Shannon entropy for discretization in such highly unbalanced datasets, a subsequence with an equal class distribution of positive and negative outcomes is assigned the maximum uncertainty value of 1. According to Shannon entropy, such a subsequence contains the least possible information about the class label of the examples in it. However, for highly unbalanced datasets, such a subsequence actually provides a lot of information. It suggests that patients that belong to the subsequence are at relatively high risk of cardiovascular death. Shannon entropy fails to convey this information. Therefore, a cut point that generates subsequences that contain a lot of information about the class might be rejected when identifying a cut point with the least joint Shannon entropy. Taking this under consideration, we present two entropy measures: asymmetric entropy and warped entropy that take into account the distribution of classes in the dataset. In Section 3.3, we discuss all three entropy measures.

While supervised methods have advantages over unsupervised methods, they are prone to overfitting. In supervised learning, overfitting occurs when the model learned from the training set describes the training data well but does not generalize well on unseen data. Overfitted models are generally a consequence of noise in the training data. When we use the training set for learning, we hope to derive a best cutoff that can be usefully applied to new unseen examples. However, an optimal solution based on a single training sample set may not necessarily be optimal for other data [46]. With this in consideration, in [46], Qureshi et.al present a resampling approach to reduce the discretization variance during the learning process. We propose a supervised entropy-based discretization approach that integrates a similar resampling approach to address the problem of overfitting. We present the proposed discretization algorithm in Section 3.2.

In Section 3.4, we evaluate the stability of cutoffs generated using different entropy measures. We also compare the recall and the precision of Naive Bayes Classifiers generated using cutoffs proposed in the literature and the cutoffs generated using entropy-based methods.

3.2 Proposed Entropy-based Supervised Discretization Algorithm

We propose a supervised discretization algorithm that finds a user-specified number of cutoffs for a continuous variable.

Let V be a continuous variable. Let Ω be a sequence of N examples sorted in an ascending order of the continuous variable. Each example is a pair $[v, l]$ where v is a value of the continuous variable and l is the class label.

We find the midpoint of the value of variable V for each successive pair of examples in Ω. These midpoint values are the candidate cut points. Each candidate cut point C partitions Ω into two sequences, Ω_1 and Ω_2, where Ω_1 contains examples with $v < C$ and Ω_1 contains examples with $v >= C$. Next, we find the class entropy of each subsequence Ω_j using an entropy measure. We use **W**eighted **J**oint **E**ntropy (WJE) to evaluate the quality of the partition generated by a candidate cut point C.

$$\mathbf{WJE}(C, \Omega) = \frac{|\Omega_1|}{|\Omega|} H(\Omega_1) + \frac{|\Omega_2|}{|\Omega|} H(\Omega_2) \tag{3.1}$$

Here, $H : \mathcal{M} \rightarrow \Re$, represents an entropy measure. $\mathcal{M}$ is an $N \times 2$ matrix containing N examples with the corresponding labels.

Amongst all the candidate cut points, C which minimizes $\mathbf{WJE}(C, \Omega)$ is selected as the cut point for binary discretization for Ω.

Equation 3.1 can be easily generalized to generate n cutoffs. However, the set of candidate cutoffs that needs to be evaluated is $O(N^n)$. To reduce the running time, we use a greedy approach. To find n cutoffs for $n > 1$, first we perform a binary split on the entire sequence Ω to identify the first cutoff. To find the next cutoff, we identify the subsequence Ω_{MaxEnt} of Ω which has the maximum class entropy (uncertainty). A binary split is then performed on the subsequence by picking the cutoff that minimizes $\mathbf{WJE}(C, \Omega_{MaxEnt})$. This process is repeated until n cutoffs are found.

Once the cutoff(s) are identified, to reduce the probability of overfitting to the sample set but instead generalize well on unseen examples, we perform a Bagging (**B**ootstrap **agg**regat**ing**) algorithm. Bagging is an algorithm used to acquire robust estimation in machine learning applications. In the context of discretization, we want to avoid overfitting so that the cutoffs identified for a continuous variable are applicable to unseen examples. Given a sample sequence of size N, bagging generates r new training sequences, also called replicates, of size N by uniformly sampling examples with replacement from the original sample sequence [7].

Once the cutoffs for all the replicates are identified for a fixed number of cutoffs n, we take the median of the distribution to identify the final cut point. The psuedo code for the algorithm for n cutoffs is presented in Algorithm 1.

Algorithm 1 Proposed Discretization Algorithm for a **n** cutoffs with Bootstrapping

```
Inputs: Sequence $\Omega$, Number of cutoffs n, Number of replicates r
Output: Cutoffs $\mathbf{finalC}_{1\times n}$
Generate $r$ replicates $replicate$
for each $replicate(i)$ do
    Maximum Entropy Subsequence $R_{maxEnt}$= $replicate(i)$  {By default}
    $j$=1 {By default}
    while $j <= n$ do
        Find candidate cut points $C$ from $R_{maxEnt}$
        for each candidate cutoff $C(m)$ do
            Calculate weighted joint entropy, $WJE(m)$
        end for
        Find binary cutoff with minimum WJE, $binaryC(i)$
        $bestC(i, j) = binaryC(i)$
        Partition $replicate(i)$ in subsequences using $bestC(i, \bullet)$
        $R_{maxEnt}$ = the subsequence of $replicate(i)$ with maximum entropy
        Increment $j$ by 1
    end while
end for
$finalC$ = median($bestC$)
Return $finalC$
```

3.3 Entropy measures

We present three different types of entropy measure. First we discuss Shannon entropy, a symmetric entropy measure, and then we present two different types of non-symmetric entropy measures, Asymmetric entropy and Warped entropy. The entropy measures can be generalized to the case of k class labels, however, we restrict our discussion to two class labels for ease of explanation.

3.3.1 Shannon entropy

Shannon entropy [?] is the most commonly used entropy measure. Let the class-label variable L take two different values, i.e. $l = 2$, with probability p_1 and p_2 respectively. The Shannon entropy of a subsequence S, with class distribution $p_1(S)$ and $p_2(S)$ is given by

$$H(S) = -p_1(S).log_2 p_1(S) - p_2(S).log_2 p_2(S) \tag{3.2}$$

Shannon entropy is a symmetric entropy measure and is maximized when the two classes in a S are present in equal proportions (Figure 3-1).

3.3.2 Asymmetric entropy

For the binary class case, the asymmetric entropy measure of a subsequence S derived from a parent sequence P is given by,

$$H(S,P) = \frac{p_1(S).p_2(S)}{(-2.z_1(P)+1).p_1(S) + (z_1(P))^2} \tag{3.3}$$

where, $p_1(S)$ and $p_2(S)$ are defined as above in Section 3.3.1. The variable z_1 is a function of the parent sequence P such that $z_1(P) = p_1(P)$. The value of z_1 determines the asymmetry of the entropy measure. Specifically, for a given parent sequence P, the function $H(S,P)$ is maximized when $p_1(S) = z_1(P)$.

By setting z_1 to the probability of class 1 in the parent sequence, we are essentially considering the distribution of the parent sequence to be the most uninformative. Any subsequence with $p_1(S) = z_1(P)$ has the same distribution as the parent sequence. Therefore, it does not provide any additional information and is assigned the maximum entropy value of 1 (Figure 3-1).

The concept of asymmetric entropy was first introduced by Zighed et.al. [33].

3.3.3 Warped entropy

This entropy measure is a modified or warped version of Shannon entropy (Section 3.3.1). In Section 3.1, we motivated the utility of asymmetry based on the fact that the prior distribution of classes is highly unbalanced in many medical datasets. One way to deal with the class imbalance is to assign greater weights to examples from the minority class than to those from the majority class so that the distribution of the weighted samples is balanced.

The warped entropy measure of subsequence S derived from a parent sequence P is given by,

$$H(S,P) = -\sum_{l=1}^{2} p_l^*(S,P).log_2 p_l^*(S,P) \tag{3.4}$$

where,

$$p_l^*(S,P) = \frac{p_l(S).w_l(P)}{w_1(P).p_1(S) + w_2(P).p_2(S)} \tag{3.5}$$

The variables w_1 and w_2 are weights assigned to examples of class 1 and 2 respectively. Specifically, $w_l(P) = \frac{z_1(P)}{z_l(P)}$ where $z_l(P) = p_l(P)$ as defined in Section 3.3.2.

3.4 Experiments

We performed all of our experiments on data from the MERLIN-TIMI 36 trial [40]. We used data from 4219 non-ST elevation acute coronary syndrome (NSTEACS) patients and considered cardiovascular death within 90 days as an endpoint. There were 83 ($\approx$ 2%) cardiovascular deaths within 90 days. The database contains continuous ECG signals recorded at 128 Hz within 48 hours of admission due to NSTEACS. Three electrocardiographic (ECG) risk metrics: heart rate variability (HRV) [1] , deceleration capacity (DC) [6] and morphological variability (MV) [51] were computed from the first 24 hours of ECG recording. For HRV, we computed HRV-LF/HF [1]. We measured the HRV and DC for each patient using the HRV Toolkit available from Physionet [18] and the libRASCH software provided by Technische Universitat Mnchen [6] respectively. MV was computed using the algorithm described in [51]. In

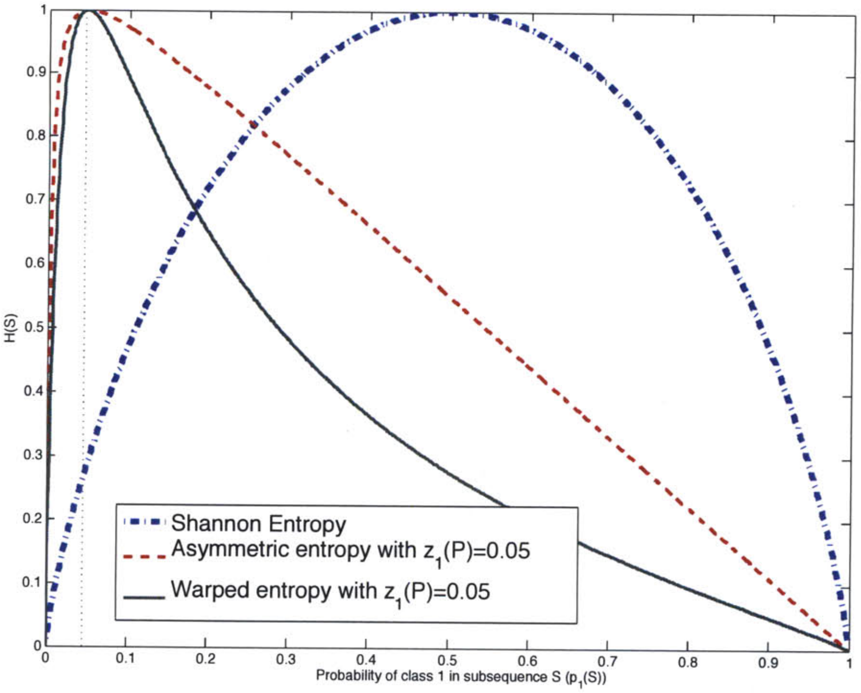

Figure 3-1: Entropy measures.

addition to these continuous ECG based metrics, we also use age, a continuous clinical variable, for the evaluation of the different entropy measures for discretization.

3.4.1 Methodology

For our experiments, we used 100 different pairs of training and test sets. Each pair was created using the holdout method where 4219 patients from MERLIN was partitioned into two disjoint sets. Each training set contains 2813 patients and its corresponding test set contains a disjoint set of 1406 patients.

3.4.2 Effectiveness of Bagging

In this experiment, we compare the stability of the cutoffs generated for each of the 100 training sets *without* Bagging with those of the cutoffs generated *with* Bagging. Since Bagging is performed to reduce overfitting to the training set, we would expect the cutoffs to be more stable than when the cutoffs are found *without* Bagging. We

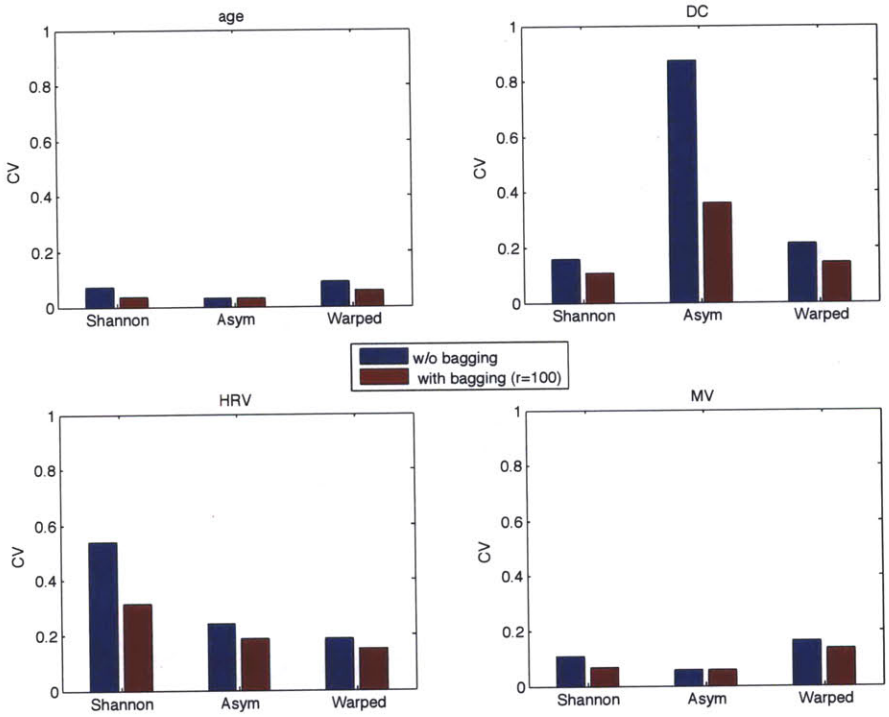

Figure 3-2: Coefficient of Variance (COV) of a single cutoff generated using our discretization algorithm *with* Bagging and *without* Bagging for different continuous variables.

use coefficient of variance (COV) as a measure of stability of the cutoffs. COV is a normalized measure of dispersion and is defined as the ratio of standard deviation (σ) to the mean (μ).

$$COV = \frac{\sigma}{\mu} \tag{3.6}$$

To evaluate the effectiveness of Bagging, we implemented our discretization algorithm *with* Bagging and *without* Bagging on 100 training sets. Binary discretization was performed on four continuous variables: age, DC, MV and HRV-LF/HF of each training set using all three entropy measures. When the algorithm was implemented with Bagging, $r = 100$ replicates were used.

Figure 3-2 shows the average coefficient of variance (COV) of the cutoffs for each variable. Lower values of COV corresponds to more stable cutoff values. On average, cutoffs found using Bagging has a lower COV than when Bagging is not used for each of the four variables. This illustrates the effectiveness of Bagging in reducing the variance of estimation of the cutoffs. Therefore, for all the following experiments that involve discretization, we perform Bagging (with $r = 100$ replicates).

As seen in Figure 3-2, the improvement in stability varies among different variables and different entropy measures. The discussion about the stability of cutoffs for different entropy measures is presented in Section 3.4.3 below.

3.4.3 Stability of Cutoffs

The COV for a single cutoff obtained with Bagging using different entropy measures is presented in Table 3.1. The worst (highest) COV for each entropy measure is highlighted in the table. Although warped entropy does not always have the least COV, it yielded the cutoffs with the best worst case COV compared to Shannon and asymmetric entropy. Among the three entropy measures, Shannon and warped entropy had the smallest average COV of 0.13.

Table 3.1: Coefficient of variance for a single cutoff using different entropy measures

	Coefficient of variance		
Risk Metric	Shannon	Asymmetric	Warped
Age	0.04	0.03	0.06
DC	0.11	**0.36**	**0.15**
HRV LF-HF	**0.31**	0.18	**0.15**
MV	0.06	0.06	0.14
Mean	0.13	0.16	0.13

The high instability exhibited by the asymmetric entropy measure was caused by its sensitivity to outliers. The outlier sensitivity can be explained by the shape of the asymmetric entropy function when a dataset is highly unbalanced (see Figure 3-1). The asymmetric entropy curve falls sharply as $p_1(S)$ approaches 0 from $p_1 = z_1 = 0.05$, but the rate of decrease is slow when we move away from the maximum entropy point towards $p_1 = 1$. The latter property causes the entropy (uncertainty) to still be high for $p_1 > z_1$. Therefore, the evaluation function favors cutoffs where one of the subsequences has $p_1 < z_1$. This makes asymmetric entropy measure susceptible to outliers from the minority class.

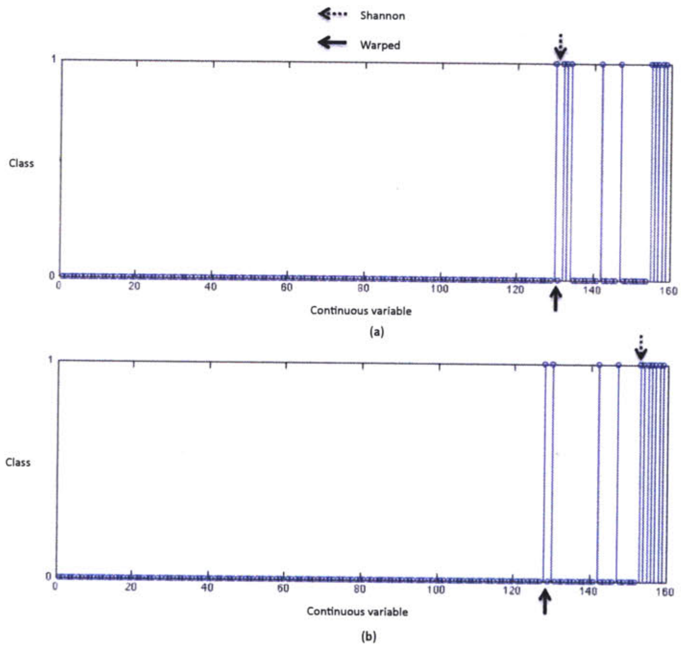

Figure 3-3: An example to illustrate why Shannon entropy measure may yield unstable cutoffs for some unbalanced distribution of classes. The x-axis represent sorted values for an arbitrary continuous variable and the y-axis correspond to the class label of the examples associated with the continuous values of the variable. Figures 3-3(a) and (b) show two possible instances of training set derived from the sample population. The class distribution is unbalanced with the minority group proportion of 7%. The dotted arrow and the solid arrow represent the position of the cutoff identified by Shannon and warped entropy measure respectively. Since Shannon entropy places equal weights on both minority and majority examples, redistribution of a few examples between instances causes the cut point to change drastically. On the other hand, warped entropy measure yields the same cut point.

The high instability shown by cutoffs derived from Shannon entropy measure (Table 3.1) can be attributed to the fact that it places equal weights on both minority and majority examples despite their unbalanced prior distribution. An example that illustrates this characteristic is shown in Figure 3-3.

3.4.4 Performance Evaluation of the Cutoffs

Next, we compared the cutoffs found using an entropy measure with those found in the literature for risk stratification of NTSEACS patients (Table 3.2). Because of its robust performance relative to other entropy measures, only the warped entropy measure is used for this experiment.

Table 3.2 show the cutoffs that are used in the literature for the four variables along with the cutoffs derived using warped entropy on the 4219 patients from the MERLIN dataset.

Table 3.2: Cutoffs for the continuous risk metrics

	Cutoffs	
Risk Metric	Literature	Warped
Age	65	60
DC	2.5 , 4.5	4.0, 6.0
HRV LF-HF	0.95	2.0
MV	50	40

The distribution of patients in the MERLIN data set for the different cutoffs is shown in Figure 3-4. Since our method used weighted warped entropy to find the cutoffs, the population size in each category is more even than for the literature cutoffs. As we'll see later, this characteristic of the warped cutoffs results in classifiers with higher recall at an expense of a smaller precision than those of the literature cutoffs.

For the performance evaluation of cutoffs, we used 100 instances of disjoint training and test sequences. We built two Naive Bayes (NB) classifiers[1] [37] from each of the training sequences using the cutoffs from the literature for one and the cutoffs derived using the warped entropy measure for the other. A NB classifier is a probabilistic classifier. Therefore, for each example in the test set, it generates a probability of death given the cutoffs of all four risk metrics: age, DC, HRV and MV. We used the death rate of the population ($\approx$2%) as the threshold such that patients

[1] The NB classifier was built using Bayes Net Toolbox by Kevin Murphy available at http://code.google.com/p/bnt/.

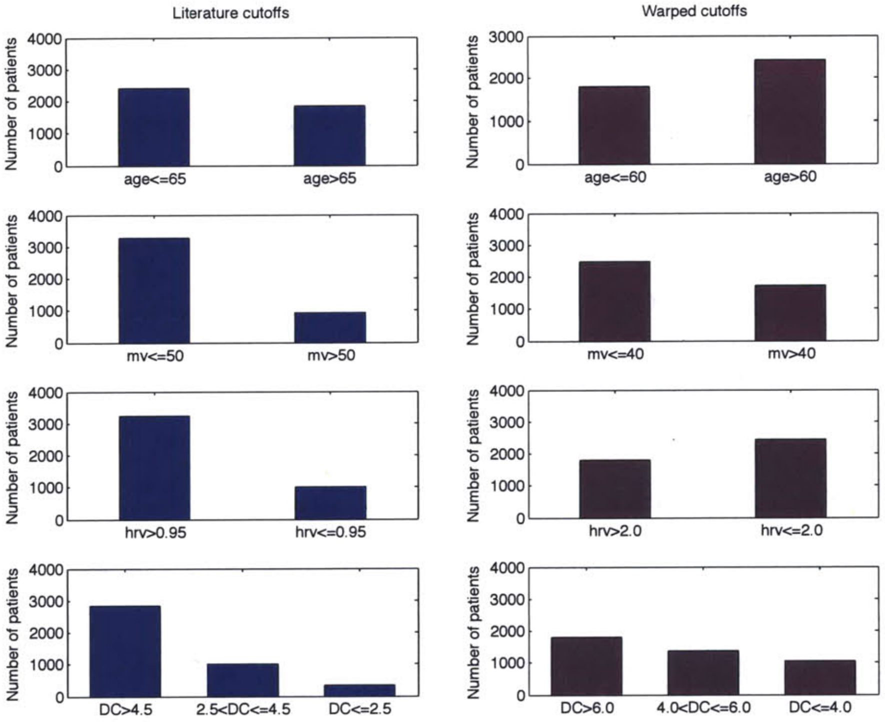

Figure 3-4: The distribution of MERLIN population in different categories for literature cutoffs and warped cutoffs.

with probability of death>2% were considered as high risk. The number of cutoffs derived using the entropy measure was the same as those used in the literature for risk stratification for cardiovascular deaths.

We evaluate the performance of the NB classifiers built from the training sequences on the corresponding test sequences based on recall and precision on the minority class:

$$Recall = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{3.7}$$

$$Precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{3.8}$$

While evaluating the performance of a classifier, it is important that both recall and precision are considered simultaneously. This is because there is an inverse relationship between them. For example, a classifier can often increase its recall by labeling most of the unseen patients as positive at the expense of large false positive rate. Therefore, it is inappropriate to evaluate classification models using only one of the measures in isolation.

We compare the recall (or precision) of the NB classifiers obtained each different types of cutoffs using the paired samples t-test [27] and the Wilcoxon paired samples signed rank test [58]. We consider the difference in recall (or precision) of the NB classifiers obtained using different cutoffs to be statistically significant if *both*, the paired t-test and the Wilcoxon test, yield p-values<0.05.

The Paired Samples T-test

The paired-samples t-test is a parametric statistical test used to decide whether there is a significant difference between the mean values of the same performance measure for two different algorithms. It calculates a pairwise difference between a performance measure for each of the instances and tests if the average difference is significantly different from zero. More formally, the paired samples t-test tests the null hypothesis that the pairwise differences are a random sample from a normal distribution with mean zero. The alternative hypothesis is that the mean is not zero.

This test assumes that the paired differences are normally distributed. Although the values of a performance measure may not be normally distributed, the pair-wise differences usually are [32].

The Wilcoxon paired samples signed rank test

The Wilcoxon paired samples signed rank test is a non-parametric test. It is used as an alternative to the paired samples t-test when the distribution cannot be assume to be normally distributed. Similar to the paired samples t-test, the Wilcoxon test also compares the pair-wise differences between a performance measure for two different algorithms.

The Wilcoxon test tests the null hypothesis that the pair-wise differences come from a symmetric distribution with zero median. The alternative hypothesis is that the distribution of pairwise differences does not have zero median. The hypothesis that the median of pair-wise differences is zero is not equivalent to the hypothesis that the median of two performance measures are equal.

Results

Table 3.3: The mean recall and precision of the *NB classifier* built using the *same* number of cutoffs as in the literature. The percentage in the parentheses next to each method is the mean percentage of patients that were labelled as high risk in the test sequence.

	Recall	Precision
Method	Mean	Mean
Warped (35%)	72%	4.1%
Literature (33%)	70%	4.2%

Table 3.3 shows the mean performance of the NB classifier on the 100 instances of test sequences as measured by recall and mean precision. Table 3.4 shows the results of the paired samples t-test and the Wilcoxon test.

The *NB classifier* built using warped entropy cutoffs have significantly higher mean recall than the classifiers built from literature cutoffs. The mean increase in recall is 2.4%. The mean precision for warped entropy based *NB classifier* is slightly

Table 3.4: Results from the paired t-test and the Wilcoxon paired samples signed rank test to compare recall and precision of NB classifier built using cutoffs derived from warped entropy measure with literature cutoffs. A positive mean difference indicates higher performance by the classifier using warped entropy cutoffs.

		p-value	
	Mean Difference	t-test	Wilcoxon
Recall	**2.4%**	**0.01**	**0.02**
Precision	-0.1%	0.03	0.07

(0.01%) lower than literature cutoff based classifier. However, the difference in not significant. Therefore, the cutoffs derived using warped entropy can yield *NB classifiers* with a higher recall for an insignificant loss in precision.

3.5 Summary

In this chapter, we reviewed the different types of discretization methods that are suggested in the literature. In addition, we proposed an entropy-based supervised discretization algorithm that uses Bagging to acquire a robust estimation of cutoffs. Next, we discussed three different types of entropy measures, namely Shannon entropy, asymmetric entropy and a novel entropy measure- warped entropy, in the context of discretization of continuous variables.

We evaluated the effectiveness of Bagging in cutoff estimation using all three entropy measures. Our results show that for each entropy measure, Bagging can improve the stability of the cutoffs, as measured by coefficient of variance (COV) of the cutoffs over 100 instances of training sets.

Next, we evaluated cutoffs obtained using different entropy measures in our proposed algorithm. The evaluation was performed based on the stability of the cutoffs generated by each of the measures over several training instances. The cutoffs were generated for four continuous variables: age, morphological variability, heart rate variability and deceleration capacity using data from the MERLIN trial [40]. We compared the COV for all the four continuous variables for each of the three entropy measures. Warped entropy and Shannon had the smallest mean COV of 0.13. In addition, warped entropy measure yielded cutoffs with the best worst case COV in the MERLIN dataset.

We then compared the cutoffs yielded using warped entropy measure to those suggested in the literature, for risk stratification of patients from the MERLIN dataset. We built Naive Bayes classifiers using discretized values of the four continuous variables using 100 instances of training sets. We show that when evaluated on test sets, classifiers built using warped entropy cutoffs have an improved classification performance, 2% increase in recall for an insignificant loss in precision, compared to those obtained from the literature cutoffs.

Chapter 4

An Integrated Approach for Risk Stratification

As described in Chapter 2, the TIMI risk score is a composite score that uses seven binary predictor variables. For each variable, a score of 1 or 0 is assigned based on the value of the predictor variable. The sum of the scores assigned to the variables represents the TIMI risk score. The GRACE [19] risk score assigns weighted scores based on the values of eight predictor variables. The sum of the weighted scores represents the GRACE risk score. The simplicity and the ease of obtaining these scores make them attractive. However, we hypothesize that more complex models have the potential to improve risk stratification. In this thesis, we investigate the application of classification trees to develop a multivariable risk stratification model for post-NSTEACS risk stratification.

4.1 Classification Trees

To illustrate the concept of a classification tree, we present a simple example. Let us consider a classification problem of determining whether a creature is a mammal or not[1] [53]. In order to get to the answer, one approach is to ask a series of questions about different characteristics of the creature. This series of questions can be

[1] The classification tree is not completely correct. It is used only for illustration purposes.

represented in the form of a classification tree shown in Figure 4-1.

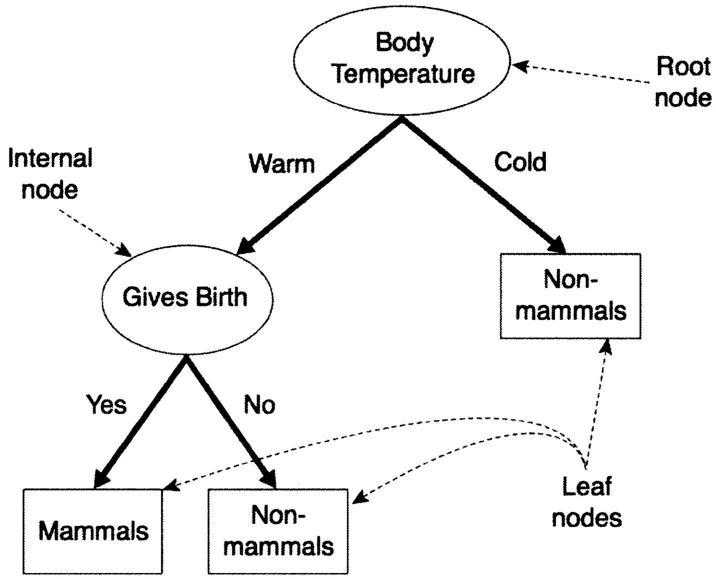

Figure 4-1: An example of a binary classification tree [53].

A classification tree is a hierarchical structure that consists of a set of nodes and edges [53]. There are three types of node in a classification tree (Figure 4-1). A *root* node is the node with no incoming edges and zero or more outgoing edges. The internal nodes have exactly one incoming edge and two or more outgoing edges. If the root node and all the interior nodes have exactly two outgoing edges, then the classification tree is a *binary* classification tree. Figure 4-1 is an example of a binary classification tree. Finally the leaf nodes, or terminal nodes, have exactly one incoming edge and no outgoing edges. Each leaf node is assigned a class label.

Given a classification tree, we can find the classification label for any new test example. We start from the root node, ask the question corresponding to the attribute at the root node and then follow the appropriate branch based on the response for the test example. We continue the process until we reach a leaf node. The new example is then assigned the class label associated with the leaf node.

In the example shown in Figure 4-1, all of the attributes/variables are categorical variables. Continuous variables can also be incorporated in a classification tree, by first discretizing them.

Classification trees provide justifications for classification of new examples. The justifications are comprised of classification rules derived from the tree. A classification rule is a series of question and answer pairs asked about different features of an example, before a conclusion is reached. The utility of justifications for a clinical decision makes classification trees more appealing than 'black-box'-type classification methods such as support vector machines. Classification trees also allow identification of interesting patterns or relationships between variables. This allows researchers to form hypotheses about the physiological mechanisms that could potentially explain the characteristics of classification rules or patterns derived from a classification tree.

We present a formal definition of a classification tree. Given an example, represented by a vector containing values of different variables incorporated in the tree model, a classification tree returns a rule associated with the example and a predicted label for that example.

Classification trees are induced from a training set. A training set consists of a set of examples. Each example is a vector whose components are values of variables that are incorporated in the model. For each example in the training set, we also have class label associated with the example. To develop classification tree models for risk stratification, we use a training set that consists of patients. Each patient is represented by a vector containing values of variables. A variable might represent a value of a test measurement, or indicate the presence or absence of symptoms. We also associate with each patient a class label that corresponds to the presence or absence of a medical outcome.

We introduce some notation to describe the components and properties of a classification tree generated from a training set of patients. Let Φ be an $N \times n$ matrix representing a dataset of N patients in a training set, where each patient is represented by a vector whose components are values of n variables. Let L be vector of length N that contains the class label of each of the N patients.

Each node $\mathfrak{N}$ of the tree contains a subset of patients from Φ. Therefore, a node $\mathfrak{N}$ is represented by a $(\Phi_{\mathfrak{N}}, L_{\mathfrak{N}})$ pair, where $\Phi_{\mathfrak{N}}$ is a $\tilde{N} \times n$ matrix that corresponds to the values of the variables for each patient at the node, and $L_{\mathfrak{N}}$ is a vector of length $\tilde{N}$ that contains the class label for each patient. Here, $\tilde{N} \leq N$.

A classification tree, generated from a training set, has the following properties:

- The child nodes of a node $\mathfrak{N}$ of the tree have mutually exclusive assignments. If a node has k child nodes, we partition the patients at a node $\mathfrak{N}$ into k disjoint set of patients based on the values of the variable at the node. Each child node contains one of the k disjoint sets.

- Each path from the root node of the tree to a leaf node (terminal node) represents a classification rule - a conjunction of variables.

- Each classification rule in the tree is assigned a class label based on the distribution of patients with positive and negative outcomes from the training set that follows the rule. The details of how this is done is described in Section 4.2.

- A classification tree is globally optimal with respect to a dataset if it has the best classification performance on the dataset compared to all the other classification trees that can be constructed using the same data.

4.2 Using Classification Tree for Risk Stratification

We use a binary classification tree to develop a risk stratification model. In Chapter 3, we explore binary classification trees constructed using both global discretization cutoffs and local discretization cutoffs.

During the construction of a classification tree, at each node we select a variable from a set of candidate variables to generate a binary split. The set of candidate variables for a node includes all the binary variables except the variables that have already been used along the path from the root node to the node.

We present the details of the algorithm in the following subsection.

4.2.1 Algorithm for Classification Tree Induction

Searching for a globally optimal tree for a training set is computationally infeasible since the number of classification trees that can be constructed from n variables increases exponentially with n.

We use a greedy algorithm for classification tree induction. The greedy algorithm selects a locally optimal variable for partitioning at each node, where optimality is defined based on a splitting criterion. The choice of variable used to split a node during tree induction depends on the splitting criterion and the discretization cutoffs of the variables. Both of these factors affect the performance of the classification tree.

The Growth Phase

The classification tree generated by the greedy algorithm in the growth phase is called the *maximal* tree. In Algorithms 2 and 3, we present the pseudocode for constructing a maximal tree.

Algorithm 2 (*GetMaximalTree*) Construction of a maximal tree

```
Input: Training data: $\mathbf{\Phi}$, $L$
Output: Maximal tree
Initialize $\mathbf{T}$ to a node $\mathfrak{N}$ with $\Phi_{\mathfrak{N}} = \mathbf{\Phi}, L_{\mathfrak{N}} = L$
Return $GrowTree(\Phi_{\mathfrak{N}}, L_{\mathfrak{N}}, \mathbf{T})$
```

Algorithm 3 (*GrowTree*) Induction of a classification tree

```
Input: Training data: $\Phi_{\mathfrak{N}}$, $L_{\mathfrak{N}}$, Tree $\mathbf{T}$
Output: Tree $\mathbf{T}$
if all examples belong to the same class then
    Return $\mathbf{T}$
else
    $\mathbf{d\Phi} = DiscretizeVariables(\mathbf{\Phi})$ {$\mathbf{d\Phi}$ contains all binarized variables.}
    $OptVariable = GetOptimalVariable(\mathbf{d\Phi})$
    for each value $v_i$ of OptVariable do
        Add a branch to node $\mathfrak{N}$ of $\mathbf{T}$ for $v_i$
        Create a child node $\mathfrak{N}_{c_i}$
        $\mathbf{T}$= $GrowTree(\Phi_{\mathfrak{N}_{c_i}}, L_{\mathfrak{N}_{c_i}}, \mathbf{T})$
    end for
end if
Return $\mathbf{T}$
```

The growth phase consists of two main components: discretization of continuous variables and selection of the 'optimal' variable as defined by a splitting criterion.

- **Discretization of continuous variables**: Since classification trees can only handle categorical variables, we need to discretize continuous variables. As described in Chapter 3, the goal of discretization is to identify cutoff points that partition the range such that the examples that belong to each category exhibit good class coherence. In Chapter 5, we explore classification trees with both global and local discretization.

- **Splitting Criterion based selection of 'optimal' variable**: A greedy classification tree induction algorithm uses the splitting criterion to determine which variable to use to split a particular node during the growth of the tree. The measures used for selecting the best splitting variable at a node are based on some measure of the class coherence of the resulting child nodes. The most commonly used measures include Shannon entropy, Gini index and classification error (Figure 4-2) [53]. For unbalanced datasets, these measures are biased towards the variables that incorrectly assigns many minority examples, i.e. patients who suffer from adverse outcome such as death, to the child node that has a lower risk of adverse outcome.

  Several measures have been proposed as splitting criteria for constructing classification trees from unbalanced data. DKM [13], Hellinger distance (HD) [12] and Class Confidence Proportion (CCP) [30] are examples of such measures. Various studies have shown that classification trees that use one of these measures as the splitting criterion showed improved performance on unbalanced datasets compared to those using Shannon entropy or Gini index [17, 12, 30].

  In this thesis, we investigate the use of non-symmetric entropy measures as splitting criteria for induction of classification trees from unbalanced datasets. We compare their performances with Shannon entropy, DKM, HD and CCP.

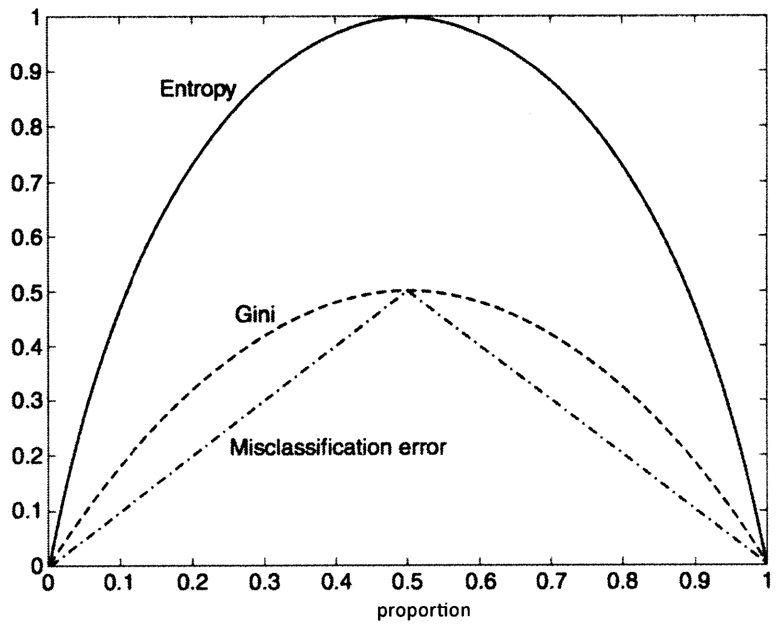

Figure 4-2: Comparison of the class coherence measures for binary classification problems [53]. The x-axis is the proportion of examples that belong to class 1.

The Pruning Phase

Maximal trees are usually large and complex. Large trees are hard to interpret and are also susceptible to overfitting to the training data [44, 34, 30]. Previous studies have shown that pruned maximal trees generalize better on unseen examples and also have shorter classification rules that are easier to interpret [35, 36, 45].

Traditional pruning algorithms are based on error estimations. A node is pruned if the predicted error is decreased after removing the node from the tree [30]. Such pruning methods can have a detrimental effect on constructing decision trees from imbalanced datasets [11] . Therefore, we use Fisher's exact test based pruning approach as proposed by Chawla et al. in [30]. Fisher's exact test (FET) is a statistical significance test that is used to determine if there exists non-random associations between two categorical variables [57]. We present a detailed explanation of FET below. The pruning approach checks if each path (rule) in the tree is statistically significant and if it is not, the path is pruned. Chawla et al. have shown that decision trees pruned using FET based pruning outperformed the trees pruned using an error estimation based method. The classification trees were evaluated using the area under the ROC curve (AUC) [22] analysis.

Fisher's Exact Test

Fisher's Exact test is a statistical significance test that is used to test the null hypothesis that two categorical variables are independent of each other [55].

Let us consider a binary attribute X and a binary class label Y for a training dataset with N examples. The information in the training set can be summarized using a 2×2 contingency table where the number of examples are counted for each value of X and Y (Table 4.1).

Table 4.1: A 2×2 contingency table for two binary categorical variables.

	X=0	X=1	Total
Y=0	a	b	a+b
Y=1	c	d	c+d
Total	a+c	b+d	N

Given a contingency table, the probability of observing the data, under the null hypothesis, is obtained using a hypergeometric distribution [42]:

$$probability = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!N!} \tag{4.1}$$

Let us consider the contingency table shown in Table 4.2. In this example, the two binary categorical variables are 'Treatment' and 'Outcome'. A sample of 19 patients are divided into 'Placebo' and 'Drug' treatment groups and their outcomes are presented in the contingency table. Using Equation 4.1, the probability of this contingency table is 0.01754.

In order to find the statistical significance of the data, FET calculates the total probability of observing contingency tables with an equal or a stronger dependence between the two variables relative to the observed contingency table, under the null hypothesis that the variables are independent. The total probability is the p-value.

Table 4.2: A example contingency table for two categorical variables.

	Treatment	Treatment	Total
	Placebo	Drug	
Outcome= alive	2	7	9
Outcome= dead	8	2	10
Total	10	9	19

In the example shown in Table 4.2, the question of interest is: given the contingency table, is there a statistically significant dependence between ‘Treatment’ and ‘Outcome’. The contingency tables with an equal or stronger dependence between the two variables, along with their probability, are shown in Table 4.3. The p-value, given by the sum of the probabilities, is 0.01852. This is a one-sided test. Given a threshold for the p-value, the dependence between the two variables is considered significant if the p-value if less than the pre-determined threshold.

Table 4.3: Contingency tables with with equal or stronger dependence between the variables ‘Treatment’ and ’Outcome’, relative to Table 4.2. The probability of each contingency table is also shown.

	Treatment	Treatment	Total
	Placebo	Drug	
Outcome= alive	2	7	9
Outcome= dead	8	2	10
Total	10	9	19

probability = 0.01754

	Treatment	Treatment	Total
	Placebo	Drug	
Outcome= alive	1	8	9
Outcome= dead	9	1	10
Total	10	9	19

probability =0.00097

	Treatment	Treatment	Total
	Placebo	Drug	
Outcome= alive	0	9	9
Outcome= dead	10	0	10
Total	10	9	19

probability = 0.00001

When using FET for pruning a classification tree, we calculate the p-value of an internal node. To do so, we form a contingency table based on the binary split generated at the node. Next, based on the contingency table we calculate the p value associated with the node using FET. We use the one-sided FET. A low p value means that the null hypothesis is rejected. Given a threshold for the p-value, we keep the branches that are statistically significant and discard the ones that are not [30].

FET-based Pruning Algorithm

We present the pseudo code for the FET-based pruning in Algorithm 4. Algorithms 5 and 6 are the subroutines of the pruning algorithm. The pseudocode is heavily borrowed from the original publication [30].

Algorithm 4 (*Prune*) Pruning based on FET

```
Input: Maximal tree $\mathfrak{T}$, p-value threshold $pVT$
Output: Pruned decision tree $p\mathfrak{T}$
for Each leaf $Leaf_i$ do
    if $Leaf_i.parent$ is not the Root of $\mathfrak{T}$ then
        $Leaf_i.parent.pruneStatus = Pruneable$, {Default}
        $SetPruneStatus(\mathfrak{T}, Leaf_i.parent, pVT)$,
    end if
end for
Initialize $p\mathfrak{T} = \mathfrak{T}$
for Each $child(i)$ of the root $\mathfrak{T}$ do
    if $child(i)$ is not a leaf then
        if $child(i).pruneStatus == Pruneable$ then
            Set $child(i)$ to be a leaf
        else
            $p\mathfrak{T} = PrunebyStatus(child(i), p\mathfrak{T})$
        end if
    end if
end for
Return $p\mathfrak{T}$
```

After the maximal tree is constructed, the significance of a split at a node in the maximal tree is measured using FET. An internal node becomes a leaf (terminal) node if and only if all its descendant nodes have a p-value greater than the threshold. FET-based pruning as proposed by Chawla et.al. is a two staged pruning algorithm. The first stage is a bottom-up process from the leaf to the root. In this stage we assign a *pruneStatus* to each node of the tree. If a node and all its descendants are non-significant, the *pruneStatus* of the node is set to *Pruneable*, otherwise, it is set as *NotPruneable*. The second stage prunes all insignificant rules from the tree to

Algorithm 5 (*SetPruneStatus*) Subroutine for setting *prune-status* of each internal node by a bottom-up search

```
Input: Maximal Tree $\mathfrak{T}$, an internal node $Node$, p-value threshold $pVT$
for Each $child(i)$ of $Node$ do
    if $child(i).pruneStatus == NotPruneable$ then
        $Node.pruneStatus = NotPruneable$,
    end if
end for
if $Node.pruneStatus == pruneable$ then
    Calculate the $p$ value of this node using Fisher's Exact Test: $Node.pValue$,
    if $Node.pValue < pVT$ then
        $Node.pruneStatus = NotPruneable$,
    end if
end if
if $Node.parent$ is not the root of $\mathfrak{T}$ then
    $Node.parent.pruneStatus = Pruneable$,
    $SetPruneStatus(\mathfrak{T}, Node.parent, pVT)$
end if
```

Algorithm 6 (*PruneByStatus*) Subroutine for pruning nodes according to their pruneStatus

```
Input: A branch represented by its top node $Node$, Tree $\mathbf{T}$
Output: Pruned Tree $\mathbf{T}$
if $Node.pruneStatus == Pruneable$ then
    Set $Node$ as a leaf in $\mathbf{T}$,
else
    for Each $child(i)$ of $Node$ do
        if $child(i)$ is not a leaf then
            $\mathbf{T}$= PruneByStatus($child(i)$,$\mathbf{T}$)
        end if
    end for
end if
```

generate a pruned classification tree.

After pruning the maximal tree, we assign a specific class label to each leaf node. Typically, one assigns a class to a leaf node based on the majority class. However, when a dataset is highly unbalanced, the majority class in the dataset is also the majority class for most of the nodes. To account for the unbalanced distribution, we use a weighted-majority rule. We derive the weights from the distribution of the training set. More specifically, the examples from the majority class in the data set get a weight of 1, while the examples from the minority class get a weight of W where,

$$W = \frac{\text{number of the majority class in the data set}}{\text{number of the minority class in the data set}} \tag{4.2}$$

4.3 Summary

In this chapter, we present the hypothesis that complex models have the potential to improve risk stratification compared to simple techniques such as TIMI and GRACE. We argued that classification trees are more suitable for developing risk stratification models than 'black-box' type models that fail to provide justifications for classification of examples. Next, we presented a classification tree induction algorithm that we use to generate a risk stratification model.

Chapter 5

Splitting criterion for classification trees

During the induction of a classification tree, a splitting criterion is used to select an variable for generating a split at a node. The most commonly used criteria are based on Shannon entropy, the Gini index and classification error [53]. For unbalanced datasets, these measures are biased towards the majority class [30]. Because of this bias, the classification trees generated using these measures may not be appropriate for correct classification of the minority class - the class that is usually of greater interest.

Imbalanced data-oriented splitting criteria have been shown to improve performance on unbalanced datasets compared to those using Shannon entropy or the Gini index [17, 12, 30]. Examples of such criteria include Hellinger distance (HD) [12], DKM [13], and Class Confidence Proportion (CCP) [30].

In this chapter, we compare HD, DKM, CCP, Shannon entropy and two non-symmetric entropy based measures - asymmetric entropy [33] and warped entropy. In Section 3.3, we proposed the concept of warped entropy in the context of discretization. Here, we investigate its performance as a splitting criterion in the construction of classification trees for unbalanced datasets.

In Section 5.2, we discuss different types of splitting criteria used for the induction of classification trees from imbalanced datasets. Section 5.3 describes the evaluation procedure and evaluates the performances of classification trees grown using different splitting criteria.

5.1 Notations

In this chapter, we use a similar notation to that used in Chapter 4.1. We represent a node $\mathfrak{N}$ by a $(\Omega_s, L_{\mathfrak{N}})$ pair. $\Omega_{\mathfrak{N}}$ is a $\tilde{N} \times n$ matrix that corresponds to the values of n variables for each one of the $\tilde{N}$ patients at the node, and $L_{\mathfrak{N}}$ is a vector of length $\tilde{N}$ that contains the class label for each patient.

Let V represent a variable. Since we are using a binary classification tree, a split generated at node $\mathfrak{N}$ using variable V induces two child nodes. $L_{\mathfrak{N},k}$ represents the class labels of patients from node $\mathfrak{N}$ which takes the k^{th} value for variable V.

5.2 Types of splitting criterion

In this section, we discuss six different types of splitting criteria. Each splitting criterion represents a different way of identifying a variable that provides the most useful partitioning of patients, based on their class labels.

Hellinger distance-based criterion measures the distance between two distributions of variable values across different classes. We select the variable that maximizes the distance as the splitting variable at a node.

The splitting criteria based on the other five measures (DKM, Shannon entropy, asymmetric entropy, warped entropy and class confidence proportion) use *information gain* (IG) to select the splitting variable at a node. *Information gain* (IG) is defined as the difference between the class incoherence of a node before a split and the weighted average class incoherence of the child nodes after a split. Each of the five measures represent different ways to quantify class incoherence.

$$IG = \text{Class Incoherence before split} - \text{Class Incoherence after split} \tag{5.1}$$

When using IG for induction of classification trees, the variable that maximizes IG is selected as the splitting variable at a node. At a given node, the class incoherence *before* split is constant for all variables. Hence, maximizing IG is equivalent to minimizing class incoherence *after* a split. We therefore select the variable that minimizes class incoherence *after* a split.

5.2.1 Hellinger distance

The Hellinger distance is a measure of divergence of distributions [24, 47]. When using the Hellinger distance as the splitting criterion, one calculates the divergence between the distributions of variable values across different classes for each of the candidate variables at a node [30]. The larger the Hellinger distance for a given variable, the higher is its propensity to separate the classes [30]. Therefore, during classification tree induction, the variable that maximizes the Hellinger distance is selected as the splitting variable at a node.

Assuming a two-class problem (class 1 and 2), the Hellinger distance for an variable V is given by:

$$\mathbf{HD} = \sqrt{\sum_{k=1}^{2}\left(\sqrt{\frac{|L_{\mathfrak{N},k}=1|}{|L_{\mathfrak{N}}=1|}} - \sqrt{\frac{|L_{\mathfrak{N},k}=2|}{|L_{\mathfrak{N}}=2|}}\right)^{2}} \tag{5.2}$$

5.2.2 DKM

DKM is a measure proposed by Dietterich, Kearns and Mansour [13]. Using DKM, the class incoherence of a node $\mathfrak{N}$ *after* a split is given by Equation 5.3.

$$\text{Class Incoherence } \textit{after} \text{ a split} = \sum_{k=1}^{2}\frac{|L_{\mathfrak{N},k}|}{|L_{\mathfrak{N}}|}.2.\sqrt{p_{\mathfrak{N},k}(1).p_{\mathfrak{N},k}(2)} \tag{5.3}$$

where, $p_{\mathfrak{N}}(1)$ and $p_{\mathfrak{N}}(2)$ are the probability of belonging to class 1 and 2 in node $\mathfrak{N}$, and $p_{\mathfrak{N},k}(l)$ is defined as:

$$p_{\mathfrak{N},k}(l) = \frac{|L_{\mathfrak{N},k}=l|}{|L_{\mathfrak{N},k}|} \tag{5.4}$$

5.2.3 Entropy Measures

In Section 3.3, we presented three different types of entropy as measures of the class incoherence of a (sub)sequence in the context of discretization. The same measures can be used to quantify the class incoherence of a node, in the context of classification trees.

For a given entropy measure H, the class incoherence of a node $\mathfrak{N}$ *after* a split is calculated using Equation 5.5:

$$\text{Class Incoherence } \textit{after} \text{ a split} = \sum_{k=1}^{K} \frac{|L_{\mathfrak{N},k}|}{|L_{\mathfrak{N}}|}.H(L_{\mathfrak{N},k}) \tag{5.5}$$

Based on the entropy measure used, Equations 3.2, 3.3 or 3.4 can be used for *Shannon, asymmetric* and *warped* entropy respectively.

In the context of discretization, asymmetric and warped entropy functions take two inputs: subsequence S and the parent sequence P. In the context of classification tree induction, we input $L_{\mathfrak{N},k}$ and $L_{\mathfrak{N}}$ to compute the class incoherence of the node $\mathfrak{N}$ after a split.

5.2.4 Class Confidence Proportion

Class Confidence Proportion (CCP) is a class incoherence measure proposed by Chawla et al. [30]. Using CCP, the class incoherence of the node *after* a split is given by Equation 5.6.

$$\text{Class Incoherence } \textit{after} \text{ a split} = \sum_{k=1}^{2} \frac{|L_{\mathfrak{N},k}|}{|L_{\mathfrak{N}}|}.\left(\sum_{l=1}^{2} -\tilde{p}_{\mathfrak{N},k}(l).\log_2 \tilde{p}_{\mathfrak{N},k}(l)\right) \tag{5.6}$$

where $\tilde{p}_{\mathfrak{N},k}(l)$ is defined as:

$$\tilde{p}_{\mathfrak{N},k}(l) = \frac{|L_{\mathfrak{N},k} = l|}{|L_{\mathfrak{N}} = l|} \tag{5.7}$$

When constructing CCP-based classification trees, Chawla et al. proposes using

CCP-based information gain in conjunction with Hellinger distance to find the splitting variable at a node. In particular, if two variables have the same CCP-based information gain, they select the variable with a greater Hellinger distance [30].

The equation for CCP-based class incoherence is similar to that based on Shannon entropy. Shannon entropy uses $p_{\mathfrak{N},k}(l)$, the probability that an example with $V = k$ belong to class l, to calculate the class incoherence in each child node. CCP, on the other hand, uses $\tilde{p}_{\mathfrak{N},k}(l)$, the probability that an example that belongs to class l has $V = k$. Chawla et al. argues that this results in a class incoherence measure that is unbiased when the data is unbalanced [30].

5.3 Experiments

We performed all the experiments on data from the MERLIN-TIMI 36 trial (Chapter 3.4). The risk variables that are available for the patients in MERLIN are listed in Table 5.1. Only 9 out of 12 risk variables had a significant (p<0.05) univariate association with the class label of CVD within 90 days, on the MERLIN population. In all the following experiments, we use those 9 variables for the induction of classification trees.

5.3.1 Methodology

In each of the experiments, we use 100 different instances of training and test datasets. Each training dataset contains 2813 patients and its corresponding test dataset contains a disjoint set of 1406 patients. In each of the 100 instances, given a splitting criterion, we use the algorithm described in Chapter 4.2.1 to induce a classification tree based on the training dataset. We evaluate the performance of the classification tree on the corresponding test data, as measured by its recall and precision. If two splitting criteria have the recall or precision that is not significantly (p<0.05) different based on either the paired samples t-test or the Wilcoxon test (Chapter 3.4.4) , we refer to them as having a *comparable* recall or precision. For each splitting criteria, we used the same training and test datasets to construct and evaluate the trees.

Table 5.1: Risk variables available to us for the patients in MERLIN. The variables that are highlighted had a significant ($p < 0.05$) univariate association with CVD within 90 days, in the MERLIN population and are used for the induction of decision tree.

Variable	Description
age	Age in years
gender	Gender (0=male, 1=female)
bmi	Body mass index
diabetes	Diabetes (0=no, 1=yes)
hyptert	Hypertension (0=no, 1= yes)
smoker	History of smoking (0=no, 1=yes)
priormi	Prior incidence of myocardial infarction (0=no, 1= yes)
chf	Congestive heart failure (0=no, 1=yes)
stdep	ST depression>0.55mm (0=no, 1=yes)
MV	Morphological variability
HRV	Heart rate variability LF-HF
DC	Deceleration capacity

5.3.2 Classification Trees with Global Discretization

We evaluated the performance of classification trees generated using splitting criteria based on HD, DKM, CCP and the three entropy measures.

We pre-discretized (global discretization) the continuous variables in both the training and the test sets using the cutoff values from in the literature as shown in Table 3.2. Since DC has two cutoffs, we 'binarize' the variable. Binarization refers to a process in which any numerical variable with more than two distinct values are represented by several binary variables [20]. For example, if a continuous variable V has two cutoffs C_1 and C_2, after binarization we obtain two binary variables V_1 and V_2. We assign a value of 0 to all the examples in the variable V_i with $V < C_i$ and 1 to the others. After the variable V is binarized, we only consider the binarized variables.

For FET-based pruning, we set the p-value threshold to 0.05.

For every pair of splitting criteria, we performed the paired samples t-test and the Wilcoxon test on the recall and the precision values for the 100 test sets. The results are shown in Table 5.3. For each splitting criteria, we have highlighted the rows where its performance (recall or precision) is higher than the other criterion, and

Table 5.2: The mean recall and precision of the classification trees with global discretization, using different splitting criteria. The mean percentage of patients that were labelled as high risk was about 27% for each of the splitting criteria.

	Test Set	
Splitting criterion	Recall	Precision
Hellinger distance	0.63	0.048
DKM	0.63	0.048
Shannon entropy	0.64	0.047
Asymmetric entropy	0.65	0.047
Warped entropy	0.64	0.048
CCP	0.64	0.048

the difference is statistically significant . The differences in the precision for every pair of splitting criteria yielded statistically insignificant results. The differences in the recall, on the other hand, are statistically significant for some pairs of splitting criteria. Shannon entropy-based tree had a higher recall than the trees induced using Hellinger or DKM, with a mean difference of 1.64% and 1.62% respectively. Similarly, warped entropy-based trees and CCP-based trees, also had a higher recall that the trees generated using Hellinger or DKM. Asymmetric entropy-based trees had a higher recall than the trees generated using Hellinger or DKM or CCP. Moreover, our results show that relative to DKM or Hellinger, the mean difference in recall is higher for asymmetric entropy ($\approx$ 2.7%) than for Shannon entropy ($\approx$1.6%), or warped entropy ($\approx$1.8%) or CCP ($\approx$1.4%).

Figure 5.2 shows the mean recall and precision for each of the splitting criterion.

Our results suggest that, for literature cutoffs-based global discretization, classification trees that use asymmetric entropy as the splitting criterion performs the best compared to the other five criteria. However, we note that although the higher recall values yielded by asymmetric entropy are statistically significant, the largest the mean difference is only 2.7%. Therefore, the magnitude of improvement in recall is small.

Table 5.3: The mean difference in recall and precision along with the p-values of paired samples t-test and the Wilcoxon test for classification trees using global discretization. The results show the comparison for every pair of splitting criteria. For example, the first table (**Hellinger vs.**) compares the recall and precision of classification trees obtained using Hellinger distance with all the other splitting criteria. A positive mean difference in that table suggests that Hellinger performed better than the other criterion. The highlighted rows indicated significantly different recall (or precision). For instance, in the third table (**Shannon vs.**), the two highlighted rows show that the recall of Shannon entropy-based trees is significantly higher than those of Hellinger-based trees and DKM-based trees respectively.

Hellinger vs.	Recall			Precision		
	Mean Diff	t-test	Wilcoxon	Mean Diff	t-test	Wilcoxon
	(%)	*p-value*	*p-value*	(%)	*p-value*	*p-value*
DKM	0.10	0.90	1.00	0.00	0.98	1.00
Shannon	-1.64	0.01	0.01	0.11	0.04	0.01
Asym	-2.71	0.00	0.00	0.11	0.08	0.17
Warped	-1.81	0.01	0.02	0.00	0.97	0.79
CCP	-1.39	0.04	0.03	-0.02	0.81	0.56

DKM vs.	Recall			Precision		
	Mean Diff	t-test	Wilcoxon	Mean Diff	t-test	Wilcoxon
	(%)	*p-value*	*p-value*	(%)	*p-value*	*p-value*
Hellinger	0.00	0.90	1.00	0.00	0.98	1.00
Shannon	-1.62	0.01	0.01	0.11	0.04	0.01
Asym	-2.70	< 0.001	< 0.001	0.11	0.08	0.18
Warped	-1.84	0.01	0.02	0.00	0.97	0.81
CCP	-1.38	0.04	0.03	-0.02	0.81	0.56

Shannon vs.	Recall			Precision		
	Mean Diff	t-test	Wilcoxon	Mean Diff	t-test	Wilcoxon
	(%)	*p-value*	*p-value*	(%)	*p-value*	*p-value*
Hellinger	**1.64**	**0.01**	**0.01**	-0.11	0.04	0.01
DKM	**1.62**	**0.01**	**0.01**	-0.11	0.04	0.01
Asym	-1.08	0.01	0.08	-0.01	0.91	0.54
Warped	-0.19	0.78	0.92	-0.12	0.07	0.05
CCP	0.24	0.72	0.74	-0.13	0.06	0.07

Asym. vs.	Recall			Precision		
	Mean Diff	t-test	Wilcoxon	Mean Diff	t-test	Wilcoxon
	(%)	*p-value*	*p-value*	(%)	*p-value*	*p-value*
Hellinger	**2.71**	**< 0.001**	**0.001**	-0.11	0.08	0.17
DKM	**2.70**	**< 0.001**	**< 0.001**	-0.11	0.08	0.18
Shannon	1.08	0.01	0.05	0.01	0.91	0.54
Warped	0.89	0.06	0.01	-0.11	0.00	0.06
CCP	**1.32**	**0.02**	**0.01**	-0.12	0.02	0.08

Warped vs.	Recall			Precision		
	Mean Diff	t-test	Wilcoxon	Mean Diff	t-test	Wilcoxon
	(%)	*p-value*	*p-value*	(%)	*p-value*	*p-value*
Hellinger	**1.81**	**0.01**	**0.02**	0.00	0.97	0.79
DKM	**1.84**	**0.01**	**0.02**	0.00	0.97	0.81
Shannon	0.19	0.78	0.92	0.12	0.07	0.05
Asym	-0.89	0.06	0.01	0.11	0.00	0.06
CCP	0.43	0.43	0.64	-0.01	0.79	0.92

CCP vs.	Recall			Precision		
	Mean Diff	t-test	Wilcoxon	Mean Diff	t-test	Wilcoxon
	(%)	*p-value*	*p-value*	(%)	*p-value*	*p-value*
Hellinger	**1.39**	**0.04**	**0.03**	0.02	0.81	0.59
DKM	**1.38**	**0.04**	**0.03**	0.02	0.83	0.56
Shannon	-0.24	0.72	0.74	0.13	0.06	0.07
Asym	-1.32	0.02	0.01	0.12	0.02	0.08
Warped	-0.43	0.43	0.64	0.01	0.79	0.92

5.3.3 Classification Trees with Local Discretization

Why Local Discretization?

Local discretization preserves context sensitivity since it finds subset-specific cutoffs for discretization. Context sensitivity is important in medicine because the course of action that needs to be taken as a response to a particular clinical test result may be different for patients with different clinical attributes. Similarly, in the context of discretization, it is likely that for a risk metric, the cutoff that best separates the high and low risk group of patients is different for patients with different clinical history. Therefore, in this section we investigate context-sensitive, or subset-specific, cutoff values.

We performed experiments to evaluate whether subset-specific cutoffs found using local discretization can improve classification performance relative to global cutoffs. To do so, we found subset-specific binary cutoffs for each of the continuous variables: age, MV, *HRV* and DC. For example, we found subset-specific cutoffs for MV for patients with and without ST-depression. In this example, patients with ST-depression belong to one subset and patients without ST-depression belongs to another. Besides ST-depression (*stdep*), we also investigated subsets based on the values of the binary variables *hypert*, *smoker*, *priormi* and *chf* (Table 5.1).

The global and the subset-specific cutoffs were derived from each of the 100 instances of training sets. Next, we tested the performance of the univariate classifier, based on each of the continuous variables, on the corresponding test sets.

Given, a continuous risk variable, we compare the recall and precision of each subset-specific cutoffs to the recall and precision of the global cutoff for the same risk variable. Subset-specific cutoffs are considered to have an improved classification performance relative to the global cutoffs if any of the following is satisfied:

1. If the subset-specific cutoffs have a higher recall and precision than those yielded by the global cutoff, and the improvements are statistically significant.

2. If the subset-specific cutoffs have a higher recall, but a comparable precision,

Table 5.4: The mean recall and precision of the binary cutoff for different continuous variables using global and subset-specific cutoffs. The highlighted rows indicate the instances where the subset-specific cutoffs had an improved classification performance relative to the global cutoffs for the same variable.

	Test Set	
Age cutoffs	Recall	Precision
Global	95%	2.3%
Subset Specific		
hypert	92%	2.6%
smoker	**95%**	**2.5%**
priormi	92%	2.7%
chf	85%	2.5%
stdep	87%	2.5%

MV cutoffs	Recall	Precision
Global	83%	2.7%
Subset Specific		
hypert	79%	3.4%
smoker	**93%**	**2.7%**
priormi	**85%**	**3.4%**
chf	77%	3.7%
stdep	**91%**	**3.6%**

HRV cutoffs	Recall	Precision
Global	86%	2.8%
Subset Specific		
hypert	**90%**	**2.9%**
smoker	**93%**	**2.7%**
priormi	**90%**	**2.7%**
chf	82%	2.9%
stdep	**93%**	**2.7%**

DC cutoffs	Recall	Precision
Global	82%	2.6%
Subset Specific		
hypert	80%	2.7%
smoker	**87%**	**2.5%**
priormi	84%	2.4%
chf	85%	2.3%
stdep	**88%**	**3.4%**

relative to that of the global cutoff, and if the improvement in recall is statistically significant.

3. If the subset-specific cutoffs have a higher precision, but a comparable recall, relative to that of the global cutoff, and if the improvement in precision is statistically significant.

   We consider the global cutoffs to have an improved classification performance relative to subset-specific cutoffs using analogous criteria.

Table ?? shows the mean recall and precision values obtained from these experiments. The highlighted rows indicate the instances where the subset-specific cutoffs had an improved classification performance relative to the global cutoffs for the same variable. The global cutoffs, on the other hand, did not yield an improved classification performance relative to any of the subset-specific cutoffs.

Our results demonstrate that identifying subset-specific cutoffs can enhance classification performance. Therefore, in the following subsection, we generate classification trees by including local discretization in our classification tree algorithm. We then evaluate their performance for different splitting criteria.

Evaluation of Classification trees with Local Discretization

In Section 5.3.2, we performed discretization as a preprocessing step before the induction of the decision tree. Now, we incorporate discretization as a part of the decision tree induction algorithm to perform local discretization. At each node of a tree, we locally derive binary cutoffs for continuous candidate variables during the growth phase of the tree. If selected as the best splitting variable at a node, its local cutoff is used to generate a split.

We generated classification trees using local discretization and splitting criteria based on HD, DKM, CCP and the three entropy measures. Table 5.5 shows the mean recall and precision for different splitting criteria obtained from this experiment.

As in Section 5.3.2, we performed the paired samples t-test and the Wilcoxon test on the recall and the precision values for every pair of splitting criteria. The results are shown in Table 5.6. For each splitting criteria, we have highlighted the rows where its performance (recall or precision) is significantly higher than the other criterion. Our results show that asymmetric entropy-based trees have an improved recall and precision compared to trees based on Hellinger, DKM, Shannon and warped entropy. It also has a significantly improved recall compared to CCP, with a mean difference of $\approx 3\%$, for a comparable precision. All the above improvements were statistically significant.

Figure 5.5 shows the mean recall and precision for each of the splitting criterion.

Table 5.5: The mean recall and precision of decision trees induced using local discretization and different types of splitting criteria. The mean percentage of patients that were labelled as high risk was 27% for all the criteria.

	Test Set	
Splitting criterion	Recall	Precision
Hellinger Distance	63%	5.5%
DKM	64%	5.5%
Shannon entropy	61%	5.3%
Asymmetric entropy	74%	6.3%
Warped entropy	60%	5.1%
CCP	71%	6.3%

5.3.4 Comparison of classification trees with global discretization versus local discretization

Here, we compare the best classification performance obtained using global discretization, with the best classification performance obtained using local discretization. According to our results in Section 5.3.2 and Section 5.3.3, splitting criterion based on asymmetric entropy yielded the best classification performance for global as well as local discretization. We repeat the results in Table 5.7.

Asymmetric entropy with local discretization yielded a mean recall of 74% and a mean precision of 6.3%, while asymmetric entropy with global discretization yielded a mean recall of 66% and a mean precision of 4.7%. Both, the mean recall and the mean precision of the classification trees generated using asymmetric entropy with

Table 5.6: The mean difference in recall and precision along with the p-values of paired samples t-test and the Wilcoxon test for classification trees using local discretization. The results show the comparison for every pair of splitting criteria. For example, the first table (**Hellinger vs.**) compares the recall and precision of classification trees obtained using Hellinger distance with all the other splitting criteria. A positive mean difference in that table suggests that Hellinger performed better than that criterion. The highlighted rows indicated significantly different recall (or precision). For instance, in the first table (**Hellinger vs.**), the highlighted row shows that the difference in recall of Hellinger-based trees is significantly higher than of warped entropy-based trees respectively.

Hellinger vs.	Recall			Precision		
	Mean Diff	t-test	Wilcoxon	Mean Diff	t-test	Wilcoxon
	(%)	*p-value*	*p-value*	(%)	*p-value*	*p-value*
DKM	-0.72	0.24	0.39	-0.05	0.46	0.54
Shannon	2.26	0.03	0.07	0.14	0.24	0.10
Asym	-10.89	< 0.001	< 0.001	-0.84	< 0.001	< 0.001
Warped	**4.16**	< **0.001**	< **0.001**	**0.43**	< **0.001**	< **0.001**
CCP	-7.88	< 0.001	< 0.001	-0.79	< 0.001	< 0.001

DKM vs.	Recall			Precision		
	Mean Diff	t-test	Wilcoxon	Mean Diff	t-test	Wilcoxon
	(%)	*p-value*	*p-value*	(%)	*p-value*	*p-value*
Hellinger	0.72	0.24	0.39	0.05	0.46	0.54
Shannon	**2.97**	**0.01**	**0.02**	0.19	0.13	0.07
Asym	-10.17	< 0.001	< 0.001	-0.79	< 0.001	< 0.001
Warped	**4.88**	< **0.001**	< **0.001**	**0.48**	< **0.001**	< **0.001**
CCP	-7.17	< 0.001	< 0.001	-0.74	< 0.001	< 0.001

Shannon vs.	Recall			Precision		
	Mean Diff	t-test	Wilcoxon	Mean Diff	t-test	Wilcoxon
	(%)	*p-value*	*p-value*	(%)	*p-value*	*p-value*
Hellinger	-2.26	0.03	0.07	-0.14	0.24	0.10
DKM	-2.97	0.01	0.02	-0.19	0.13	0.07
Asym	-13.15	< 0.001	< 0.001	-0.98	< 0.001	< 0.001
Warped	1.90	0.09	0.05	0.29	0.01	0.09
CCP	-10.14	< 0.001	< 0.001	-0.93	< 0.001	< 0.001

Asym. vs.	Recall			Precision		
	Mean Diff	t-test	Wilcoxon	Mean Diff	t-test	Wilcoxon
	(%)	*p-value*	*p-value*	(%)	*p-value*	*p-value*
Hellinger	**10.89**	< **0.001**	< **0.001**	**0.84**	< **0.001**	< **0.001**
DKM	**10.17**	< **0.001**	< **0.001**	**0.79**	< **0.001**	< **0.001**
Shannon	**13.15**	< **0.001**	< **0.001**	**0.98**	< **0.001**	< **0.001**
Warped	**15.05**	< **0.001**	< **0.001**	**1.27**	< **0.001**	< **0.001**
CCP	**3.01**	< **0.001**	< **0.001**	0.05	0.64	0.96

Warped vs.	Recall			Precision		
	Mean Diff	t-test	Wilcoxon	Mean Diff	t-test	Wilcoxon
	(%)	*p-value*	*p-value*	(%)	*p-value*	*p-value*
Hellinger	-4.16	< 0.001	< 0.001	-0.43	< 0.001	< 0.001
DKM	-4.88	0.00	0.00	-0.48	0.00	0.00
Shannon	-1.90	0.09	0.05	-0.29	0.01	0.09
Asym	-15.05	< 0.001	< 0.001	-1.27	< 0.001	< 0.001
CCP	-12.04	< 0.001	< 0.001	-1.22	< 0.001	< 0.001

CCP vs.	Recall			Precision		
	Mean Diff	t-test	Wilcoxon	Mean Diff	t-test	Wilcoxon
	(%)	*p-value*	*p-value*	(%)	*p-value*	*p-value*
Hellinger	**7.88**	< **0.001**	< **0.001**	**0.79**	< **0.001**	< **0.001**
DKM	**7.17**	< **0.001**	< **0.001**	**0.74**	< **0.001**	< **0.001**
Shannon	**10.14**	< **0.001**	< 0.001	**0.93**	< **0.001**	< **0.001**
Asym	-3.01	< 0.001	< 0.001	-0.05	0.64	0.96
Warped	**12.04**	< **0.001**	< **0.001**	**1.22**	< **0.001**	< **0.001**

Table 5.7: The best mean recall and precision of classification trees induced using different types of discretization.

	Test Set	
Splitting criterion[*Discretization*]	Recall	Precision
Asymmetric entropy[*local*]	74%	6.3%
Asymmetric entropy[*global*]	66%	4.7%

Table 5.8: Results from the paired t-test and the Wilcoxon paired samples signed rank test to compare recall and precision of the classification trees using asymmetric entropy with *local* discretization versus those of the classification trees using asymmetric entropy with *global* discretization. A positive value implies that trees with local discretization had a higher value.

		p-value	
	Mean Difference	t-test	Wilcoxon
Recall	**9.15%**	**< 0.001**	**< 0.001**
Precision	**1.6%**	**< 0.001**	**< 0.001**

local discretization was higher than those obtained using global discretization. The improvements in the recall and precision were statistically significant. The mean difference in recall was 9.15% and the mean difference in precision was 1.6% (Table 5.8). These results demonstrate the importance of subset-specific cutoffs for classification tree-based risk stratification models for post-ACS patients.

5.3.5 Evaluation of Classification Stability

The classification tree model constructed using our algorithm depends on the training set. It is important, however, that the trees, constructed using small perturbations of the training set, give similar classification labels for a given set of test patients. We refer to this notion as 'classification stability'. In this subsection, we evaluate the classification stability of the trees generated from our proposed algorithm using asymmetric entropy with warped entropy-based local discretization.

Methodology

For this experiment, we divided the MERLIN set into a training set and a holdout set that consisted of 2813 and 1406 patients respectively. Using the training set, we generated 100 bootstrapped training replicates using Bagging, as described in Chapter 3.2. For each of the 100 replicates, we constructed a classification tree model. Next, we used each of the tree models to predict outcomes of patients in the holdout set.

We constructed a 100 × 1406 matrix, D, that contains the classification labels for all the holdout patients using each one of the 100 tree models. In order to quantify, classification emphinstability, we evaluated the Hamming distance between the predicted labels generated for each pair of trees, i.e. between each pair of rows of matrix D. The hamming distance between two rows of matrix D is the number of patients for which the classification labels are different. A total of $\binom{100}{2} = 4950$ comparisons were made.

Fractional hamming distance is defined as:

$$\text{fractional Hamming distance} = \frac{\text{Hamming Distance}}{\text{Size of the holdout set}} \tag{5.8}$$

Figure 5-1 shows the fractional Hamming distance for each of the 4950 comparisons. The graph is a symmetric plot. From the plot, we see some outlier cases, indicated by the red lines in the plot, that have a high Hamming distance. These classification tree models had a much higher recall rate but a lower precision on the holdout set, compared to the mean recall and precision. The outlier trees predicted many of the patients who did not die as high risk patients. Therefore, the Hamming distance for these are much higher than the classification trees with recall and precision comparable to the mean values.

Figure 5-2 shows the histogram of the fractional Hamming distance for every pairwise comparison of the rows of matrix D. Based on 4950 comparisons, on average $\approx$ 16% (231 patients: 227 patients who were alive and 4 patients who died) of the patients in the holdout set were assigned different classification labels by the two

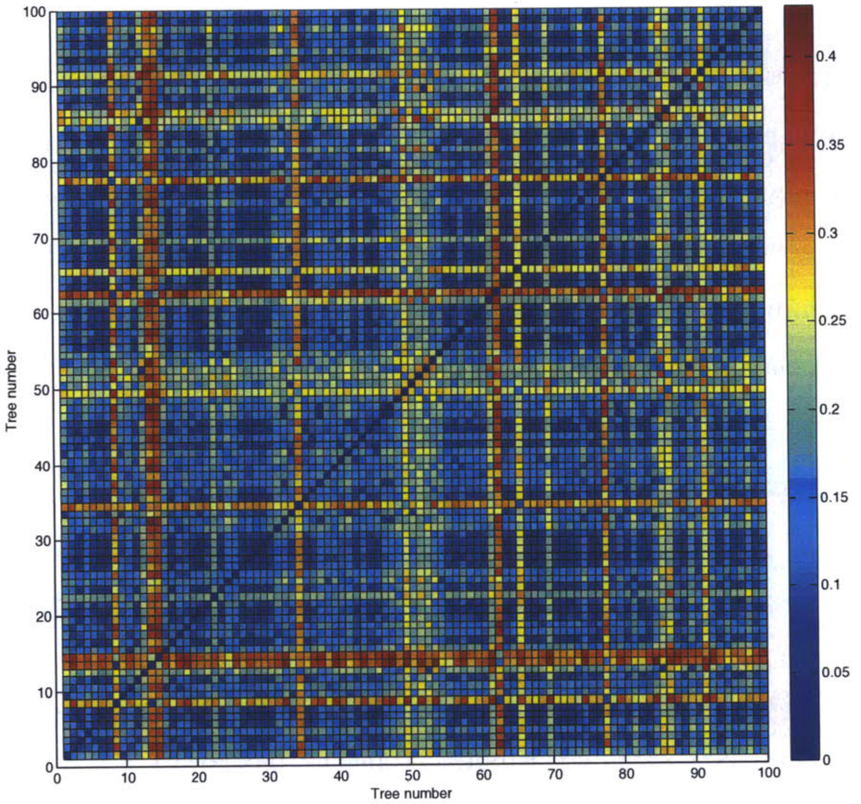

Figure 5-1: Pair-wise fractional Hamming distance between the predicted labels for the holdout patients using the 100 classification tree models generated using the 100 training samples.

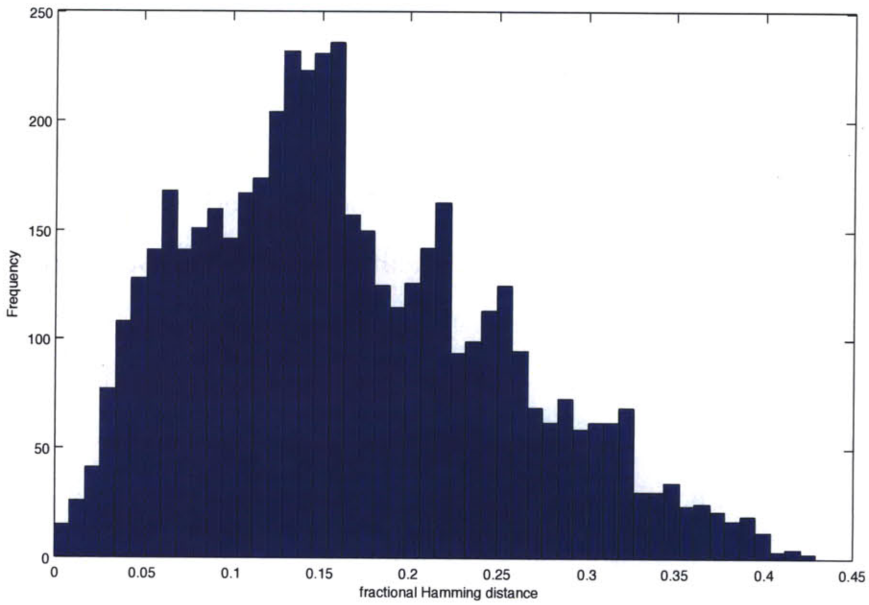

Figure 5-2: Histogram of the fractional Hamming distance for every pairwise comparison of the rows of matrix D.

classification tree models that form a pair. Out of the 27 patients in the holdout set who died, on average only 4 patients were assigned different classification labels by the two models in a pair.

5.4 Summary

In this chapter, we reviewed different types of splitting criteria proposed in the literature for constructing classification trees from unbalanced data. In addition, we also presented non-symmetric entropy measures as splitting criteria for induction of classification trees. We evaluated the performance of non-symmetric entropy-based measures with other splitting criteria that are discussed in the literature. For global discretization of continuous variables, we found that the performance of classification trees constructed using different splitting criteria were comparable for risk stratification of post-ACS patients. The classification performance was measured by recall and precision.

We also showed that subset-specific cutoffs can improve the classification performance for a univariate classifier over using global cutoff derived from the entire patient population with NTSEACS. Therefore, we incorporated local discretization in our classification tree algorithm. We show that for risk stratification of post-ACS patients, classification trees constructed using local discretization coupled with asymmetric entropy yields significantly better results than other splitting criteria proposed in the literature. Our results also show that for post-ACS risk stratification, those classification models outperform models generated using global discretization.

In the next chapter, we compare the performance of classification tree models with the TIMI risk score and multivariate logistic regression models.

Chapter 6

Comparison of different models for post-ACS risk stratification

In this chapter, we compare the performance of asymmetric entropy-based classification trees with local discretization to those of the TIMI risk score and the multivariate linear logistic regression models for risk stratification of post-ACS patients.

6.1 Methodology

For all of the experiments, we use 100 different instances of training and test datasets. Each training dataset contains 2813 patients and its corresponding test dataset contains a disjoint set of 1406 patients. For each of the 100 instances, we use the algorithm described in Section 4.2.1 to induce a classification tree based on the training dataset. We use asymmetric entropy as the splitting criterion and warped entropy based local discretization. We also generate a logistic regression model for each of the 100 instances of the training sets. The logistic regression models were developed using the generalized linear model regression toolbox in Matlab.

Next, we evaluate the performance of the classification trees and logistic regression models on the corresponding test data, as measured by their recall, precision and the odds ratio. For both classification trees and logistic regression models, we used the same training and test datasets to construct and evaluate them. To evaluate the risk stratification based on the TIMI risk score, we simply used the TIMI risk score of the patients in the test set. We considered the test patients with TIMI risk score of 5-7 as the high risk patients ($\approx$ 20% of the MERLIN population) and the others as low risk. Based on the risk labels assigned to the patients and their true outcomes, we computed the recall, precision and odds ratio.

Recall and precision are the same as defined in Section 3.4.4. We present the definition of the odds ratio below.

The Odds Ratio

The odds of an event is the probability of an event occurring divided by the probability an event not occurring. The odds ratio is the ratio of the odds of an event in two separate groups. In case of risk stratification for death, the odds ratio is the ratio of odds of death in the low risk group to the odds of death in the high risk group.

An odds ratio>1 suggests that the odds of death in the high risk group is higher than the odds of death in the low risk group.

Given a performance measure, two types of models have a significantly different value for that measure if both the paired samples t-test and the Wilcoxon test (Chapter 3.4.4) yield p-values<0.05.

For convenience, we'll refer to the classification tree models as 'AsymCT', the TIMI risk score as 'TIMI' and the logistic regression models as 'LRM'.

6.2 Classification Tree Model versus TIMI

As described in Chapter 2.2.1, the TIMI risk score is a simple post-ACS risk stratification tool that incorporates several binary clinical risk variables. Table 6.1 shows the mean recall, precision and odds ratio for TIMI and AsymCT.

AsymCT performs significantly better than the TIMI risk score in terms in recall, precision and the odds ratio. On average, the difference in recall between AsymCT and TIMI was 30.9%. The mean difference in precision was 1.94% and the mean difference in the odds ratio was 9.20.

Table 6.1: The mean recall, precision and odds ratio of AsymCT model and TIMI risk score-based risk stratification model.

	Test Set		
	Recall	Precision	Odds Ratio
TIMI	44%	4.4%	3.44
AsymCT	74%	6.3%	12.7

Table 6.2: Results from the paired t-test and the Wilcoxon paired samples signed rank test to compare recall, precision and odds ratio of AsymCT with those of the TIMI risk score on the test sets. A positive value implies that AsymCT had a higher value than the TIMI score.

		p-value	
	Mean Difference	t-test	Wilcoxon
Recall	**30.9%**	$< \mathbf{0.001}$	$< \mathbf{0.001}$
Precision	**1.94%**	$< \mathbf{0.001}$	$< \mathbf{0.001}$
Odds Ratio	**9.20**	$< \mathbf{0.001}$	$< \mathbf{0.001}$

It is important to note that the some of the risk attributes that are used to construct the classification tree-based model (AsymCT) are not considered when calculating the TIMI risk score and vice versa. In particular, none of the electrocardiographic risk measures such as HRV, DC and MV are incorporated in the TIMI risk score. On the other hand, the TIMI risk score uses information about cardiac markers such as CK-MD or troponin that are not used to construct the tree model. We did not use these markers was because we did not have access to all the data used to derive the TIMI risk score. Therefore, the improvement in recall, precision and the odds ratio obtained from the classification tree-based model could be the result of the additional information provided by the electrocardiographic risk attributes or the method used to create the integrated risk stratification model.

Next, we present compare the classification tree-based model with the model derived using multivariate logistic regression model (LRM) obtained using an identical set of risk attributes.

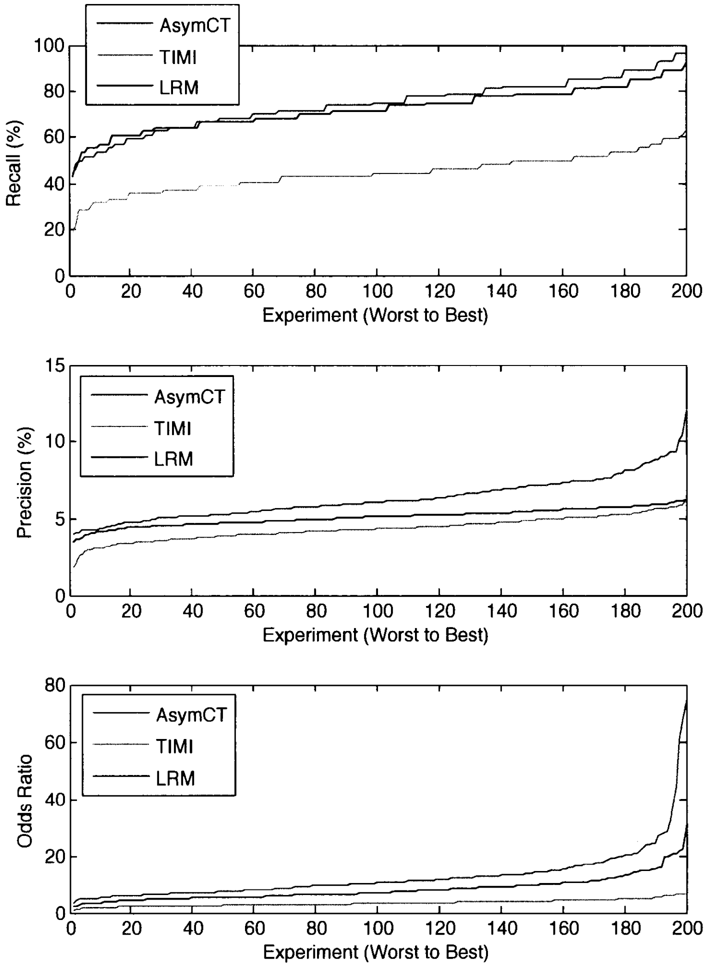

Figure 6-1: Comparison of the distribution of recall, precision and odds ratio for three different classification models. It shows the sorted values of each classification performance measures for different models.

6.3 Classification Tree Model versus Linear Multivariate Logistic Regression Model

Logistic regression is the most commonly used technique in the literature for generating multivariate risk stratification model [3, 5, 10]. Using the same set of risk attributes that were included in the classification tree-based model, AsymCT, we generated a logistic regression model for each of the training sets. To generate a logistic regression model (LRM), all the risk attributes were normalized to fall in the range of 0 to 1.

The logistic regression model generates the probability of death for each of the patients in the test set. To calculate recall, precision and the odds ratio, we used 2% (average death rate in the MERLIN population) as the threshold such that patients with probability of death $> 2\%$ were considered as high risk. Table 6.3 shows the mean recall, precision and odds ratio for LRM and AsymCT. Both the average recall and precision for LRM is lower than those obtained from AsymCT.

Table 6.3: The mean recall, precision and odds ratio of the AsymCT model and the logistic regression model (LRM).

	Test Set		
	Recall	Precision	Odds Ratio
LRM	72%	5.1%	8.3
AsymCT	74%	6.3%	12.7

Table 6.4: Results from the paired t-test and the Wilcoxon paired samples signed rank test to compare recall, precision and odds ratio of AsymCT with those of LRM on the test sets. A positive value implies that AsymCT had a higher value than LRM.

		p-value	
	Mean Difference	t-test	Wilcoxon
Recall	**2.1%**	**0.01**	**0.005**
Precision	**1.18%**	**< 0.001**	**< 0.001**
Odds Ratio	**4.38**	**< 0.001**	**< 0.001**

AsymCT yields significantly higher recall, precision and odds ratio than the logistic regression model. Table 6.4 shows the mean difference and the the corresponding p-values for each.

For LRM, using a threshold value greater than 2% to classify patients as low versus high risk, will result in a lower recall and higher precision relative to those obtained using 2% as the threshold. Conversely, using a threshold value smaller than 2% to classify patients as low versus high risk, will result in a higher recall and lower precision relative to those obtained using 2% as the threshold. Therefore, no matter which threshold we use, our results show that LRM cannot yield a higher recall and a higher precision relative to those of AsymCT.

Our results show that classification tree-based models can generate risk stratification models that yield improved classification performance relative to the logistic regression based models - the standard method of generating multivariate risk stratification models in medicine.

Figure 6-1 shows the distribution of recall, precision and odds ratio for AsymCT, TIMI and LRM models. It compares the sorted values of each classification performance measure obtained using different classification models.

6.4 Summary

Our results also show that classification tree based models can generate risk stratification models with an improved recall, precision and odds ratio than the TIMI risk score and logistic regression based models that are popularly used in the medical literature.

Chapter 7

Understanding the rules of a classification tree

In this chapter, we present a classification tree model generated from the entire MERLIN dataset using warped entropy-based local discretization and the asymmetric entropy-based splitting criterion. We also analyze the properties of the tree and the classification rules derived from it. In addition, we discuss how, given a classification tree model, a doctor should interpret and communicate the rules derived from the model to make a clinical decision.

7.1 Classification tree model for MERLIN

The classification tree model developed using the MERLIN dataset is a risk stratification model for identifying NTSEACS patients at high risk of cardiovascular mortality. We used the nine variables highlighted in Table 5.1 to generate the model. There are other risk attributes, such as left ventricular ejection fraction and the level of cardiac biomarkers, that are known to provide useful prognostic information for risk stratification of NTSEACS patients. Since we did not have access to such information, those attributes were not incorporated in our model. To develop a comprehensive classification tree-based risk stratification model for NTSEACS patients, one should include those risk attributes.

Figure 7-1 shows the classification tree based risk stratification model generated from the dataset. The cutoffs for the continuous variables were rounded to the lowest precision, such that the set of patients with variable values on either side the cutoff remain unchanged before and after the rounding.

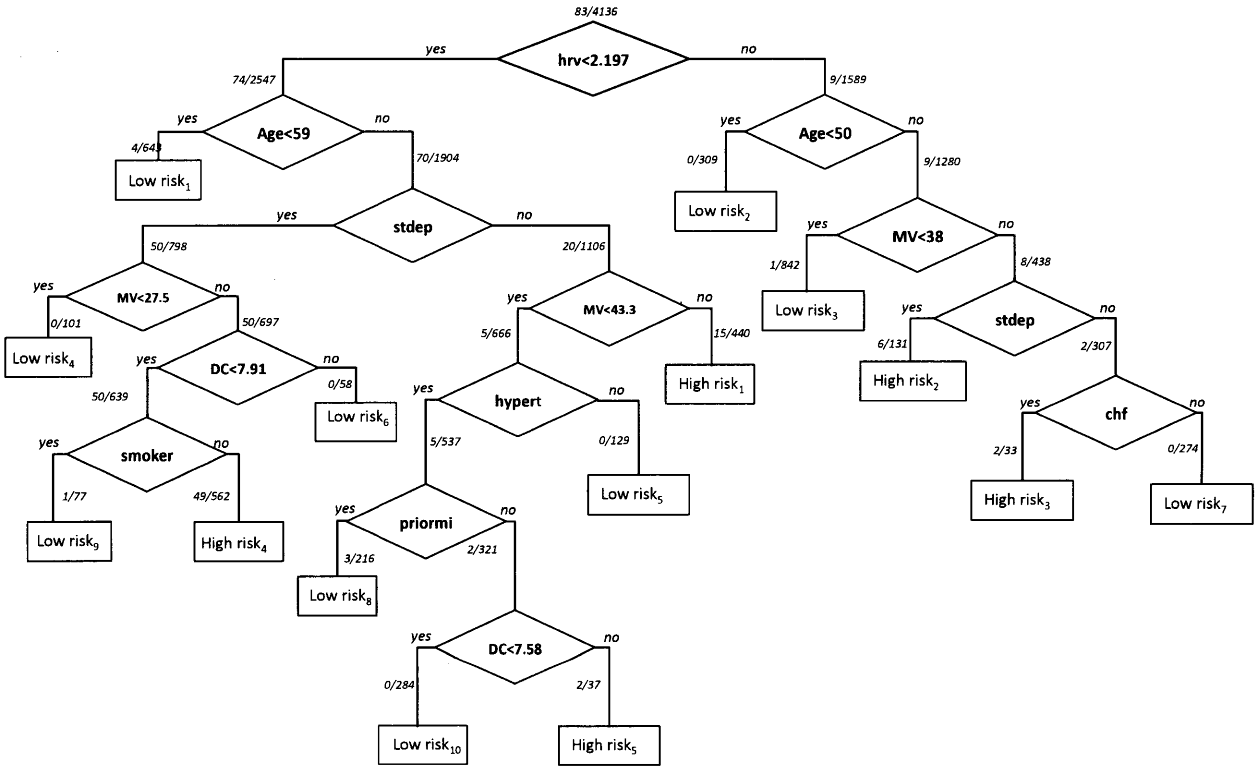

Figure 7-1: The classification tree based risk stratification model generated from the MERLIN dataset. We used our algorithm with warped entropy-based local discretization and asymmetric entropy-based splitting criterion to construct the tree. The numbers above each node show the number of dead patients and the number of alive patients at each node i.e. *number of dead patients/number of alive patients.* The superscripted number next to the 'Low risk' or 'High risk' label is the rule number. For example, 'Low risk$_1$' means *the first low risk rule.*

7.2 Model Analysis and Discussion

7.2.1 Analysis of Classification Rules

Each path starting from the root node to a terminal node represents a rule in a classification tree. The classification tree model derived from the MERLIN consists of ten low risk rules and five high risk rules as shown in Figure 7-1.

Different risk attributes contain information about different stages of medical state of a patient. For example, the risk attribute *priormi* describes the past medical condition of a patient. On the other hand, the risk attribute *stdep* tells us whether or not a patient had ST-depression at admission.

We categorize the nine variables included in our classification model into three categories:

- Category A: *Untreatable* (e.g. *age*)
- Category B: *Likely not yet treated* (e.g. *stdep, HRV, MV, DC*)
- Category C: *Likely to have been treated before* (for e.g. *priormi, chf, hypertension, smoker*)

Category A and B describe the current state of the patient. Category C describes the past medical condition of a patient that might have been treated. If the variable that contains information about the treatment is not included in the model, it could potentially be a confounding factor when analyzing the classification rules that includes attributes from Category C. It is important for a doctor to interpret such rules with caution.

Next, we will discuss the classification rules derived from the classification tree model in Figure 7-1.

1. Let us first analyze Low risk$_1$. This rule suggests that despite having a low hrv (usually associated with elevated risk), patients with NSTEACS who are less than 59 years of age are at a low risk of CVD. High risk$_2$, on the other hand, suggests that a patients with high hrv (considered to be at lower risk based on

hrv alone) are still at high risk of cardiovascular death if the patient is older than 50, have MV>38 and also have ST-depression. These rules illustrate the advantage of integrated models in improving risk stratification. All variables that are included in Low risk_1 or High risk_2 belongs to Category A or Category C. Therefore, these rules are likely to be free of confounding factors.

2. Next, we consider rules Low risk_4 and High risk_1. Both of these rules include *hrv*<2.2 and *age*>=59. Patients that satisfy Low risk_4 have ST depression and those who satisfy High risk_1 don't. The MV cutoffs that separates patients at lower risk from those at higher risk is 27.5 and 43.3 for Low risk_4 and High risk_1, respectively. This suggests that a patient with ST-depression has to have a very low MV (< 27.5) to be considered at low risk. For patients without ST-depression, a patient must have MV of at least 43.3 to be considered a high risk patient. The different MV cutoffs are a consequence of local discretization in our algorithm.

3. Another notable rule in the risk stratification model is Low risk_8. This rule includes the risk attribute *priormi* (history of prior MI), a Category C attribute. Patients who have a history of MI are likely to be on medication such as beta-blockers. At first glance, Low risk_8 appears to suggest that a history of prior MI is protective. However, this unexpected rule is most probably because the information about past medications of the patients is not included in the model. This causes *medication history* to potentially be a confounding factor while interpreting Low risk_8.

7.2.2 Using the classification tree model in a medical setting

In this subsection, we will discuss how a doctor could use a classification tree-based model in a medical setting.

Consider a scenario in which a doctor consults the classification model in Figure 7-1 to identify the risk profile of a new patient. We will assume that information about all the risk attributes, included in the model, is available. Based on the rule that the patient satisfies, the doctor can infer why a patient should be considered to be at high or low risk of CVD. In addition, the doctor may also use these rules to guide the therapy for the patient. For example, a patient who satisfies High risk$_2$ has high MV and ST-depression. ST depression is a sign of myocardial ischemia [41]. High MV is also believed to be associated with the myocardial instability caused by ischemia. Based on this information, a doctor might prescribe a drug that is aimed to treat ischemia.

The classification model also contains information about the number of patients at each node of the tree. A doctor should consider a rule satisfied by a larger population to be more reliable than the one satisfied by a smaller population. For example, High risk$_1$ can be considered more reliable than High risk$_5$.

7.3 Summary

In this chapter, we present the classification tree model generated using our proposed algorithm using the entire MERLIN dataset.

We also analyzed the rules represented in the classification tree model. The classification rules derived from the tree illustrate the power of integrating information from multiple variables in improving risk stratification. We highlighted that confounding attributes may lead to rules that may appear counter-intuitive.

Finally, we discussed how a doctor might use a classification tree model in a medical setting to help guide clinical decisions.

Chapter 8

Conclusions and Future Work

We conclude the thesis with a summary of its goals, major contributions of our work and discussion of proposed improvements for future work.

8.1 Summary

Advances in medical technology have greatly increased the amount of information that can be used to aid clinical decision making. Electronic health records (EHR) make this vast amount of data easily accessible. This offers an opportunity to use the data to create integrated models that can improve risk stratification.

8.1.1 Goals and Contributions

The goal of this thesis was to develop an automated, data-driven algorithm to create an integrated risk stratification model that can potentially be used in a clinical setting. We focused on development of classification tree based models. Classification trees provide justification for classification of examples, which makes these models appealing for use in a clinical setting.

We developed a classification tree induction algorithm in two main stages. In the first stage, we investigated different discretization methods. Discretization is important because classification trees can only handle discrete variables. In our context, the purpose of discretization is to identify cutoffs that partition the range of continuous attribute into intervals such that members of each category exhibit class coherence. In this thesis, we investigated different non-symmetric entropy based methods to identify discretization cut points for continuous attributes in unbalanced data.

In the second stage, we proposed and evaluated classification tree algorithms that use different splitting criteria. For each splitting criterion, we also explored classification tree induction using global and local discretization cutoffs. For local discretization, we implemented the warped entropy-based discretization developed in the first stage.

The major contributions of our work are:

- **Concept of Warped entropy**: In Chapter 3, we proposed a novel non-symmetric entropy measure called warped entropy. This measure allows the maximum entropy point to be set to a distribution other than the one where all the classes are present in equal proportions. We presented this measure in the context of discretization and classification trees for unbalanced data.

- **Entropy based discretization using Bagging**: In Chapter 3.2, we propose an entropy-based supervised discretization algorithm that incorporates Bagging to avoid overfitting of cutoffs to the training set. Our results demonstrated that Bagging can improve the stability of cutoffs derived using different entropy-based methods on unbalanced datasets. The stability of the cutoffs was measured using the coefficient of variance (COV) of the cutoffs over 100 instances of training sets.

- **Evaluation of symmetric and non-symmetric entropy measures for discretization of continuous attributes in unbalanced datasets**: In Chapter 3.4, we evaluated Shannon entropy and two non-symmetric measures: warped entropy and asymmetric entropy [33] for discretization of continuous variables in unbalanced datasets. The cutoffs were generated for four continuous variables: age, morphological variability (MV), heart rate variability (HRV) and deceleration capacity (DC), using data from the MERLIN trial [40].

Among the three entropy measures, we showed that the warped entropy measure yielded cutoffs with the smallest average COV and the best worst case COV in the MERLIN dataset.

- **A novel approach of building classification trees from highly unbalanced dataset**: We presented an algorithm for development of classification trees from unbalanced data. Our algorithm performs binary discretization (*with* Bagging) of continuous variables at each node during the tree induction process. Given a class coherence measure, the algorithm selects the variable that has maximizes class coherence (or minimizes class incoherence) after a split.

- **Evaluation of different class coherence measures for induction of classification tree models with local discretization:** We evaluated our algorithm using different class coherence measure-based splitting criteria. Our results showed that using asymmetric entropy to generate the classification tree yields models with significantly higher recall and precision than other splitting criterion proposed in the literature. We also compared the performance of these classification models with those generated using global discretization. Our results showed that the recall and precision obtained using asymmetric entropy with local discretization is significantly higher than the best recall and precision obtained using global discretization.

- **Comparison of classification tree based models with other models:** We compared the risk stratification performance of the classification tree based models developed using non-symmetric entropy measure with that of the TIMI risk score. Our results show that the classification tree models yield a significantly higher recall, precision and odds ratio than the TIMI risk score. However, the TIMI risk score and the classification tree models uses different set of risk attributes with some overlap. We also developed multivariable logistic regression models (LRM) using the same variables used in the classification tree model. Again, the performance of the classification tree based models were significantly higher than LRMs.

8.2 Future Work

- **Further evaluation of proposed algorithms**: We speculate that our algorithms can be applied successfully on other medical and non-medical datasets. In the near future, we will perform a more rigorous evaluation of our discretization and classification tree algorithm in multiple datasets.

- **Development of stopping criterion for discretization**: At present, our entropy based supervised discretization algorithm takes the number of cutoffs as an input. As future work, we plan to develop an appropriate stopping criteria that can be integrated with our proposed algorithm so that it can automatically determine the appropriate number of cutoffs based on the characteristics of the data.

- **Handling missing values in classification trees**: In all of our analysis of classification trees, we have only used patients with no missing data. However, most of the real world datasets have missing data. In the future, we want to incorporate data imputation in our algorithm so that it handle datasets with missing values.

Bibliography

[1] Heart rate variability: standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology. *Circulation*, 93(5):1043–1065, March 1996.

[2] American heart association. heart disease and stroke statistics- 2009 update. manuscript, Dallas, Texas, 2009.

[3] Gareth Ambler, Rumana Z. Omar, Patrick Royston, Robin Kinsman, Bruce E. Keogh, and Kenneth M. Taylor. Generic, Simple Risk Stratification Model for Heart Valve Surgery. *Circulation*, 112(2):224–231, 2005.

[4] Elliott M. Antman, Marc Cohen, Peter J. L. M. Bernink, Carolyn H. McCabe, Thomas Horacek, Gary Papuchis, Branco Mautner, Ramon Corbalan, David Radley, and Eugene Braunwald. The TIMI Risk Score for Unstable Angina/Non-ST Elevation MI: A Method for Prognostication and Therapeutic Decision Making. *JAMA*, 284(7):835–842, 2000.

[5] Steven I. Aronin, Peter Peduzzi, and Vincent J. Quagliarello. Community-Acquired Bacterial Meningitis: Risk Stratification for Adverse Clinical Outcome and Effect of Antibiotic Timing. *Annals of Internal Medicine*, 129(11 Part 1):862–869, 1998.

[6] Axel Bauer, Jan W Kantelhardt, Petra Barthel, Raphael Schneider, Timo Mäkikallio, Kurt Ulm, Katerina Hnatkova, Albert Schömig, Heikki Huikuri, Armin Bunde, Marek Malik, and Georg Schmidt. Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study. *Lancet*, 367(9523):1674–81, May 2006.

[7] Eric Bauer and Ron Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. In *Machine Learning*, pages 105–139, 1998.

[8] Shlomo A. Ben-Haim, Bruno Becker, Yeouda Edoute, Mira Kochanovski, Orly Azaria, Elieser Kaplinsky, and Yoram Palti. Beat-to-beat electrocardiographic morphology variation in healed myocardial infarction. *The American Journal of Cardiology*, 68(8):725 - 728, 1991.

[9] GE Billman, PJ Schwartz, and HL Stone. Baroreceptor reflex control of heart rate: a predictor of sudden cardiac death. *Circulation*, 66(4):874–880, 1982.

[10] James E. Calvin, Lloyd W. Klein, Betsy J. VandenBerg, Peter Meyer, Joseph V. Condon, R. Jeffrey Snell, Luz Maria Ramirez-Morgen, and Joseph E. Parrillo. Risk Stratification in Unstable Angina: Prospective Validation of the Braunwald Classification. *JAMA*, 273(2):136–141, 1995.

[11] Nitesh V. Chawla. C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In *In Proceedings of the ICML03 Workshop on Class Imbalances*, 2003.

[12] David A. Cieslak and Nitesh V. Chawla. Learning decision trees for unbalanced data. In *In European Conference on Machine Learning*, 2008.

[13] Tom Dietterich, Michael Kearns, and Yishay Mansour. Applying the weak learning framework to understand and improve c4.5. In *In Proceedings of the Thirteenth International Conference on Machine Learning*, pages 96–104. Morgan Kaufmann, 1996.

[14] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. pages 194–202. Morgan Kaufmann, 1995.

[15] N. El-Sherif, B. J. Scherlag, R. Lazzara, and R. R. Hope. Re-entrant ventricular arrhythmias in the late myocardial infarction period. 1. conduction characteristics in the infarction zone. *Circulation*, 55(5):686–702, May 1977.

[16] Barnaby J. Feder. Defibrillators are lifesaver, but risks give pause. New York Times, 2008.

[17] Peter A. Flach. The geometry of roc space: understanding machine learning metrics through roc isometrics. In *in Proceedings of the Twentieth International Conference on Machine Learning*, pages 194–201. AAAI Press, 2003.

[18] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet : Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23):e215–220, 2000.

[19] Christopher B. Granger, Robert J. Goldberg, Omar Dabbous, Karen S. Pieper, Kim A. Eagle, Christopher P. Cannon, Frans Van de Werf, Alvaro Avezum, Shaun G. Goodman, Marcus D. Flather, and Keith A. A. Fox. Predictors of Hospital Mortality in the Global Registry of Acute Coronary Events. *Arch Intern Med*, 163(19):2345–2353, 2003.

[20] Peter Hammer and Tibrius Bonates. Logical analysis of dataan overview: From combinatorial optimization to medical applications. *Annals of Operations Research*, 148:203–225, 2006.

[21] Healthcare Information and Management Systems Society. Electronic health record, November 2010.

[22] Swets J.A. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics.* Addison Wesley, 1996.

[23] ME Josephson and AL Wit. Fractionated electrical activity and continuous electrical activity: fact or artifact? *Circulation*, 70(4):529–532, 1984.

[24] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *Communication Technology, IEEE Transactions on*, 15(1):52 –60, 1967.

[25] Robert E. Kleiger, J.Philip Miller, J.Thomas Bigger Jr., and Arthur J. Moss. Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. *The American Journal of Cardiology*, 59(4):256 – 262, 1987.

[26] Sotiris Kotsiantis and Dimitris Kanellopoulos. Discretization techniques: A recent survey, 2006.

[27] E. Kreyszig. *Introductory Mathematical Statistics.* John Wiley, 1970.

[28] Lukasz A. Kurgan and Krzysztof J. Cios. Caim discretization algorithm. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 16(2):145–153, 2004.

[29] D J Kurz, A Bernstein, K Hunt, D Radovanovic, P Erne, Z Siudak, and O Bertel. Simple point-of-care risk stratification in acute coronary syndromes: the AMIS model. *Heart*, 95(8):662–668, 2009.

[30] Wei Liu, Sanjay Chawla, D.A. Cieslak, and N.V. Chawla. A robust decision tree algorithms for imbalanced data sets. In *Proceedings of the Tenth SIAM International Conference on Data Mining*, pages 766–777. Society for Industrial and Applied Mathematics, 2010.

[31] N. R. Lomb. Least-squares frequency analysis if unequally spaced data. *Astrophysics and Space Science.*

[32] Grover H Visweswaran S. Lustgarten J.L, Gopalakrishnan V. Improving classification performance with discretization on biomedical datasets. In *Proceedings of American Medical Informatics Association*, 2008.

[33] Simon Marcellin, Djamel-Abdelkader Zighed, and Gilbert Ritschard. Detection of breast cancer using an asymmetric entropy measure. In *Computational Statistics*, volume 25, pages 975–982. Springer, 2006.

[34] Manish Mehta, Jorma Rissanen, and Rakesh Agrawal. Mdl-based decision tree pruning. pages 216–221. AAAI Press, 1995.

[35] John Mingers. An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, 4:227–243, 1989. 10.1023/A:1022604100933.

[36] John Mingers. An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, 3:319–342, 1989. 10.1007/BF00116837.

[37] T. M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[38] David A. Morrow, Elliott M. Antman, Andrew Charlesworth, Richard Cairns, Sabina A. Murphy, James A. de Lemos, Robert P. Giugliano, Carolyn H. McCabe, and Eugene Braunwald. TIMI Risk Score for ST-Elevation Myocardial Infarction: A Convenient, Bedside, Clinical Score for Risk Assessment at Presentation : An Intravenous nPA for Treatment of Infarcting Myocardium Early II Trial Substudy. *Circulation*, 102(17):2031–2037, 2000.

[39] David A. Morrow, Elliott M. Antman, Lori Parsons, James A. de Lemos, Christopher P. Cannon, Robert P. Giugliano, Carolyn H. McCabe, Hal V. Barron, and Eugene Braunwald. Application of the TIMI Risk Score for ST-Elevation MI in the National Registry of Myocardial Infarction 3. *JAMA*, 286(11):1356–1359, 2001.

[40] David A. Morrow, Benjamin M. Scirica, Ewa Karwatowska-Prokopczuk, Sabina A. Murphy, Andrzej Budaj, Sergei Varshavsky, Andrew A. Wolff, Allan Skene, Carolyn H. McCabe, Eugene Braunwald, and For the MERLIN-TIMI 36 Trial Investigators. Effects of Ranolazine on Recurrent Cardiovascular Events in Patients With Non-ST-Elevation Acute Coronary Syndromes: The MERLIN-TIMI 36 Randomized Trial. *JAMA*, 297(16):1775–1783, 2007.

[41] Kors JA van Herpen G Crow RS Fabsitz RR Howard BV. Okin PM, Devereux RB. Computerized st depression analysis improves prediction of all-cause and cardiovascular mortality: the strong heart study. *Annals of noninvasive electrocardiology : the official journal of the International Society for Holter and Noninvasive Electrocardiology, Inc*, pages 107–116, 2001.

[42] K.J. Preacher and N.E. Briggs. Calculation of fisher's exact test: An interactive calculation tool for fisher's exact probability for 2×2 contingency tables., May 2001.

[43] Fu-Ming Qiu, Jie-Kai Yu, Yi-Ding Chen, Qi-Feng Jin, Mei-Hua Sui, and Jian Huang. Mining novel biomarkers for prognosis of gastric cancer with serum proteomics. *Journal of Experimental and Clinical Cancer Research*, 28(1):126, 2009.

[44] J. R. Quinlan. Simplifying decision trees, 1986.

[45] J. R. QUINLAN. Simplifying decision trees. *International Journal of Human-Computer Studies*, 51(2):497 – 510, 1999.

[46] Taimur Qureshi and Djamel-Abdelkader Zighed. Using resampling techniques for better quality discretization. In *6th International Conference on Machine Learning and Data Mining (MCDM'09), Leipzig, Germany*, pages 68–81, July 2009.

[47] C. Radhakrishna Rao. A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. 1995.

[48] PJ Schwartz, E Vanoli, M Stramba-Badiale, GM De Ferrari, GE Billman, and RD Foreman. Autonomic mechanisms and sudden death. New insights from analysis of baroreceptor reflexes in conscious dogs with and without a myocardial infarction. *Circulation*, 78(4):969–979, 1988.

[49] Philip Pohong Sung. Risk stratification by analysis of electrocardiographic morphology following acute coronary syndrome. Master of engineering, Massachusetts Institute of Technology, 2009.

[50] Zeeshan Syed and Ilan Rubinfield. Personalized risk stratification for adverse surgical outcomes. *American Medical Association (AMA)-IEEE Medical Technology Conference on Individualized Healthcare.*, march 2010.

[51] Zeeshan Syed, Benjamin M Scirica, Satishkumar Mohanavelu, Phil Sung, Eric L Michelson, Christopher P Cannon, Peter H Stone, Collin M Stultz, and John V Guttag. Relation of death within 90 days of non-st-elevation acute coronary syndromes to variability in electrocardiographic morphology. *Am J Cardiol*, 103(3):307–11, Feb 2009.

[52] Zeeshan Hassan Syed. Computational methods for physiological data. Doctor of philosophy, Massachusetts Institute of Technology, 2009.

[53] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison Wesley, u.s. edition edition, May 2005.

[54] H Tsuji, Jr Venditti, FJ, ES Manders, JC Evans, MG Larson, CL Feldman, and D Levy. Reduced heart rate variability and mortality risk in an elderly cohort. The Framingham Heart Study. *Circulation*, 90(2):878–883, 1994.

[55] Florian Verhein and Sanjay Chawla. Using significant, positively associated and relatively class correlated rules for associative classification of imbalanced datasets. *Data Mining, IEEE International Conference on*, 0:679–684, 2007.

[56] Lawrence Weinstein, Todd Radano, Timothy Jack, Philip Kalina, and John S. Eberhardt. Application of Multivariate Probabilistic (Bayesian) Networks to Substance Use Disorder Risk Stratification and Cost Estimation. *Perspectives in Health Information Management*, 2009.

[57] Eric W. Weisstein. Fisher's exact test. From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/FishersExactTest.html.

[58] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.