RISK ESTIMATION BY MAXIMIZING AREA UNDER RECEIVER OPERATING CHARACTERISTICS CURVE WITH APPLICATION TO CARDIOVASCULAR SURGERY

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BİLKENT UNIVERSITY
IN PARTIAL FULLFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Murat Kurtcephe
July 2010

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. H. Altay Güvenir (Supervisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. A. Rüçhan Akar

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Çiğdem Gündüz Demir

Approved for the Institute of Engineering and Sciences:

Prof. Dr. Levent Onural
Director of Institute of Engineering and Sciences

ABSTRACT

RISK ESTIMATION BY MAXIMIZING AREA UNDER RECEIVER OPERATING CHARACTERISTICS CURVE WITH APPLICATION TO CARDIOVASCULAR SURGERY

Murat Kurtcephe
M.S. in Computer Engineering
Supervisor**:** Prof. Dr. H. Altay Güvenir

June 2010

Risks exist in many different domains; medical diagnoses, financial markets, fraud detection and insurance policies are some examples. Various risk measures and risk estimation systems have hitherto been proposed and this thesis suggests a new risk estimation method. Risk estimation by maximizing the area under a Receiver Operating Characteristics (ROC) curve (REMARC) defines risk estimation as a ranking problem. Since the area under ROC curve (AUC) is related to measuring the quality of ranking, REMARC aims to maximize the AUC value on a single feature basis to obtain the best ranking possible on each feature. For a given categorical feature, we prove a sufficient condition that any function must satisfy to achieve the maximum AUC. Continuous features are also discretized by a method that uses AUC as a metric. Then, a heuristic is used to extend this maximization to all features of a dataset. REMARC can handle missing data, binary classes and continuous and nominal feature values. The REMARC method does not only estimate a single risk value, but also analyzes each feature and provides valuable information to domain experts for decision making. The performance of REMARC is evaluated with many datasets in the UCI repository by using different state-of-the-art algorithms such as Support Vector Machines, naïve Bayes, decision trees and

boosting methods. Evaluations of the AUC metric show REMARC achieves predictive performance significantly better compared with other machine learning classification methods and is also faster than most of them.

In order to develop new risk estimation framework by using the REMARC method cardiovascular surgery domain is selected. The TurkoSCORE project is used to collect data for training phase of the REMARC algorithm. The predictive performance of REMARC is compared with one of the most popular cardiovascular surgical risk evaluation method, called EuroSCORE. EuroSCORE is evaluated on Turkish patients and it is shown that EuroSCORE model is insufficient for Turkish population. Then, the predictive performances of EuroSCORE and TurkoSCORE that uses REMARC for prediction are compared. Empirical evaluations show that REMARC achieves better prediction than EuroSCORE on Turkish patient population.

Keywords: Risk Estimation, AUC Maximization, AUC, Ranking, Cardiovascular Operation Risk Evaluation

ÖZET

RECEIVER OPERATING CHARACTERTICS EĞRİSİ ALTINDAKİ ALANI MAKSİMİZE EDEREK RİSK TAHMİNİ VE KARDİYOVASKÜLER CERRAHİ UYGULAMASI

Murat Kurtcephe
Bilgisayar Mühendisliği Bölümü Yüksek Lisans
Tez Yöneticisi: Prof. Dr. H. Altay Güvenir
Temmuz 2010

Risk birçok farklı alanda mevcuttur; tıbbi tanı, finansal piyasalar, dolandırıcılık tespiti ve sigorta poliçeleri bunların birkaçıdır. Çeşitli risk ölçütleri ve risk tahmin sistemleri bugüne kadar önerildi ve bu tez yeni bir risk tahmini yöntemi sunmaktadır. Receiver Operating Characteristics (ROC) eğrisi altındaki alanı maksimize ederek risk tahmin yöntemi (REMARC), risk tahmini bir sıralama sorunu olarak tanımlar. ROC eğrisi altındaki alan (AUC) değeri sıralama kalitesini ölçme ile ilgili olduğundan, REMARC tek bir öznitelik üzerinde en yüksek AUC'yi elde ederek her öznitelik üzerinde mümkün olabilecek en iyi sıralamayı sağlamayı hedeflemetedir. Verilen bir kategorik öznitelik için, herhangi bir risk yordamının en yüksek AUC'yi elde etmek için sağlaması gereken şartın ne olduğunu ispatladık. Sayısal öznitelikler de ölçüt olarak AUC'yi kullanan bir yöntemle ayrıklaştırılmıştır. Sonra, sezgisel bir yaklaşımla AUC'nin maksimize eldilmesi tüm veriseti üzerine genişletilmiştir. REMARC eksik verileri, ikili sınıfları, sürekli ve kategorik öznitelikleri işleyebilir. REMARC yöntemi sadece risk değeri tahmin etmekle kalmaz aynı zamanda her bir öznitelik üzerinde analiz yapar ve karar verme esnasında alan uzmanlarına değerli bilgiler sağlar. REMARC'ın performansı, UCI veriseti deposundan elde edilen birçok veri seti ile support vector machine naïve Bayes, decision trees (karar ağaçları) ve boosting (arttırma) yöntemleri gibi modern

algoritmalar kullanılarak değerlendirilmiştir. AUC ölçütüyle yapılan değerlendirmeler göstermektedir ki REMARC diğer birçok makina öğrenmesi yönteminden önemli derecede daha iyi tahmin performansına sahiptir ve diğer yöntemden çoğundan daha hızlı çalışmaktadır.

Kardiyovasküler cerrahi alanı, REMARC yöntemi ile yeni risk tahmini çerçevesi oluşturmak amacıyla seçilmiştir. TurkoSCORE projesi, REMARC algoritmasının öğrenme aşaması için veri toplamak amacıyla kullanıldı. REMARC'ın tahmin performansı, en popüler kardiyovasküler cerrahi riski değerlendirme yöntemlerinden biri olan EuroSCORE ile karşılaştırıldı. EuroSCORE Türk hastalar üzerinde değerlendirildi ve EuroSCORE modelinin Türk nüfusü için yeterli olmadığı gösterildi. Sonra, EuroSCORE ve tahmin için REMARC kullanan TurkoSCORE'un tahmin performansı karşılaştırldı. Deneysel değerlendirmeler göstermektedir ki REMARC Türk hasta popülasyonunda EuroSCORE'a göre daha iyi tahmin performansı göstermektedir.

Anahtar Kelimeler: Risk Tahmini, AUC azamileştirme, AUC, Sıralama, Kardiyovasküler operasyon risk değerlendirmesi

Acknowledgements

This thesis would not have been possible without the great support I have received from some special people. First of all, I am heartily thankful to my supervisor Prof. Dr. Halil Altay Güvenir. During these two years I realized that I could not have asked for a better person to guide me in my research. He has been a great mentor from whom I learned a lot. It was a great pleasure to work with him in this thesis.

I grateful to know these incredible people, Serkan Durdu and Çağın Zaim, in the Department of Cardiovascular Surgery at Ankara University who supported me in my research. Specially, I am indebted to Assoc. Prof. Dr. Rüçhan Akar for his invaluable suggestions during my research. Without his dedication, some parts of this thesis could not be completed.

I would like to thank to Asst. Prof. Dr. Çiğdem Gündüz Demir for accepting to read and review this thesis. I would like to show my gratitude to TUBITAK (The Scientific and Technological Research Council of Turkey) since they have supported me in my master studies.

I would like to thank members of room EA511; Can, Volkan, Funda for their valuable friendship. I would specially like to thank to my collage Gönenç for his valuable comments during my researches.

Last but not least, I would like to thank to my parents, Pervin and Yahya for their love and support that always kept me motivated. Without them I would not be able to come so far. Finally, I would like to thank my beloved fiancée Betül

for being in my life and for her undying support. Nothing would be the same without her.

To My Family,

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Accurate prediction of risk is essential for life. Avoiding or being aware of risks in domains such as finance or medicine can save money and lives, respectively. The main motivation behind the research on risk-prediction systems is to improve system performance to avoid unwanted events or negative consequences.

This thesis proposes a new risk measure and a supervised machine learning algorithm to estimate the values of this measure. The algorithm, learning from training instances, develops a mode of the domain based on receiver operating characteristics (ROC) analysis, so that the area under ROC curves (AUC) of ordering the instances will be maximized [1]; hence, the algorithm is called Risk Estimation by Maximizing the Area under ROC Curve (REMARC).

Specific risk estimation methods have been developed for finance [2] medicine [3, 4] and insurance [5] to name some examples. Some methods are dependent on statistical models while some are based on machine learning algorithms. The machine learning algorithms are usually classification

algorithms that can associate a certainty factor with their classification. The certainty factor for a predicted unwanted case is taken as the value of risk.

The word “risk” is generally taken to mean “an unwanted situation” [6]. Although these unwanted cases may be severe, their likelihood of occurrence is usually rare. Therefore, datasets for such domains usually are unbalanced and the costs of misclassification are not symmetric. Classification algorithms that aim to maximize accuracy are not suitable for such unbalanced datasets [7, 8, 9]. Instead, an alternative metric called AUC, proposed by Bradley, is the evaluation metric to maximize [10]. AUC has important features such as insensitivity to class distribution and cost distributions [10, 11, 9], which make it suitable for risk domains.

In risk domains, representing the risk score as a real value between 0 and 1 may not be sufficient, and even misleading; relatively ordering instances in terms of risk values may be much more informative. For example, instances can be located on a single dimension, where the safest cases are on one side and the riskiest cases are on the other side. Since it has been shown by Hanley and McNeil that AUC is able to qualify ranking instances, maximizing AUC also leads to the best ranking [1]. Recent research on maximizing AUC by Toh *et al.* [12] and Rakotomamonjy also shows the importance of ranking instances [13].

The REMARC method is not able to handle continuous data without preprocessing. All continuous features should be discretized first. In this thesis in addition to the REMARC method, a discretization method called Maximum Area under ROC curve based Discretization (MAD) is proposed.

The main contributions of the REMARC algorithm can be shown in three different ways. First, we show the conditions a risk scoring function must possess in order to achieve maximum AUC for a single feature dataset case. Second, the maximization of AUC is extended over the whole dataset by using a

simple heuristic, which also depends on AUC's metric. Lastly, the human readable model formed by REMARC helps domain experts by indicating what features and how their particular values affect the risks.

Cardiovascular surgery domain is selected as a test domain for REMARC. There are important reasons behind this choice. First of all, risk evaluation methods are being used in order to inform cardiac patients properly about the mortality risk of surgery by taking into consideration risk factor of patients. The predictions obtained by using these methods are also valuable for monitoring the surgical care and checking the surgical quality with the accepted norms. Since the patients risk factors are taken into consideration, operative mortality can be used as a measure of surgical quality. Therefore, different machine learning approaches have been proposed to predict mortality risks of patients undergoing cardiovascular surgeries [14, 15, 16, 17].

EuroSCORE risk model is learned by using nearly 20 thousand patients from 128 hospitals in eight European countries [14]. EuroSCORE method has been used in Turkish cardiovascular surgery departments in order to assess mortality risk of patients. Validation of EuroSCORE has been analyzed in countries outside of Europe [18, 19]. According to these researches, there exist crucial differences between the patient populations across the nations. As a result, the EuroSCORE risk prediction model is not validated in some patient populations. Therefore, in this thesis the evaluation of EuroSCORE model on Turkish patients is analyzed. After analyzing EuroSCORE model on Turkish population, the predictive performance of REMARC used in TurkoSCORE system is compared with EuroSCORE. Since REMARC performs better than EuroSCORE, the REMARC algorithm is proposed as a new cardiovascular surgery risk estimation system.

In the next chapter, literature summary about the risks, risk domains, ROC, AUC, AUC maximization, discretization are given. Chapter 3 covers the

theoretical background of the REMARC method, implementation details and empirical evaluation of REMARC. In Chapter 4, REMARC is applied to cardiovascular surgery domain and compared by EuroSCORE model. Finally, Chapter 5 concludes with some directions for future work.

Chapter 2

Background

In this chapter, the background information needed to understand the concepts in the following chapters is provided. The risk subject is investigated in detail. The ROC and AUC subjects are given since they are essential in REMARC. AUC maximization subject is discussed in this chapter, as well. Discretization subject is also investigated in order to provide background information for the MAD method.

2.1 Risks

Risk has always been a normal occurrence. Risks such as a complication from surgery, a fraudulent financial transaction, a firm going into financial distress and an e-mail being spam are all part of today's world. Giddens claims that the ideas of risk and responsibility are closely linked in a risk society, and suggests that legal theorists and practitioners should also concern themselves with the idea and reality of risk [6]. The word "risk" is commonly used in daily life, because of its popularity in the media, however, a formal definition is needed.

2.1.1 Definitions of Risk

Hansson gives five definitions of risk commonly used in different disciplines [20]. Hansson's third definition is the most suitable for defining the risk used in this thesis: "The probability that an unwanted event may or may not occur". For example, the risk of a credit card transaction being fraudulent is 17%.

2.1.2 Risk Domains

Risk implies an unwanted situation. In medicine, mortality and morbidity are two unwanted situations. In finance, money loss and bankruptcy are examples. Since the consequences of these situations are crucial, in order to avoid them extensive research continues on this subject. As an example, it is possible to find books written on specific domains such as process management systems risk estimation [21].

According to Shishkin and Savkov some of the most popular commercial risk analysis tools for financial domains are "Risk Watch" (www.riskwatch.com, USA) and "Commercial Risk Analysis and Management Methodology-CRAMM" (www.cramm.com) [22]. Other than the commercial tools, concepts such as Value-At-Risk (VAR) and other models can be found in literature [2], [23]. Stoyan *et al.* provide a survey on stochastic models for risk estimations [24]. Recently, Ferrari and Paterlini proposed a new risk estimation method that claims a better performance than VAR [25].

In medicine, a risk scoring system based on logistic regression for cardiovascular surgery is proposed by Roques *et al.* [26]. Other scoring systems for the same domain also exist [3, 27]. A recent study by D'Agostino *et al.* shows that some of these scoring systems [4] use Cox regression methods, which is proposed by Cox [28].

2.1.3 Risk Estimation in Machine Learning

Risk estimation is not yet a major subarea of machine learning literature. Classification algorithms, which are able to output the confidence or probability of classification results, can be used to approximate risk estimation.

In a risk estimation system, a risk function that assigns higher values to risky instances than safer instances is crucial. In such a system, risk will be computed as a real value between 0 and 1, where 1 indicates the definite risk while 0 represents the safest situation. However, the absolute value of this risk score is also very important for the user. Assume *risk*() is a function that returns a real number between 0 and 1 as the estimation of the risk. Another risk function, *risk'*(), defined as $\sqrt{risk()}$, also returns a value between 0 and 1. Both of these functions will rank the instances in the same order, although their absolute risk values are different.

On any dataset gathered from a risk domain, two classes should be determined in order to distinguish a risky situation from a safe one. In this thesis, we will define these class labels as **p** (positive, unwanted class) and **n** (negative, safe class). For example, in a loan dataset, the class label **p** indicates a default, while label **n** indicates that the loan amount has been paid back.

Machine learning techniques have been applied to different domains in order to predict risk. In medicine, Colombet *et al.* evaluated three different machine learning algorithms in order to predict cardiovascular surgery risk [29]. Biagioli *et al.* used Bayesian models to predict risks in coronary artery surgery operations [30] and Gamberger *et al.* evaluated machine learning results on a heart database [31]. Financial domains have also taken advantage of machine learning algorithms. Galindo and Tamayo evaluated machine learning and statistical methods in order to predict credit risks [32]. Kim proposed a financial time series prediction system by using a support vector machine (SVM) [33] and

Min and Lee tried to predict bankruptcy risk by using optimal kernel functions for SVM [34]. However, to the best of our knowledge, a risk estimation system that aims to maximize the AUC metric has never been proposed. The ROC curves and AUC metric will be examined in detail before explaining the REMARC method. The next section elaborates on the features of ROC and AUC and their appropriateness for this thesis.

2.2 ROC, AUC and AUC maximization

Since their application to machine learning, ROC graphs and the AUC metric have become popular; AUC is used in evaluating machine learning algorithms and as a learning criterion. We explain the properties that make AUC a better metric than accuracy and discuss the existing research on AUC maximization.

2.2.1 Receiver Operating Characteristics (ROC)

The first application of ROC graphs dates back to World War II, where they were used to analyze radar signals [35]. Since then, they have been used in areas such as signal detection and medicine [36, 37, 38]. The first application to machine learning is done by Spackman [39]. According to Fawcett's definition, the ROC graph is a tool that can be used to visualize, organize and select classifiers based on their performance [9]. It has become a popular performance measure in the machine learning community after it has been realized that accuracy is often a poor metric to evaluate classifier performance [40, 41, 11].

The ROC literature is more established to deal with binary classification (two classes) problems than multi-class ones. At the end of the classification phase, some classifiers simply map each instance to a class label (discrete output). Some classifiers are able to estimate the probability of an instance belonging to a specific class such as naïve Bayes or neural networks (continuous valued output, also called score). Classifiers produce a discrete output represented by only one point in the ROC space, since only one confusion matrix is produced

from their classification output. Continuous-output-producing classifiers can have more than one confusion matrix by applying different thresholds to predict class membership. In this thesis, all instances with a score greater than the threshold are predicted to be **p** class and all others are predicted to be **n** class. Therefore, for each threshold value, a separate confusion matrix is obtained. The number of confusion matrices is equal to the number of ROC points on an ROC graph. With the method proposed by Domingos, it is possible to obtain ROC curves even for algorithms that are unable to produce scores [42].

ROC space is two dimensional space with a range of (0.0, 1.1) on both axes. In ROC space the y-axis represents the true positive rate (*TPR*) of a classification output and the x-axis represents the false positive rate (*FPR*).

To calculate *TPR* and *FPR* values, the definitions of the elements in the confusion matrix must be given. The structure of a confusion matrix is shown in Figure 2.1. True positives (*TP*) and false positives (*FP*) are the most important elements of the confusion matrix for ROC graphs. For each threshold value, *TP* is equal to the number of positive instances (those that have been classified correctly) and *FP* is equal to the number of negative instances (those that have been misclassified).

		Actual Class	
		p	**n**
Predicted Class	**p**	TP	FP
	n	FN	TN
Column Totals:		**P**	**N**

Figure 2.1 ROC curves of the REMARC method with TurkoSCORE risk factors

TPR and *FPR* values are calculated by using Eq. 2.1. In this equation *N* is the number of total negative instances and *P* is the number of total positive instances.

$$TPR = TP / P \qquad \textbf{Eq. 2.1}$$

$$FPR = FP / F$$

As mentioned above, the classifiers producing continuous output can form a curve since they are represented by more than one point in the ROC graph. To draw the ROC graph, different threshold values are selected and different confusion matrices are formed.

By varying the threshold between $-\infty$ and $+\infty$, an infinite number of ROC points can be produced for a given classification output. However, this operation is computationally costly and it is possible to form the ROC curve more efficiently with other approaches.

As proposed by Fawcett, in order to calculate the ROC curve efficiently, classification scores are sorted in an increasing order first [9]. Starting from $-\infty$, each distinct score element is taken as a threshold; TPR and FPR values are calculated using Eq. 2.1.

As an example, assume that the score values for test instances and actual class labels for a toy dataset are given in Table 2.1. The ROC curve for this toy dataset is shown in Figure 2.2. In this figure, each ROC point is given with the threshold value used to calculate it. In a dataset with *S* distinct classifier scores, there are *S*+1thresholds including $-\infty$ and the same number of ROC points. Since there are eight distinct score values in this toy dataset, there are nine ROC points. With this simple method it is possible to calculate the ROC curve in linear time.

Class Label	n	n	n	p	p	n	p	p	p
Score	-7	-3	0	0	4	7	8	10	11

Table 2.1 A Toy dataset given with hypothetical scores

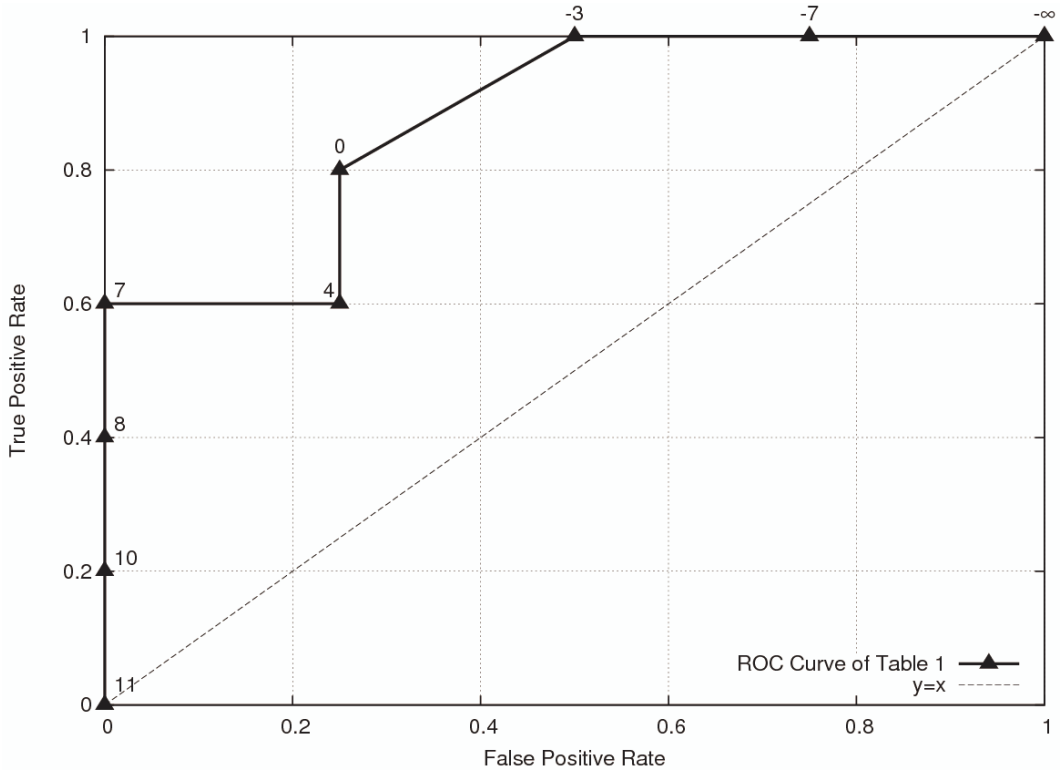

Figure 2.2 ROC graph of the given toy dataset in Table 1 including the *y=x* line in order to show random performance.

It is possible to divide the ROC space into three regions: the region above *y=x* line, the area below *y=x* line and the points on the *y=x* line. The points on *y=x* line represent random performance. As an example, a classifier that has a point on (0.6,0.6) guesses the positive class 60% correctly, however it also has a 60% false positive rate. The points above the *y=x* line are those belonging to the classifiers that have an acceptable trade-off between the positive and negative classes; similarly, the points below the *y=x* line correspond to an unacceptable classification performance. A classifier's ROC point below the diagonal line can be negated by simply inverting the decision criteria of the classifier, replacing all **p** class labels with **n** class labels and vice versa. According to Flach and Wu

classifiers below the diagonal have valuable information, but they are not able to use it [43].

2.2.2 Area under the ROC Curve (AUC)

ROC graphs are useful to visualize the performance of a classifier but a scalar value to compare classifiers is needed. In the literature, Bradley proposes the area under the ROC curve as a performance measure [10]. According to the AUC measure, the classifier with a higher AUC value performs better in general. A classifier can be outperformed by another classifier in some regions of ROC space, for some specific threshold values, even though the classifier, which has larger AUC, is better than the other.

The ROC graph space is a one-unit square. The highest possible AUC value is 1.0, which represents the perfect classification. In ROC graphs a 0.5 AUC value means random guessing has occurred and values below 0.5 are not realistic as they can be negated by changing the decision criteria of the classifier.

The AUC value of a classifier is equal to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Hanley and McNeil show that this is equal to the Wilcoxon test of ranks [1].

2.2.3 Why AUC is More Proper than Accuracy

There are several reasons why we chose AUC as a learning criterion in this thesis. The first reason is the independence of the decision threshold of the AUC metric. Since the risk estimation methods are not actual classifiers, unless a threshold is fixed it is not possible to calculate an accuracy value. As mentioned in Section 2.1.3, the first task of a risk estimation method is ranking instances correctly. Since AUC has the ability to measure the quality of ranking, it is better than accuracy metric on this basis.

Another reason regards the discrimination power of the accuracy and AUC metrics. Bradley was the first author to question the applicability of accuracy metrics in classifier algorithms and to recommend the use of AUC instead [10]. Provost *et al.* also questioned the applicability of accuracy metrics in classification algorithms and suggested ROC analysis as a powerful alternate tool [41]. Rosset claimed that even if the goal is to maximize accuracy, AUC may be better than empirical error for discriminating between models [44]. The formal proof of the superiority that AUC has over accuracy is later given by Huang and Ling [11]. In their work, the authors showed that AUC is a statistically consistent and more discriminating metric than accuracy. These works clearly show the discriminatory power of the AUC metric.

Skewed (unbalanced) datasets is another reason to prefer AUC as a metric. This situation occurs when the difference between class priors is high. Risk areas such as medicine [45, 8] and fraud detection [46] are examples of skewed datasets. For example, a classifier that predicts all instances as negative even though a few of the instances achieve very high accuracies is misleading [13]. In addition, class distribution can change over time. For example, if in a financial crisis a large number of banks claim bankruptcy, this can change class distribution drastically. In order to solve such problems, AUC, which is insensitive to class distributions, is preferred.

Lastly, misclassification costs cannot be determined for most risk domains. As noted above, skewed datasets are common in real life. In a domain with unbalanced class distribution, when the true misclassification cost is higher than implied by the distribution of training set examples, this situation becomes problematic [47]. Since AUC is also insensitive to misclassification cost, it is preferred in this thesis [48].

2.2.4 AUC Maximization

Most classification algorithms are designed to maximize accuracy (or error rate). Since accuracy is a classification performance criterion, algorithms that maximize it give better predictive performance. However, because of the aforementioned drawbacks to the accuracy metric for some domains, AUC has become more popular. It has been shown that maximizing accuracy does not lead to maximizing AUC [49, 50]. As a result, new algorithms maximizing AUC have been proposed.

Some approximation methods to maximize the global AUC value have been proposed by researchers [51, 50, 52]. Ferri *et al.* proposed a method to locally optimize AUC in decision tree learning [53], and Cortes and Mohri proposed boosted decision stumps [49]. To maximize AUC in rule learning, several new algorithms have been proposed [54, 55, 56]. A nonparametric linear classifier based on the local maximization of AUC was proposed by Marrocco *et al.* [57]. A ROC-based genetic learning algorithm has been proposed by Sebag *et al.* [7]. Marrocco *et al.* used linear combinations of dichotomizers for the same purpose [58]. Freund *et al.* gave a boosting algorithm combining multiple rankings [59]. Cortes and Mohri showed that this approach also aims to maximize AUC [49]. A method by Tax *et al.* that weighs features linearly by optimizing AUC has been proposed and applied to the detection of interstitial lung disease [8]. Ataman *et al.* advocate an AUC-maximizing algorithm with linear programming [60]. Rakotomamonjy suggested rank optimizing kernels for SVMs to maximize AUC [13]. Ling and Zhang compare AUC-based Tree-Augmented Naïve Bayes (TAN) and error-based TAN algorithms; the AUC-based algorithms are shown to produce more accurate rankings [61]. More recently, Calders and Jaroszewicz proposed a polynomial approximation of AUC to optimize it efficiently [62]. Linear combinations of classifiers are used to maximize AUC in biometric scores fusion in Toh *et al.* [12]. Han and Zhao propose a linear classifier based on active learning, which maximizes AUC [63].

2.3 Discretization

Discretization methods aim to find the cut-points that form the intervals in the process of discretization. A continuous attribute is then treated as a discrete attribute whose number of intervals is known on the continuous space.

Liu et al. categorized discretization algorithms on four axes [64]. These categories include *supervised vs. unsupervised, splitting vs. merging, global vs. local, and dynamic vs. static*.

Simple methods such as equal-width or equal-frequency binning algorithms do not use class labels for instances during the discretization process [65]. These methods are called *unsupervised discretization methods*. To improve the quality of the discretization, methods that use class labels are proposed; they are referred to as s*upervised discretization methods*. Splitting methods take the given continuous space and try to divide it into small intervals by finding proper cut-points, whereas merging methods handle each distinct point on the continuous space as an individual candidate for a cut-point and merges them into larger intervals. Some discretization methods process localized parts of the instance space during discretization. As an example, the C4.5 algorithm handles numerical values by using a discretization (binarization) method that is applied to localized parts of the instance space [66, 67]; these methods are called *local methods*. Methods that use the whole instance space of the attribute to be discretized are called *global methods*. Dynamic discretization methods use the whole attribute space during discretization and perform better on data with interrelations between attributes. Conversely, static discretization methods discretize attributes one by one and assume that there are no interrelations between attributes. According to the categories defined above, MAD is a supervised, merging, global, and static discretization method.

Splitting discretization methods usually aim to optimize measures such as entropy [68, 69, 70, 71, 72], which aims to obtain pure intervals, dependency [73] or accuracy [74] of values placed into the bins. On the other hand, the merging algorithms proposed so far use the chi-square statistic [75, 76, 77]. As far as we know, the ROC Curve has never been employed in the discretization domain.

Chapter 3

REMARC

This chapter presents detailed information about the REMARC method. First of all, a brief introduction to REMARC is given. Then, the risk function designed for categorical features to maximize AUC and the details of the MAD method and its application to REMARC is given. The REMARC method and its implementation are detailed. Finally, in the empirical evaluations REMARC method is compared with other machine learning on real life datasets and results are discussed.

3.1 REMARC Introduction

REMARC is a risk estimation method designed to maximize the AUC metric. The REMARC algorithm reduces the problem of finding a risk function for the whole set of features into finding a risk function for a single categorical feature, and then combines these functions to form one risk function covering all features. We will show here that it is possible to determine risk functions that achieve the maximum AUC for a single categorical feature. REMARC discretizes the numerical features by an algorithm called MAD, proposed by

Kurtcephe and Guvenir [78]. The MAD method discretizes a continuous feature in a way that results in a categorical feature by maximizing the AUC.

For a given query, REMARC outputs a real value *r* in the range of [0,1] as the estimated risk of being the unwanted state. This *r* value is roughly the probability that the query instance will be in the **p** class. It is only a rough estimate of probability, since it is very likely that no other instance with exactly the same feature values has been observed in the training set. The REMARC algorithm determines this estimated probability by computing the weighted average of probabilities computed on single features. The weight of a feature is a linear function of its AUC value calculated by the risk estimates for each instance in the training set. A higher value of AUC for a feature is an indication of its higher relevance in determining the class label.

3.2 Single Categorical Feature Case

A categorical feature has a finite set of choices. Let $V = \{v_1, v_2, \ldots v_n\}$ be a categorical feature and v_i be a categorical value that feature V can take. The dataset D is a set of instances represented by a vector of n values and class label as $<v,c>$, where $v \in V$ and $c \in \{\mathbf{p},\mathbf{n}\}$

Given a dataset D with a single categorical feature whose value set is $V = \{v_0, v_1,\ldots, v_n\}$, a risk function r: $V \rightarrow [0,1]$ can be defined to rank the values in V. According to this risk function, a value v_i comes after a value v_j if and only if $r(v_i) > r(v_j)$; hence r defines a partial ordering on the set V. A pair of consecutive values v_i and v_{i+1} defines a ROC point R_i on the ROC space. The coordinates of the point R_i are (FPR_i, TPR_i).

Theorem 1: Let D be a dataset with a single categorical feature whose value set is $V = \{v_0, v_1, \ldots, v_n\}$. Let r: $V \rightarrow [0,1]$ be the risk function that orders the values of V, as v_{i+1} comes after v_i if $r(v_{i+1}) > r(v_i)$, for all values of $0 \leq i < n$. If the values

of the risk function for two consecutive values v_i and v_{i+1} are swapped, then the only change in the ROC curve is that the ROC point corresponding to the v_i and v_{i+1} values moves to a new location so that the slopes of the line segments adjacent to that ROC point are swapped.

Proof: The slope of the line segment between two consecutive ROC points R_i and R_{i+1} is

$s_i = \frac{TPR_i - TPR_{i+1}}{FPR_i - FPR_{i+1}}$. Since $TPR_i = \frac{TP_i}{P}$ and $FPR_i = \frac{FP_i}{N}$,

$s_i = \frac{N}{P}\frac{TP_i - TP_{i+1}}{FP_i - FP_{i+1}}$.

Further replacing $TP_i = P_i + TP_{i+1}$ and $FP_i = N_i + NP_{i+1}$, where P_i is the number of **p**-labeled instances with value v_i, and N_i is the number of **n**-labeled instances with value v_i.

$s_i = \frac{N}{P}\frac{P_i}{N_i}$.

Similarly, the slope of the line segment connecting the ROC points between R_{i+1} and R_{i+2} is

$s_{i+1} = \frac{N}{P}\frac{P_{i+1}}{N_{i+1}}$.

When the ranking of values v_i and v_{i+1} are changed, only the following changes take place:

$P'_{i+1} = P_i, \quad P'_i = P_{i+1}$,
$N'_{i+1} = N_i, \quad N'_i = N_{i+1}$,

$\forall_{j \neq i, i+1} \quad P'_j = P_j$ and $N'_j = N_j$.

With this change, only the ROC point R_i at (FPR_i, TPR_i) is replaced with a new ROC point R'_i at (FPR'_i, TPR'_i). The slopes of the new line segments adjoining R'_i are

$s'_i = \frac{N}{P}\frac{P'_i}{N'_i}$ and $s'_{i+1} = \frac{N}{P}\frac{P'_{i+1}}{N'_{i+1}}$.

Replacing the new count values with the old ones,

$s'_i = s_{i+1}$ and $s'_{i+1} = s_i$ are obtained. ■

For example, consider the dataset given below:

D={(a,n), (b,p), (b,n), (b,n), (b,n), (c,p), (c,p), (c,n), (c,n), (d,p), (d,p), (d,n)}, where *V* = {a, b, c, d}. If a risk function *r* orders the values of *V* as r(a) < r(b) < r(c) < r(d), the ROC curve shown in Figure 3.1a will be obtained. On the other hand, if the rankings of values b and c are swapped, the ROC curve shown in Figure 3.1b will be obtained. A similar technique was used earlier by Flach and Wu to create better prediction models for classifiers [43].

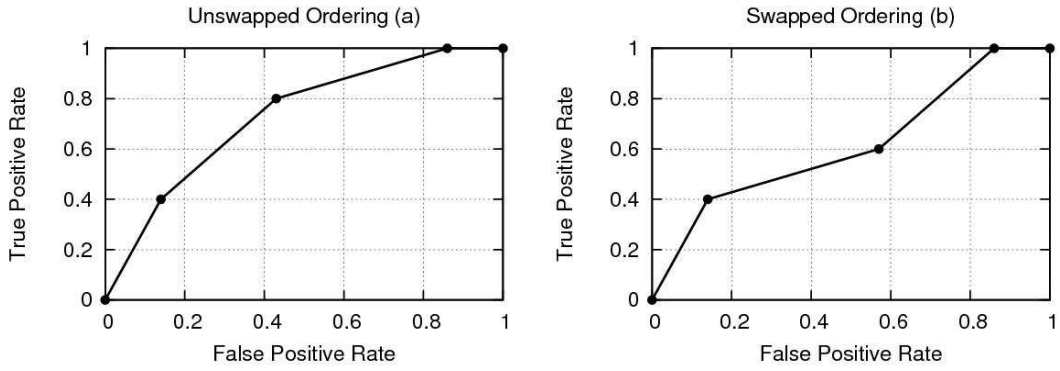

Figure 3.1 Effect of swapping the risk values of two feature values

Theorem 1 shows how concavities in a ROC curve can be removed, resulting in a larger AUC. The next question is how to form the convex ROC curve. The following theorem sets the necessary and sufficient condition for risk functions to satisfy so that their ROC curves are convex.

Theorem 2: Let *D* be a dataset with a single categorical feature that takes values from the set $V = \{v_0, v_1, \ldots, v_n\}$. Let r: $V \rightarrow [0,1]$ be the risk function that orders the values of *V*, as v_{i+1} comes after v_i if $r(v_{i+1}) > r(v_i)$ for all values of $0 \leq i < n$. In order for the ROC curve of the ordering by *r* to be convex, the following condition must be satisfied:

$$\forall_{0 \le i < n} \; \frac{P_{i+1}}{N_{i+1}} \ge \frac{P_i}{N_i}, \qquad \textbf{Eq. 3.1}$$

where P_i is the number of **p**-labeled instances with value v_i, and N_i is the number of **n**-labeled instances with value v_i.

Proof: In order for the ROC curve to be convex, the slopes of all line segments connecting consecutive ROC points starting from the ROC point (1,1) must be non-decreasing.

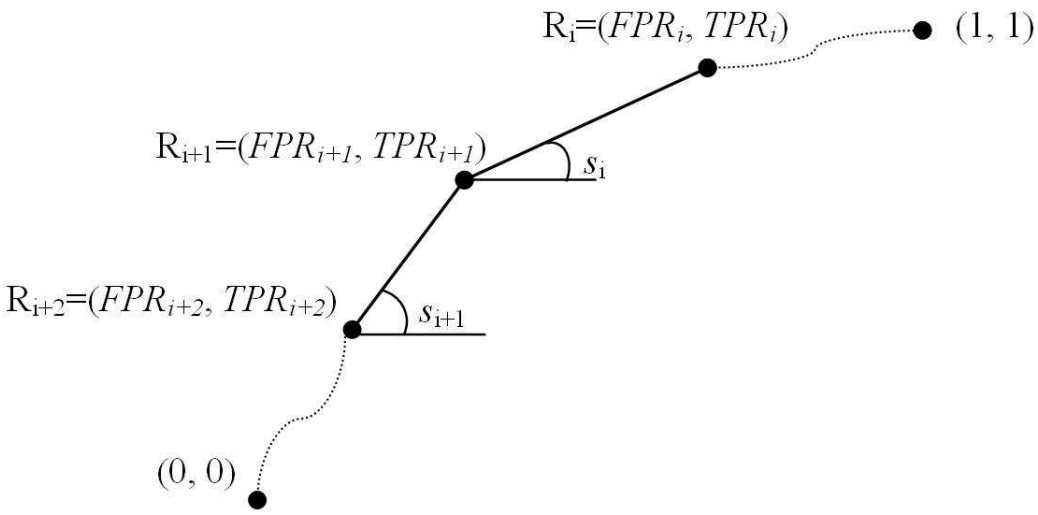

Figure 3.2 Relation between the slopes of two consecutive line segments in a convex ROC curve

Therefore, the condition for a convex ROC curve is

$$\forall_{0 \le i < n} \; s_i \ge s_{i+1} \qquad \textbf{Eq. 3.2}$$

$$\forall_{0 \le i < n} \; \frac{TPR_{i+1} - TPR_{i+2}}{FPR_{i+1} - FPR_{i+2}} \ge \frac{TPR_i - TPR_{i+1}}{FPR_i - FPR_{i+1}}$$

By definition, $TPR_i = \frac{TP_i}{P}$.

Further, due to the ordering of values, $TP_i = P_i + TP_{i+1}$.

Hence, $TPR_i - TPR_{i+1} = \frac{TPR_i}{P} - \frac{TPR_{i+1}}{P} = \frac{1}{P}\left(P_i + TP_{i+1} - TP_{i+1}\right) = \frac{P_i}{P}$

Similarly,

$FPR_i - FPR_{i+1} = \frac{N_i}{N}$, $TPR_{i+1} - TPR_{i+2} = \frac{P_{i+1}}{P}$ and $FPR_{i+1} - FPR_{i+2} = \frac{N_{i+1}}{N}$.

Therefore, the inequality in Eq. 3.1 can be rewritten as

$$\forall_{0 \le i < n} \quad \frac{P_{i+1}/P}{N_{i+1}/N} \ge \frac{P_i/P}{N_i/N} .$$

Finally, $\forall_{0 \le i < n} \quad \frac{P_{i+1}}{N_{i+1}} \ge \frac{P_i}{N_i}$. ■

Therefore, according to Theorem 2, any risk function r that assigns a higher value to v_{i+1} than to v_i when $\frac{P_{i+1}}{N_{i+1}} \ge \frac{P_i}{N_i}$, for all values of V, will result in a convex ROC curve. For example, a risk function defined as $r(v_i) = \frac{P_i}{N_i}$ will result in a convex ROC curve.

Theorem 3. Let D be a dataset with a single categorical feature whose value set is $V = \{v_0, v_1, \ldots, v_n\}$. Ignoring the ineffective ROC points that lie on a line, there exists exactly one convex ROC curve.

Proof: Since there exists only one possible ordering of values of V that satisfies the condition given in Theorem 1, there exists only one convex ROC curve. ■

The general assumptions for risk estimation problems are given in Eq. 3.3:

$$\underset{0\le i<n}{\forall i} \quad P_i \ge 0 \quad , \quad \underset{0\le i<n}{\forall i} \quad N_i \ge 0 \qquad \textbf{Eq. 3.3}$$

$$P = \sum_{0}^{n-1} P_i > 0 \quad , \quad N = \sum_{0}^{n-1} N_i > 0$$

Although the dataset is guaranteed to have at least one instance with class label **p** and one instance with label **n**, it is possible that for some values of i, N_i may be 0. In such cases the risk function defined above will have undefined values. In order to avoid such problems, the risk can be defined as

$$r(v_i) = \frac{P_i}{P_i + N_i} \qquad \textbf{Eq. 3.4}$$

Lemma 1. $\underset{0\le i<n}{\forall i} \quad \frac{P_{i+1}}{P_{i+1} + N_{i+1}} \ge \frac{P_i}{P_i + N_i}$ iff $\frac{P_{i+1}}{N_{i+1}} \ge \frac{P_i}{N_i}$

Proof : if $\underset{0\le i<n}{\forall i} \quad \frac{P_{i+1}}{P_{i+1} + N_{i+1}} \ge \frac{P_i}{P_i + N_i}$, then $\underset{0\le i<n}{\forall i} \quad P_{i+1}(P_i + N_i) \ge P_i(P_{i+1} + N_{i+1})$,

and $\underset{0\le i<n}{\forall i} \quad \frac{P_{i+1}}{N_{i+1}} \ge \frac{P_i}{N_i}$.

The same arithmetic operations can be applied in the reverse direction to show that

if $\underset{0\le i<n}{\forall i} \quad \frac{P_{i+1}}{N_{i+1}} \ge \frac{P_i}{N_i}$, then $\frac{P_{i+1}}{P_{i+1} + N_{i+1}} \ge \frac{P_i}{P_i + N_i}$ ■

Since, if both P_i and N_i are 0 for some i, the corresponding value v_i can be completely removed from the dataset $\underset{0\le i<n}{\forall i} P_i + N_i > 0$, and this risk function is defined for all values of i.

The risk function $r(v_i)=\frac{P_i}{(P_i+N_i)}$ has another added benefit in that it is simply the probability of the **p** label among all instances of value $v_{i,}$ which is easily interpretable.

Corollary: For a dataset *D* with a single categorical feature whose value set is $V = \{v_0, v_1, \ldots, v_n\}$, the risk function defined as $r(v_i)=\frac{P_i}{(P_i+N_i)}$ gives the maximum possible AUC.

Therefore, the REMARC algorithm uses $r(v_i)=\frac{P_i}{(P_i+N_i)}$ as the risk function for categorical features.

3.2.1 The Effect of the Class Label Choice on a Feature's AUC

In order to calculate the *P* and *N* values one of the classes should be labeled as **p** and the other class as **n**, but one can question the effect this choice has on the AUC value. It is possible to show that the AUC value of a categorical feature is independent from the choice of class labels by using the value from the Wilcoxon-Mann-Whitney statistics.

In Eq. 3.5, the AUC formula based on the Wilcoxon-Mann-Whitney statistics is given. *P* is the number of instances that have the **p** class label and *N* represents the number of **n**-class-labeled instances. The set D_p represents the **p**-labeled instances and D_n represents the **n**-labeled instances. An element belonging to D_p set, which is D_{pi}, is the ranking of the i^{th} instance, which is labeled **p**. Inversely, an element belonging to D_n set, such as $D_{ni,}$ is the ranking of the i^{th} instance, which is labeled **n**.

$$AUC = \frac{\sum_{i=1}^{P}\sum_{j=1}^{N} f(D_{pi}, D_{ni})}{PN} \qquad \textbf{Eq. 3.5}$$

$$f : \begin{bmatrix} D_{pi} > D_{ni} = 1 \\ D_{pi} < D_{ni} = 0 \\ D_{pi} \equiv D_{ni} = 0.5 \end{bmatrix}$$

The dividend part of the AUC formula in Eq. 3.5 counts the number of **p**-labeled instances for each element of the D_p set whose ranking is higher than any element of the D_n set. Then, AUC is calculated by dividing this summation by the multiplication of the **p**-labeled and **n**-labeled elements.

The effect of the class label choice on the AUC calculation should be investigated. First of all, it is straightforward that the divisor part of the AUC formula is independent of class choice. Then, assume that the risk score $r_i = \frac{P_i}{P_i + N_i}$ is used on the D dataset and D_p and D_n sets are formed. Let n_i be the number of **n**-labeled instances whose ranking is lower than the i^{th} element of the D_p set and let r_i be the score assigned to this element. When the classes are swapped, the new risk value r'_i is equal to 1- r_i. With this property all instance scores are negated. However, negating scores does not change the relative ranking but inverses it. So, the AUC formula in Eq. 3.5, which calculates AUC depending on the ranking of the instances, is independent of the class-label decision when the proper risk scoring is used.

3.2.2 An Example Toy Dataset

Assume that a toy training dataset with a single categorical feature is given in Table 3.1. In order to calculate the AUC value of this particular feature, risk values are needed. The risk values are calculated by the proposed risk function.

The sorted version of the dataset according to the risk estimates is given in Table 3.2. The AUC value of this feature is calculated by using Eq. 3.1. The *P* value is 7 and the *N* value is 6. The AUC value is $\frac{34.5}{7*6} = 0.82$. In order to calculate this AUC value, for each **p**-labeled instance all **n**-labeled instances whose risk (ranking) is smaller or equal should be counted. When the class labels are swapped the risks are also swapped. The sorted version of the swapped toy dataset is given in Table 3.3. Since the relative ranking of the instances does not change the new AUC value is also $\frac{34.5}{6*7} = 0.82$.

Class Label	n	n	p	n	n	p	n	p	p	n	p	p	p
Feature Value	a	a	a	a	b	b	b	c	c	c	d	d	d

Table 3.1 Toy training dataset with one categorical feature

Risk	0.25	0.25	0.25	0.25	0.33	0.33	0.33	0.66	0.66	0.66	1.00	1.00	1.00
Class Label	n	n	p	n	n	p	n	p	p	n	p	p	p
Feature Value	a	a	a	a	b	b	b	c	c	c	d	d	d

Table 3.2 Training datasets risk values are calculated and instances are sorted in ascending order

Risk	0.0	0.0	0.0	0.33	0.33	0.33	0.66	0.66	0.66	0.75	0.75	0.75	0.75
Class Label	n	n	n	n	n	p	p	n	p	p	p	n	p
Feature Value	d	d	d	c	c	c	b	b	b	a	a	a	a

Table 3.3 Negated version of the training dataset. The risk values are calculated again and instances are sorted in ascending order

3.3 Handling Continuous Features

Having found the necessary and sufficient conditions for the risk function for a categorical feature to result in the maximum possible AUC, the next problem is to determine a mechanism for handling the continuous features. An obvious and trivial risk function maps any real value seen in the training set with the class

value **p** to 1 and any real value with the class value **n** to 0. This risk function will result in the maximum possible value for AUC, which is 1.0. However, such a risk function will over fit the training data, and will be undefined for unseen values of the feature, which are very likely to be seen in the query instance. So, our first requirement for a risk function for a continuous feature is that it must be defined for all possible values of that continuous feature. A straightforward solution to this requirement is to discretize the continuous feature by grouping all consecutive values with the same class value to a single categorical value; the cut off points can be set to the middle point between feature values of differing class labels. The risk function, then, can be defined using the risk function given in Eq. 3.4 for categorical features. Although this would result in a risk function that is defined for all values of a continuous function, it would still suffer from the over fitting problem. In order to overcome this problem, the REMARC algorithm makes the following assumption:

Assumption 1: The risk values are either non-increasing or non-decreasing for the increasing values of a continuous feature.

Although there exist some features in real-world domains that do not satisfy this assumption, in the datasets we examined this assumption is satisfied in general.

This assumption is also consistent with the interpretations of the values of continuous features in many real-world applications. For example, in a medical domain, a high value of fasting blood glucose is an indication for a high risk of diabetes. On the other hand, low fasting blood glucose is an indication of a risk for another heath problem, called hypoglycemia.

3.3.1 The MAD Method

The REMARC algorithm requires all features to be categorical. Therefore, the continuous features in a dataset need to be categorized. The aim of a

discretization method is to find the proper cut-points in order to categorize a given continuous feature. After the discretization process a continuous feature is treated as a discrete feature whose number of intervals is known on the continuous.

The MAD method is designed to maximize the AUC value by checking the ranking quality of values of a continuous feature. The MAD algorithm given in Kurtcephe and Guvenir is defined for multi-class datasets [78]. A special version of the MAD method, called MAD2C and defined for two-class problems, is used in REMARC.

In order to measure the ranking quality of a continuous feature, the instances are sorted in ascending order. Sorting is essential for all discretization methods in order to produce unambiguous intervals. After the sorting operation, feature values are used as hypothetical score values and the ROC graph of the feature is drawn. The AUC of the ROC curve shows the overall ranking quality of the continuous feature. In order to obtain the maximum AUC value, only the points on the convex hull must be selected. The minimum number of points that form the convex hull is found by eliminating the points that cause concavities on the graph. In each pass, the MAD method compares the slopes in the order of the creation of the hypothetical lines, finds the junction points (cut-points) that cause concavities and eliminates them. This process is repeated until there is no concavity on the graph. The points left on the graph are the cut-points, which will be used to discretize the feature.

It has been proven that the MAD method finds the cut-points and the AUC value of the feature independently from the class choice. It is shown that the cut-points found by MAD never separate two consecutive instances of the same class. This is an important property, as it shows that a discretization method works properly. The implementation details, formal proofs and empirical evaluation of MAD can be found in Kurtcephe and Guvenir [78].

3.3.2 A Toy Dataset Discretization Example

It is possible to visualize the discretization process by using the MAD method. A toy dataset for the discretization is given in Table 3.4. After the sorting operation, the ROC points are formed. This ROC graph is given in Figure 3.3. Since the risk values are either non-increasing or non-decreasing for the increasing values of a continuous feature, two ROC graphs are formed. As can be seen in Figure 3.3 one of these graphs is below the diagonal line since the risk is increasing with increasing values of the continuous feature.

Class Value	n	n	p	n	n	p	n	p	p	n	p	p	p
F1	1	2	3	4	5	6	6	7	8	9	10	11	12

Table 3.4 A toy dataset for visualizing MAD in two-class problems. The name of the attribute to be discretized is F1

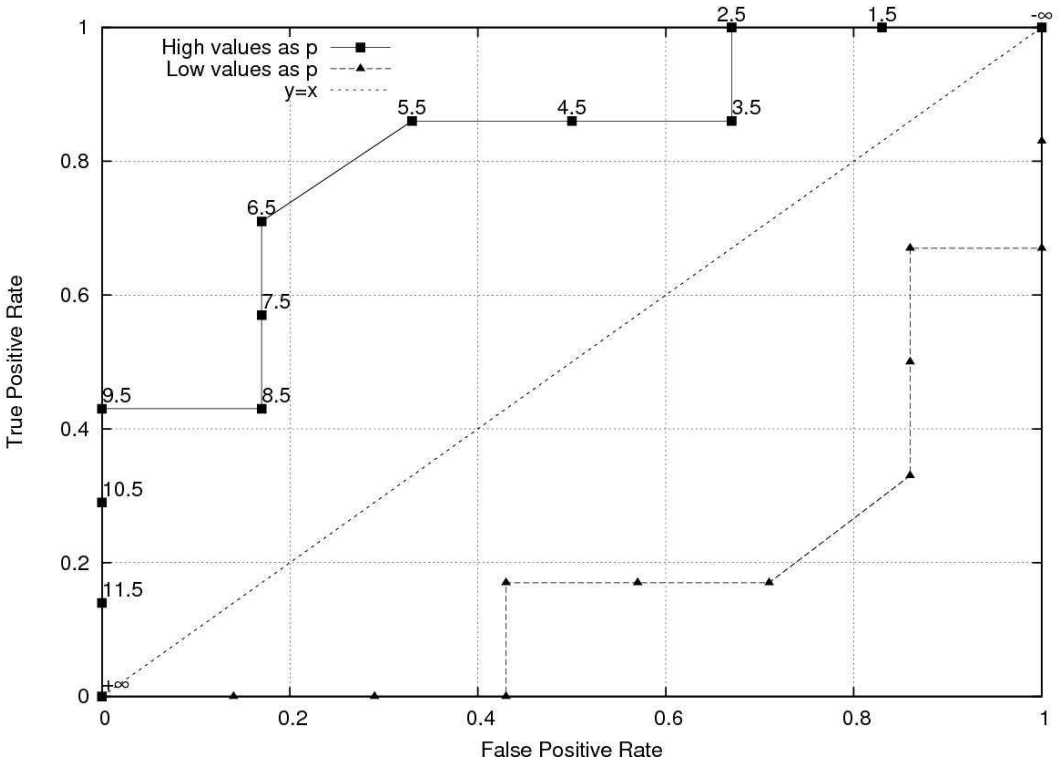

Figure 3.3 Visualization of the ROC points in a two-class discretization.

The first pass of the MAD method is shown in Figure 3.4. All points below or on the diagonal are ignored since they have no positive effect on the maximization of AUC. Then the points causing concavities are eliminated. MAD converged to the convex hull in one pass for this example. The points left on the graphs are the discretization cut-points.

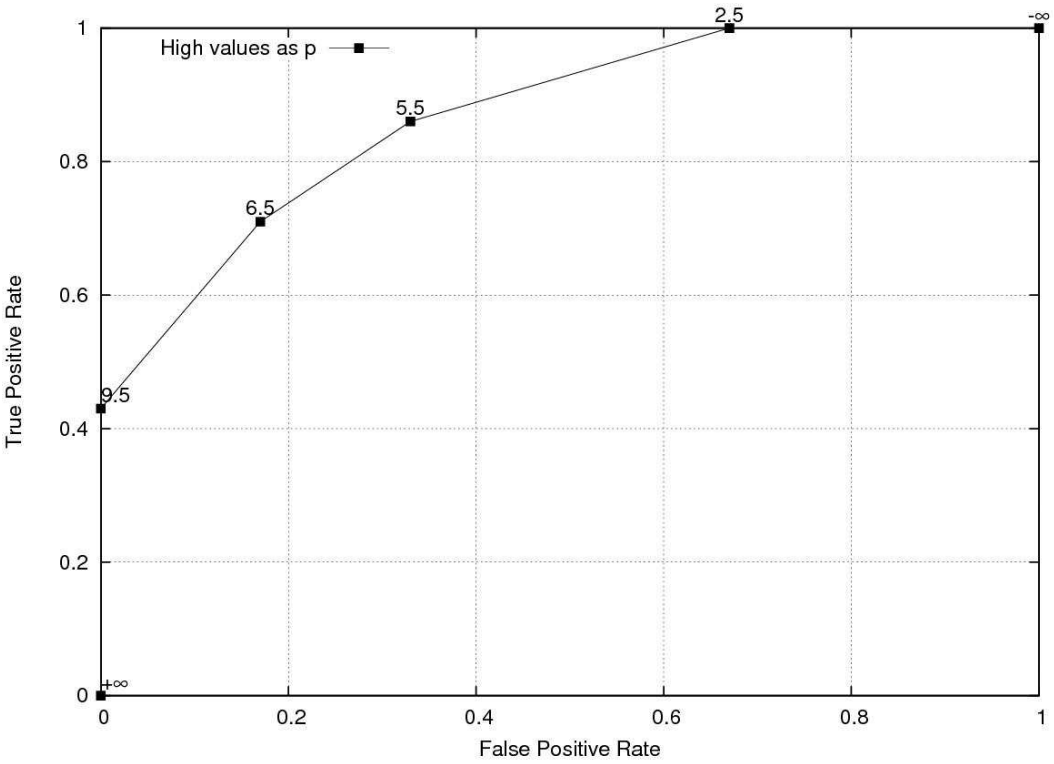

Figure 3.4 Final cut-points after the first pass of convex hull algorithm

3.4 REMARC Algorithm

The training phase of the REMARC algorithm is given in Figure 3.5. In the training phase all continuous features are discretized. In order to discretize continuous features, MAD2C, which is shown on the fifth line of Figure 3.5, is used. Risk values are calculated for each value of a given categorical feature (discretized continuous features are included). In this step, the risk function defined in Eq. 3.4 is used in order to obtain the optimal ranking for categorical features. Then, training instances are sorted according to the risk values

calculated in the previous step. Since the risk function used by REMARC always results in a convex ROC curve, the AUC is always equal to or greater than 0.5. Therefore, the REMARC algorithm learns a weight w_i for a feature f_i as

$$w = (AUC_f - 0.5) * 2 \qquad \textbf{Eq. 3.6}$$

The ROC curve of an irrelevant feature is simply a diagonal line from (0,0) to (1,1), with $AUC = 0.5$. The weight function in Eq. 3.6 assigns 0 to such irrelevant features in order to eliminate them. The risk values and weights of the features are stored for the testing phase.

```
1 :REMARCTrain (trainSet[M][N]) // Includes M features and N train
instances
2 :   Begin
3 :      for i=0 to M-1
4 :         if(isContinuous(trainSet[i]));
5 :            cutPoints=MAD2C(trainSet[i][0..N-1]);
6 :            numericalValuesToCatVal(cutPoints,trainSet[i]);
7 :         risks[i]<-computeCategoricalRisk(trainSet[i][0..N-1]);
8 :         sortInstancesByRisk(trainSet[i][0..N-1]);
9 :         aucValues[i]<-computeAUC(trainSet[i][0..N-1]);
10:         featureWeights[i]=(aucValues[i]-0.5)*2;
11:      end
12:   end
```

Figure 3.5 Algorithm of the REMARC method's training phase.

The testing phase of the REMARC method is straightforward, as for each feature, the risk value corresponding to the value of the feature in the test instance is used. Then the risk of this feature is weighted by its weight, which is calculated in the training phase. The computation of the risk for a query instance q is in Eq. 3.7. The maximization of AUC for whole dataset is a challenging problem. Cohen et al. showed that the problem of finding the ordering that agrees best with a learned preference function is NP-Complete [79]. This weighting mechanism is used as a simple heuristic in order to extend this maximization over the whole feature set.

$$\text{risk}(q) = \frac{\sum_f w_f * P(p \mid q_f)}{\sum_f w_f} \qquad \textbf{Eq. 3.7}$$

$$w_f = \begin{cases} 2(AUC_f - 0.5) & q_f \text{ is known} \\ 0 & q_f \text{ is missing} \end{cases}$$

where $P(p \mid q_f)$ is the probability of q being **p**-labeled, given that the value of feature f in q is q_f, and w_f is the weight of the feature f, calculated by using Eq. 3.6

Finally, in order to obtain the weighted average, all risk and weight values are summed and final risk is calculated by dividing the cumulative.

```
1 :REMARCTest (testInstance[M][1])
2 :   Begin
3 :      for i=0 to M-1
4 :         oneFeatureRisk= risks[i][testInstace[i][0]];
5 :         totalRisk+= oneFeatureRisk * featureWeights[i];
6 :         totalWeight+= featureWeights[i];
7 :      end
8 :      return totalRisk/totalWeight;
9:   end
```

Figure 3.6 Testing phase algorithm of the REMARC method

The time complexity of the MAD algorithm is given as $O(n^2)$, where n is the number of training instances. After discretizing the numerical features the time complexity of the REMARC algorithm is $O(m*v\lg v+n)$, where m is the number of features and v is the average number of values per feature. As a result, REMARC is bounded by the MAD algorithm's time complexity.

3.5 Interpretation of the REMARC Predictive Model

As mentioned above, the REMARC method does not only provide risk estimation as a single real value, but the predictive model used in order to

estimate risk can provide useful information to domain experts. A high weight value indicates that the corresponding feature is a highly effective risk factor in the given domain. Domain experts may choose to ignore features with low weights, potentially reducing the cost of record keeping.

Some of the categorical features are formed by discretizing continuous features. For example, age can be discretized into child, youth, adult and elderly. Assume that the impression of the feature age is investigated on a risky domain, such as medicine. The intervals should be chosen carefully since they can affect a system's predictive performance. The domain experts can provide this information. However, there can be experimental domains where this knowledge is not applicable. The MAD method used in REMARC learns the proper intervals in order to maximize AUC during the training phase. These intervals also report the risks associated with each interval. For example, consider a dataset that contains an age feature and a class label that indicates the presence of a new disease. The MAD method will find the distinct age groups in terms of this disease and the REMARC method will determine the risk for each age group.

The choice of class label during risk estimation has no effect on the feature weights. However, the risk function used by REMARC depends on this choice directly, as shown in Section 4.1, so in order to interpret risk scores correctly one must pay attention to the class label that represents the unwanted situation. Otherwise, risk scores can be misleading.

3.6 Empirical Evaluations

In order to maximize AUC the theoretical background of the REMARC method is given. In order to support the theoretical background with empirical results two different experiments are conducted. First, REMARC is compared with 26 different machine learning algorithms on an AUC basis. Then, since there can be

domains where the predictive models have to trained often, running times of the algorithms are also measured.

The real-life datasets are provided by the UCI machine learning repository (Frank and Asuncion 2010) and are two-class problems [80]. Ten datasets are selected from risk domains such as medicine and finance. The properties of the datasets are given in Table 3.5.

Dataset Name	# Instances	# Continuous Attributes	#Categorical Attributes	# Dataset Abbreviations
Australian	690	6	8	A
Bupa	345	6	0	B
Crx	653	9	6	C
Heart (Statlog)	270	7	6	H1
Hypothyroid	3164	7	18	H2
Mammographic Masses	961	1	5	M
Pima-Diabetes	768	8	0	P
Sick-Euthyroid	3163	7	18	S1
SPECTF	267	44	0	S2
Wisconsin-Breast	569	30	0	W

Table 3.5 Properties of the datasets used in the empirical evaluations of the REMARC algorithm

In order to perform the comparisons, 26 different classification algorithms are selected from the WEKA software package [81]. Only the algorithms that able to produce continuous output (confidence on the class decision) are selected. As mentioned above, the ROC graphs of algorithms producing continuous output are meaningful. Since REMARC is a non-parametric method, none of the classifiers is optimized for each dataset. All classifiers are used with default settings of WEKA for the sake of fairness. The SVM is taken from the LIBSVM package provided in WEKA [82].

3.6.1 Predictive performance

Researchers reported that some of the algorithms that aim to maximize AUC do not obtain significantly better AUC values than the ones designed to maximize

accuracy [49, 63]. Therefore, it is important to show that REMARC can outperform accuracy-maximizing algorithms statistically significantly.

Algorithms/Datasets	A	B	C	H1	H2	M	P	S1	S2	W	Average
REMARC	0.923	0.659	0.931*	0.904*	0.986	0.901*	0.827	0.942	0.857*	0.986	**0.892**
AdaBoostM1+	0.922	0.737	0.926	0.888	0.990*	0.895	0.804	0.966	0.801	0.985	0.891
Class.ViaRegr.+	0.918	0.727	0.918	0.882	0.990*	0.896	0.827	0.986*	0.763	0.989*	0.890
Bagging+	0.918	0.755*	0.910	0.872	0.980	0.888	0.822	0.972	0.795	0.977	0.889
ADTree+	0.917	0.705	0.925	0.880	0.988	0.887	0.802	0.979	0.803	0.984	0.887
Logistic+	0.912	0.714	0.915	0.900	0.970	0.893	0.831*	0.956	0.801	0.972	0.886
MultiC.Classifier+	0.912	0.714	0.915	0.900	0.970	0.893	0.831*	0.956	0.801	0.972	0.886
AODE+	0.928*	0.540	0.930	0.904*	0.989	0.900	0.823	0.963	0.820	0.988	0.879
NaiveBayes++	0.895	0.641	0.900	0.897	0.977	0.895	0.816	0.920	0.850	0.980	0.877
BayesNet+	0.920	0.540	0.928	0.901	0.989	0.899	0.818	0.959	0.825	0.986	0.876
ThresholdSelector+	0.904	0.699	0.916	0.898	0.969	0.892	0.826	0.956	0.686	0.969	0.871
MultiBoostAB+	0.908	0.673	0.908	0.865	0.988	0.886	0.790	0.955	0.709	0.981	0.866
DecisionTable+	0.917	0.574	0.910	0.883	0.989	0.876	0.801	0.971	0.678	0.972	0.857
FT++	0.898	0.721	0.853	0.824	0.943	0.874	0.751	0.907	0.752	0.984	0.851
LWL++	0.911	0.643	0.909	0.839	0.955	0.886	0.775	0.942	0.674	0.948	0.848
FilteredClassifier++	0.899	0.540	0.893	0.836	0.958	0.863	0.794	0.949	0.683	0.939	0.835
REPTree++	0.879	0.666	0.871	0.824	0.963	0.846	0.768	0.957	0.631	0.924	0.833
PART++	0.867	0.645	0.853	0.785	0.966	0.882	0.778	0.954	0.652	0.937	0.832
Att.Sel.Classifier++	0.869	0.584	0.875	0.801	0.952	0.867	0.786	0.914	0.624	0.938	0.821
END++	0.865	0.648	0.877	0.777	0.940	0.868	0.758	0.939	0.593	0.939	0.821
OrdinalC.Classifier++	0.865	0.648	0.877	0.777	0.940	0.868	0.758	0.939	0.593	0.939	0.821
J48 (C4.5) ++	0.865	0.648	0.877	0.777	0.940	0.868	0.758	0.939	0.593	0.939	0.821
VFI++	0.913	0.562	0.910	0.871	0.782	0.836	0.550	0.755	0.853	0.946	0.798
DecisionStump++	0.833	0.572	0.848	0.688	0.951	0.788	0.696	0.936	0.623	0.886	0.782
IBk++	0.801	0.634	0.798	0.743	0.766	0.799	0.648	0.752	0.592	0.947	0.748
RBFNetwork++	0.732	0.509	0.787	0.835	0.581	0.786	0.642	0.676	0.641	0.755	0.694
SVM-RBF++	0.628	0.609	0.602	0.509	0.952	0.872	0.518	0.735	0.466	0.760	0.655

Table 3.6 The comparison of the predictive performance of REMARC algorithm with other algorithms on AUC metric. 10 datasets are used during evaluation. Algorithms marked with ++ are outperformed by REMARC method with a statistically significant difference Algorithms marked with + are outperformed by REMARC on average with no significant difference. AUC values marked with * are the best AUC values for that dataset (Higher results better)

A stratified ten-fold cross validation is employed to calculate AUC values for each datasets. As shown in Table 3.6, the REMARC method outperformed all algorithms on the average AUC. A paired t-test is used to decide whether the differences on averages are significant. According to the paired t-test on a 95% confidence level (the same level will be used for other t-tests) REMARC

statistically significantly outperforms 15 of the 26 machine learning algorithms on the average AUC. These algorithms include naïve Bayes, decision trees (part, C4.5) and SVM with a RBF kernel. REMARC outperformed the other 11 algorithms, as well, but the differences between the averages for these algorithms are not statistically significant.

One important point should be mentioned about the SVMs. As seen in Table 3.6, SVM has the worst predictive performance among all the classification algorithms because of the absence of parameter tuning. However, as mentioned before none of the algorithms is tuned for best predictive results.

The classification algorithms such as logistic (multinomial logistic regression model) and classification via regression achieve high AUC values. As mentioned above, these models are highly used in the domain of medicine, and in this work their predictive performance is validated.

The second classifier with the highest AUC was the Adaboost method. Since it is an ensembling algorithm, it uses a base classifier (default DecisionStump in WEKA). We believe that the performance of REMARC can be further improved by using an ensembling algorithm, as then, a statistically significant difference can be obtained.

3.6.2 Running Time

The REMARC method is designed to be simple, effective and fast. It handles categorical features close to the linear time. MAD requires more time since it uses sorting. Theoretically, REMARC seems fast, but empirical experiments must be conducted to support this claim.

The overall running times of the training phase of 25 different algorithms are calculated. The running times of all algorithms are measured using java virtual machines' CPU time and hundreds of results are averaged (to be objective). The

SVM algorithm is not included in the running times section since WEKA uses an outside library for this algorithm. However, it takes seconds for SVM to complete the training phase, so it is much slower than REMARC. The results of the overall running time for the other algorithms are shown in Table 3.7.

Algorithms/Datasets	A	B	C	H1	H2	M	P	S1	S2	W	Average
VFI--	16*	5*	17*	9*	104*	12*	14*	111*	23*	31*	34
DecisionStump--	24	9	23	11	113	17	30	116	43	105	49
NaiveBayes--	30	11	29	18	164	30	37	154	61	102	64
AODE--	53	16	52	27	352	40	56	350	113	280	134
BayesNet--	49	17	52	28	425	54	60	393	81	244	140
REPTree-	77	33	80	33	395	105	136	569	124	185	174
FilteredClassifier-	119	16	107	48	462	86	116	858	155	265	223
J48 (C4.5)-	160	75	153	81	635	147	181	1283	256	378	335
Att.Sel.Classifier-	173	31	159	99	761	201	162	867	419	496	337
REMARC	106	37	94	49	1131	128	164	1117	191	470	349
OrdinalC.Classifier+	162	79	150	83	681	177	197	1380	255	378	354
END+	236	119	218	126	880	219	254	1563	333	460	441
RBFNetwork++	404	136	348	148	1057	367	280	1149	712	620	522
PART++	416	102	511	193	865	230	248	2023	684	473	574
AdaBoostM1++	302	132	296	156	1584	302	394	1579	475	1135	635
MultiBoostAB++	318	133	297	164	1614	317	388	1622	477	1140	647
MultiC.Classifier++	991	79	1175	125	3698	239	226	3656	473	963	1163
Logistic++	1014	74	1215	124	3720	261	257	3667	472	935	1174
ADTree++	689	276	645	469	3149	579	973	3536	1562	2717	1459
Ibk+	172	33	169	35	7365	233	153	7465	124	222	1597
Bagging++	740	295	727	295	4484	636	961	6904	1070	1718	1783
ThresholdSelector++	1695	137	2032	227	6442	429	387	6346	1118	2028	2084
DecisionTable++	1582	156	1699	434	10824	411	635	11498	1183	2526	3095
Class.ViaRegr. ++	5340	1355	5573	1143	9943	3912	3593	15598	2218	2662	5134
FT++	4705	847	4879	927	14834	2856	2230	30959	1796	2324	6636
LWL+	2094	376	1940	412	52652	2360	2652	52887	1414	6919	12370

Table 3.7 The comparison of the average running time performance of REMARC algorithm with other algorithms (in ms) . 10 datasets are used during evaluation. Algorithms marked with ++ symbol are outperformed by REMARC method on running time basis with a statistically significant difference. Algorithms marked with -- symbol outperformed REMARC method on running time basis with a statistically significant difference. + marked algorithms are outperformed by REMARC on average and – marked algorithms outperform REMARC on average with no significant difference. AUC values marked with * are the best AUC values for that dataset (Lower results better)

REMARC outperforms 12 different algorithms significantly according to a paired t-test on a running-time basis. These outperformed methods are shown by

the ++ symbol on Table 3.7. Five algorithms outperformed REMARC statistically significantly. These algorithms are shown with a -- symbol. The differences between the other seven methods on the table and REMARC are not significant.

Chapter 4

TurkoSCORE: Turkish System for Cardiac Operative Risk Evaluation

In this chapter, the REMARC method is applied to the cardiovascular surgery domain. The data is gathered from the TurkoSCORE system. Detailed information about TurkoSCORE project is given in this chapter. The EuroSCORE project, which is one of the most popular risk evaluation systems in cardiac surgery, is evaluated on TurkoSCORE dataset that contains data about the cardiovascular operations performed in some hospitals in Turkey. The properness of EuroSCORE risk model on Turkish patient population is investigated. In empirical evaluation section, EuroSCORE and REMARC are compared by using a dataset that consist of EuroSCORE features on AUC basis. In order to propose a new risk estimation framework specially designed for Turkish patients, most likely risk factors (highly discriminative) are identified and filtered by consultant surgeons. Then, the performance of REMARC algorithm is investigated on this dataset.

4.1 The TurkoSCORE Project

One of the major aims of the TurkoSCORE project is to construct a risk estimation system in order to predict the early mortality in patients undergoing cardiovascular surgeries in Turkey on the basis of objective risk factors [83].

The TurkoSCORE project includes a database system for storing cardiovascular surgical patient's data in Turkey. A variety of parameters including personal, preoperative, postoperative, follow up and mortality have been recorded in this project. The aim of the project is not only finding risk factors of the patient and estimating mortality risk of patients but also collecting shared information about the Turkish cardiac patients nationwide. A web application is designed in this project allowing doctors enter their patient data online. The same web application is also used by doctors in order to monitor, search, and print the health profile of their patients.

The TurkoSCORE project also aims to lead the cardiovascular research by supplying a wide range of data collected from different institutions. Currently, Cardiovascular Surgery Department of Ankara University, Acıbadem Hospital and Ankara Atatürk Hospital are supplying data for TurkoSCORE. More detailed information about the structure of the TurkoSCORE database can be found in Tunca [83].

4.2 EuroSCORE

In Europe, a model called the European System for Cardiac Operative Risk Evaluation (EuroSCORE) has been developed and commonly used by European cardiovascular surgeons. This system predicts the risk of operative mortality during surgery or 30-days after the surgery. This prediction is based on the

values of some parameters measured before operation. In the development of EuroSCORE, North American and European risk model studies were investigated [14]. Initially, as candidate risk factors, 68 preoperative and 29 operative parameters were selected. The risk factors, which are most likely useful, are identified and selected by consultant surgeons. However the selected risk factors were very similar to those in other American studies. The definitions of these factors were simplified in EuroSCORE. In order to learn the model used in EuroSCORE, nearly 20 thousand patients were gathered from 128 hospitals in eight European counties (Germany, France, UK, Italy, Spain, Finland, Sweden and Switzerland)

The potential risk factors are analyzed and their effect on risk estimation is investigated. Some of these risk factors were eliminated in order to obtain a better predictive model. As a result, seventeen risk factors were found useful for calculating the early mortality risk of a cardiovascular surgery. The details of these risk factors can be found in [14].

The first scoring system proposed in EuroSCORE is called Additive (Standard) scoring. Additive scoring is designed by using the β coefficients as weights for each risk factor. During the calculation of Additive EuroSCORE, the weights are summed together according to the existence of a risk factor for a patient. However, after some validation studies of Additive EuroSCORE on other cardiac datasets outside of Europe, the deficiency of Additive scoring is noted. Since the Additive EuroSCORE can sometimes underestimate the risk in very high risk patients, logistic regression based scoring system, called Logistic EuroSCORE is proposed. The logistic β coefficients and the formula of this scoring system can be found in [26].

4.3 EuroSCORE Validation on Turkish Patients

In this section, firstly, prevalence of risk factors in Turkish patient population and EuroSCORE patient population is compared. Since the EuroSCORE scoring system is trained by using European patient population, any difference between populations can affect the performance of EuroSCORE on Turkish population.

Definitions of some risk factors were not identical with EuroSCORE definition. Therefore, some approximations were made in order to complete the analysis. These approximations are listed in the, Table A.1, Appendix A.

Statistical analysis of risk factor prevalence is performed by using chi-square test for categorical values and unpaired t-test for continuous values. *P* values less that 0.05 is considered as significant

Currently, there are 9451 patients in TurkoSCORE database. In this thesis 8018 patients are used. These patients are selected from the ones whose EuroSCORE values and EuroSCORE parameters are complete. This selection was necessary since most of the analysis is based on EuroSCORE parameters and values.

4.3.1 Demographic results

There were significant differences between Turkish and European cardiac patient populations. The prevalence of risk factors for both populations is given in Table 4.1. When patient related factors are investigated, it is seen that the Turkish cardiac patient population is younger on average. There exist significantly more patients in Turkish population whose age is less than 60 and fewer patients in any other age interval. Turkish patients have higher incidence of chronic pulmonary disease and active endocarditis. Fewer patients in Turkish population have extracardiac arteriopathy and previous cardiac surgery. Critical preoperative state is more likely to be present in Turkish patients than European

patients. In cardiac related factors, Turkish patients are more likely to have unstable angina LV function, moderate dysfunction and recent myocardial infraction.

Risk factor	Turkish prevalence (%) (n=8018)	EuroSCORE prevalence (%) (n=19030)	P-value
Patient Related Factors			
Age			
Mean	59.49 years	62.5 years	**<0.001**
Standard deviation	12.02 years	10.7 years	
<60 years	46.9	33.2	**<0.001**
60—64 years	17.3	17.8	0.325
65—69 years	16.6	20.7	**<0.001**
70—74 years	13.1	17.9	**<0.001**
75+ years	6.1	9.6	**<0.001**
Female	28.6	27.8	0.325
Chronic pulmonary disease	13.4	3.9	**<0.001**
Extracardiac arteriopathy	8.6	11.3	**<0.001**
Neurological disease	1.3	1.4	0.181
Previous cardiac surgery	4.1	7.3	**<0.001**
Serum creatinine >200 mmol/l	1.9	1.8	0.515
Active endocarditis	3.2	1.1	**<0.001**
Critical preoperative state	9.0	4.1	**<0.001**
Cardiac Related Factors			
Unstable angina LV function	9.8	8.0	**<0.001**
Moderate dysfunction	29.9	25.6	**<0.001**
Severe dysfunction	5.3	5.8	0.103
Recent myocardial infarction	23.5	9.7	**<0.001**
Pulmonary hypertension	1.9	2.0	0.565
Operation Related Factors			
Emergency	4.3	4.9	**0.035 (<0.05)**
Other than isolated CABG	23.0	36.4	**<0.001**
Surgery on thoracic aorta	3.7	2.4	**<0.001**
Postinfarct septal rupture	0.1	0.2	0.069

Table 4.1 Prevelances of risk factors in Turkish and EuroSCORE population. The risk factors that have significant difference are shown in bold face. EuroSCORE prevelance values are taken from Roques et al. [84]

Operation related factors such as emergency or other than isolated coroner artery bypass grafting (CABG) have less prevalence in Turkish population than European. Also Turkish patients are more likely to have surgery on thoracic aorta than European patients. All these differences are statistically significant. There are no significant differences in the prevalence of the risk factors sex, age

interval between 60 and 64, neurological disease, serum creatinine, pulmonary hyper tension, severe dysfunction and postinfarct septal rupture.

4.3.2 Model Calibration and Discrimination

The EuroSCORE values of 8018 patients are used in this section. Predicted mortality is both calculated using Additive and Logistic EuroSCORE. Then, observed and predicted mortality of the patients compared with 95% confidence intervals. These analysis are done for both whole cohort and isolated CABG cohort. Chi-square statistics is employed for measuring the difference between the observed and predicted mortality over risk sections.

For entire cohort, 157 deaths are observed in 8018 patients, 1.96% overall mortality rate is calculated. The Additive EuroSCORE predicted 2.98% mortality rate ($P < 0.001$ vs. observed) and 3.17% mortality rate ($P < 0.001$ vs. observed) is predicted by Logistic EuroSCORE. As shown in Table 4.2, both scoring systems overestimated mortality at each risk tertile. The predictive performances of both models are fair with 0.76 AUC value. The ROC curves are given in Figure 4.1.

	Patients (deaths)	Observed mortality rate (%95 CI)	Predicted mortality rate (%95 CI)
EuroSCORE additive			
0-3 (Low risk)	5164 (39)	0.76% (0.52-0.99)	1.52% (1.49-1.55)
4-6 (Medium risk)	2186 (65)	2.97% (2.26-3.69)	4.78% (4.75-4.82)
7+ (High risk)	668 (53)	7.93% (5.88-9.98)	8.33% (8.20-8.47)
Total	8018 (157)	1.96% (1.65-2.26)	2.98% (2.93-3.03)
EuroSCORE logistic			
Low Risk	2673 (16)	0.60% (0.31-0.89)	1.07% (1.06-1.08)
Medium Risk	2673 (26)	0.97% (0.60-1.34)	1.99% (1.76-2.22)
High Risk	2672 (115)	4.30% (3.53-5.07)	6.45% (6.22-6.68)
Total	8018 (157)	1.96% (1.65-2.26)	3.17% (3.08-3.26)

Table 4.2 Predicted and observed mortality by EuroSCORE risk level for whole cohort. In logistic EuroSCORE analysis, patients are divided into three approximately equal risk quintiles

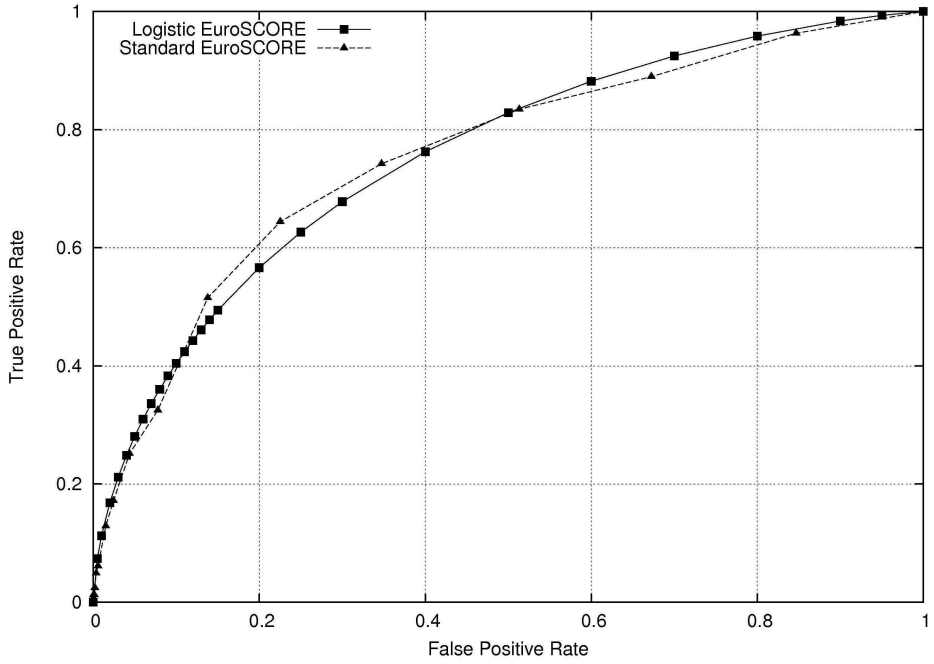

Figure 4.1 ROC curves for both Logistic and Standard EuroSCORE for whole cohort. In logistic EuroSCORE analysis, patients are divided into three approximately equal risk quintiles

.

	Patients (deaths)	Observed mortality rate (%95 CI)	Predicted mortality rate (%95 CI)
EuroScore additive			
0-3 (Low risk)	4042 (18)	0.45% (0.24-0.65)	1.54% (1.50-1.57)
4-6 (Medium risk)	1681 (31)	1.84% (1.20-2.49)	4.77% (4.73-4.81)
7+ (High risk)	448 (30)	6.70% (4.38-9.01)	8.12% (7.98-8.25)
Total	6171 (79)	1.28% (1.00-1.56)	2.89% (2.84-2.95)
EuroScore logistic			
Low Risk	2057 (11)	0.53% (0.22-0.85)	1.06% (1.05-1.07)
Medium Risk	2057 (8)	0.39% (0.12-0.66)	1.95% (1.74-2.16)
High Risk	2057 (60)	2.92% (2.19-3.64)	5.77% (5.56-5.99)
Total	6171 (79)	1.28% (1.00-1.56)	2.93% (2.84-3.02)

Table 4.3 Predicted and observed mortality by EuroSCORE risk level for isolated CABG cohort

Of 6171 patients undergoing isolated CABG, 79 deaths are observed, 1.28% overall mortality is calculated The Additive EuroSCORE predicted 2.89% mortality rate ($P < 0.001$ vs. observed) and 2.93% mortality rate ($P < 0.001$ vs. observed) is predicted by Logistic EuroSCORE. As shown in Table 4.3, both

scoring systems overestimated mortality at each risk tertile except the additive model in highest risk decile. The predictive performances of both models are fair with 0.77 AUC value for Additive and 0.76 for Logistic EuroSCORE. The ROC curves for both scoring systems are given in Figure 4.2

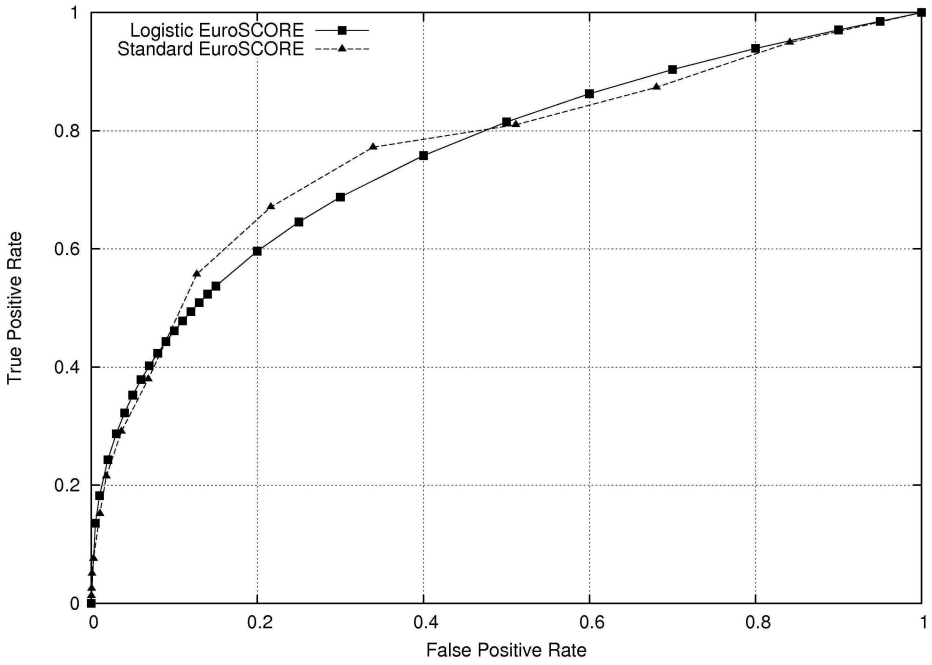

Figure 4.2 ROC curves for both Logistic and Standard EuroSCORE for isolated CAGB cohort

4.4 Comparison of REMARC and EuroSCORE

In the previous section, it is shown that the patient population, which is used in the training phase of EuroSCORE, is considerably different than Turkish cardiac patients. Since REMARC is proposed as a new scoring system for Turkish cardiac patients, it is essential to show that REMARC can predict early mortality risk better than EuroSCORE by using only EuroSCORE parameters.

In this section only the predictive performance of EuroSCORE and REMARC will be compared on AUC basis. The calibration of REMARC model is not available since the Turkish patient dataset is not large enough to create a validation set. As mentioned above some of the definitions of risk factors in

EuroSCORE are not identical with TurkoSCORE dataset. Therefore, the approximations given in Table A.1 are used in this section, as well.

There exist 9451 patients in the TurkoSCORE database currently. However, the number of patients whose EuroSCORE values are complete is 8018. The ROC curves for Additive and Logistic EuroSCORE are calculated over whole dataset, since EuroSCORE has a trained model. However, REMARC must be trained and test on the same dataset. Therefore, ten-fold cross-validation is employed in order to obtain the ROC curve.

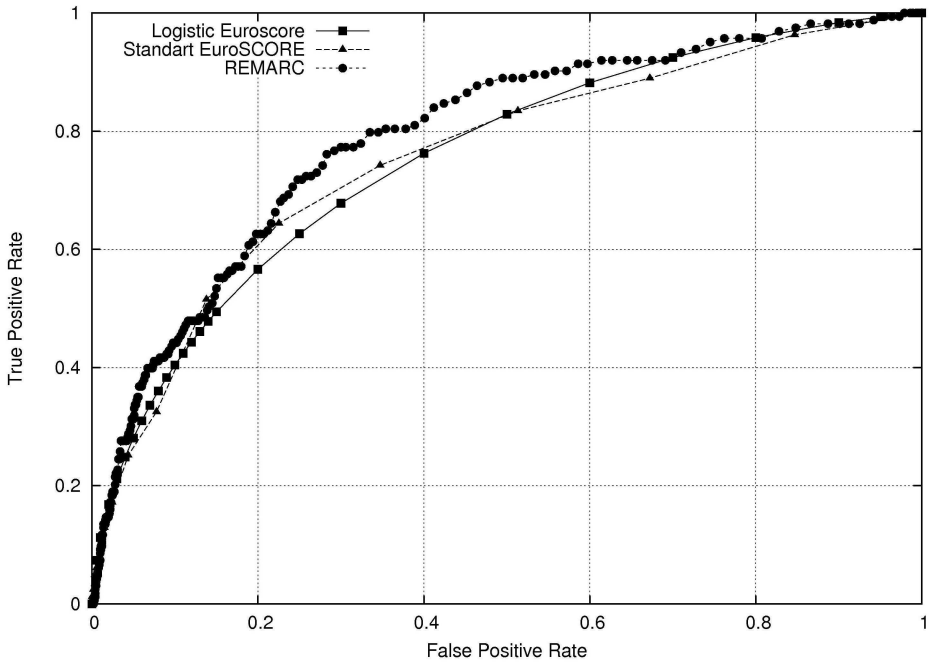

Figure 4.3 ROC curves for both Logistic EuroSCORE, Standard EuroSCORE and REMARC with EuroSCORE risk factors

Then, the AUC values are calculated using only the EuroSCORE risk factors. The AUC value for REMARC is 0.79 and 0.76 for both Additive and Logistic EuroSCOREs. The ROC curves for REMARC and EuroSCOREs can be found in Figure 4.3. Since the higher AUC represent better predictive performance, REMARC risk estimation method outperforms both EuroSCORE scoring systems over Turkish population.

4.5 REMARC Based Cardiovascular Risk Estimation System

According to the analysis done in the Section 4.4, REMARC performs better prediction than EuroSCORE by using only the risk factors used in EuroSCORE model. These risk factors are specially selected for the European surgeons by analyzing European patient population. Since REMARC enables domain experts to investigate discriminative ability of each feature by providing AUC values, EuroSCORE parameters are analyzed in this thesis, as well. Table B.1, Appendix B shows the AUC values of each EuroSCORE parameter. According to the AUC values of features, most of the features are predicting well except other than isolated CABG and postinfarct septal rupture features.

There exist 190 preoperative and 16 operative candidate risk factors in TurkoSCORE database [83]. In order to find the discrimination ability of each feature REMARC is used. The whole dataset (9451 patients) is used since the training phase of the REMARC ignores missing values. As a result, the AUC values for each feature (weights) are calculated. Most likely risk factors to be useful are identified by consultant cardiac surgeons in Cardiovascular Surgery Department, Ankara University by considering these weights. The risk factors whose AUC values are too close to 0.5 (irrelevant risk factors) and the risk factors that has few number of instances (more than %90 is missing) are eliminated. The risk factors left after these eliminations are shown in Table B.2, Appendix B.

After selecting the most important features, a new dataset with 28 features and 9451 instances is formed. The testing phase of REMARC is robust to missing values, as well. However, there exist some instances in the dataset most features are missing. In order to investigate the effect of missing values on AUC a simple experiment is conducted. 26 different datasets are formed. The first dataset, called dataset0, included only the instances which has no (0) missing

features (a complete dataset). The last dataset, called dataset25, is formed from the instance which can have 25 missing features (at most) of 28. Table 4.4 is formed by using these 26 datasets in REMARC program. During this analysis ten-fold cross validation is employed.

Dataset Names	# Instances	# P	# N	REMARC AUC	Additive AUC	Logistic AUC
Dataset 0	3584	59	3525	0,80	0,74	0,74
Dataset 1	5620	102	5518	0,83	0,76	0,77
Dataset 2	6871	127	6744	0,83	0,77	0,77
Dataset 3	7916	153	7763	0,80	0,76	0,76
Dataset 4	8179	166	8013	0,81	0,76	0,76
Dataset 5	8263	166	8097	0,81	0,76	0,76
Dataset 6	8315	167	8148	0,81	0,76	0,76
Dataset 7	8353	168	8185	0,81	0,76	0,76
Dataset 8	8394	170	8224	0,80	0,76	0,76
Dataset 9	8433	171	8262	0,80	0,76	0,76
Dataset 10	8476	174	8302	0,80	0,76	0,76
Dataset 11	8745	180	8565	0,80	0,76	0,76
Dataset 12	8773	181	8592	0,80	0,76	0,76
Dataset 13	8787	182	8605	0,80	0,76	0,76
Dataset 14	8816	182	8634	0,80	0,76	0,76
Dataset 15	8988	185	8803	0,80	0,76	0,76
Dataset 16	9055	186	8869	0,80	0,76	0,76
Dataset 17	9086	187	8899	0,80	0,76	0,76
Dataset 18	9201	190	9011	0,80	0,76	0,76
Dataset 19	9225	191	9034	0,80	0,76	0,76
Dataset 20	9236	191	9045	0,80	0,76	0,76
Dataset 21	9244	191	9053	0,79	0,76	0,76
Dataset 22	9311	191	9120	0,79	0,76	0,76
Dataset 23	9322	191	9131	0,79	0,76	0,76
Dataset 24	9334	191	9143	0,79	0,76	0,76
Dataset 25	9336	191	9145	0,79	0,76	0,76

Table 4.4 26 Different datasets are formed by eliminating the instances with missing values. Dataset*i* contains at most *i* many missing features from 28 features Number of instances, **p**, **n** values and AUC values are given. AUC values of the REMARC algorithm are calculated by ten-fold cross validation

According to the Table 4.4, the number of missing features is increases, naturally the number of instances increases, as well. The relationship between number of missing features and number of instances is given in Figure 4.4. According to this figure, the complete dataset (dataset0) has relatively low

number of instances compared to other datasets. The graphic get stabilized after dataset3.

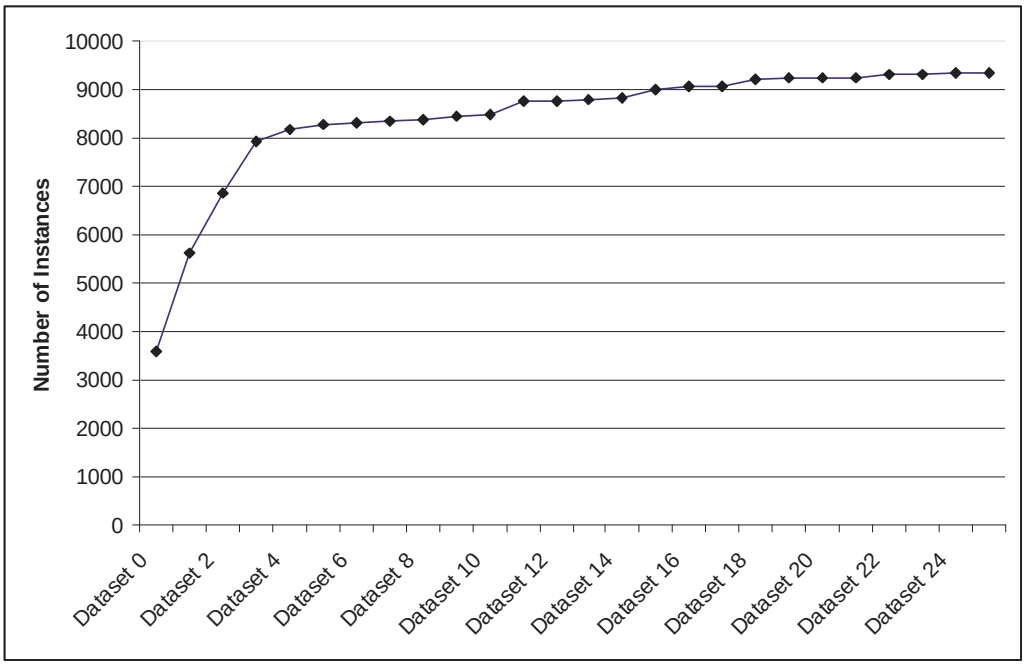

Figure 4.4 Number of missing values over different datasets, which contains different number of missing features

Another important aspect of this experiment is analyzing the AUC values over missing features. As mentioned above, the more missing features allowed, the more instances can be used in the training phase and it is expected that the model learned from more instances will perform better prediction. However, when instances with highly missing features are used in test phase, this noise will cause decrement in the predictive performance. Therefore, it is essential to choose the dataset which gives highest AUC value with relatively high number of instances. As shown in Figure 4.5, dataset1 and dataset2 has the same AUC value, 0.83. According to Table 4.4, dataset2 has higher number of instances (6871) than dataset1 (5620). As a result, in order to build the final predictive model dataset2 will be used in this section.

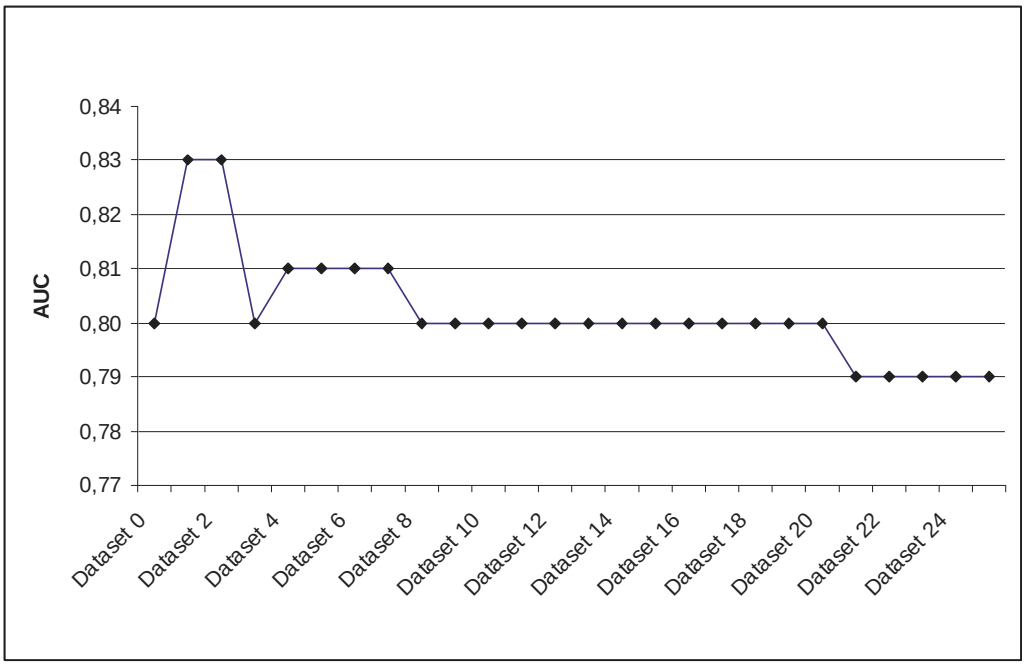

Figure 4.5 The distribution of AUC values obtained by ten-fold cross validation over 26 different datasets

As seen in Figure 4.5, the predictive performance of REMARC varies between 0.79 and 0.83. Therefore, it is possible to say that REMARC is a robust algorithm even with highly missing data.

After the performance optimizing experiment is done, 6871 (dataset2) patients of 9451 are selected. The EuroSCORE AUC values are recalculated again for these 6871 patients (0.77 for additive and logistic). The ROC curves of the REMARC and EuroSCORE Additive and Logistic are given in Figure 4.6. The AUC value of the REMARC method with TurkoSCORE parameters is 0.83.

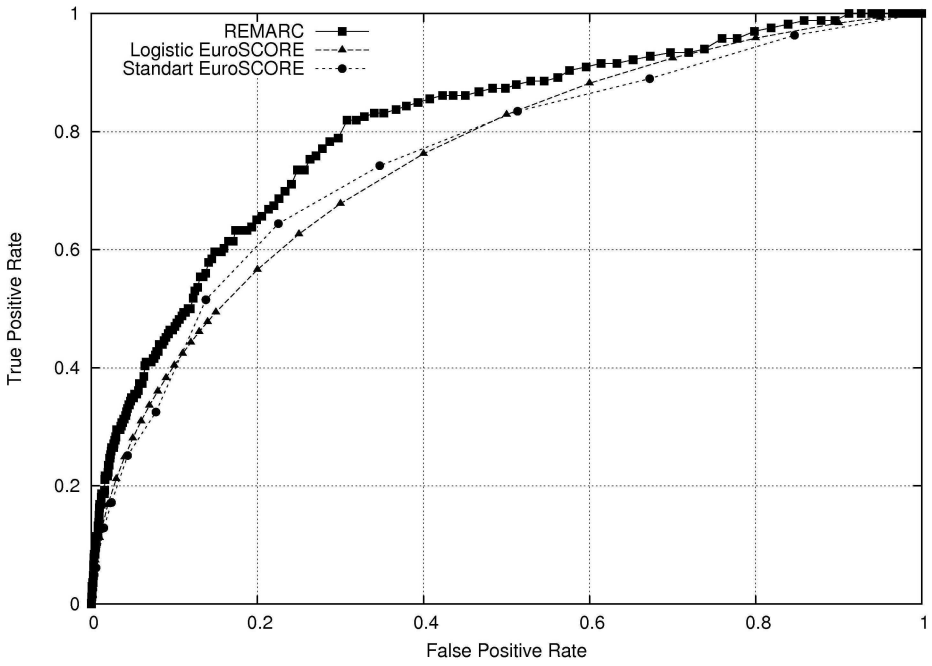

Figure 4.6 ROC curves of the REMARC method with TurkoSCORE risk factors

As a result, REMARC achieved higher performance with the TurkoSCORE risk factors when compared with the EuroSCORE factors (0.83 vs 0.79). REMARC also outperformed EuroSCORE Additive and Logistic on AUC basis (0.83 vs 0.77 on whole cohort).

In order to propose REMARC as a risk estimation system, the weights of risk factors and risk values are needed for each risk factor. Therefore, all instances in the selected dataset are used in the training phase to learn the weights and the rules. The weights of the risk factors are shown in Table B.3 of Appendix B. The knowledge learned by REMARC comprises the weights of features and the risk scores for each feature value, is shown in Table B.4 in the same appendix.

Chapter 5

Conclusion and Future Work

In this thesis, we gave a discussion of risk in real-life domains. Different risk domains are analyzed and some of the methods used specially in these domains are given. Then we showed how the risk estimation problem can be modeled as a two-class classification problem in machine learning.

We argued the effectiveness of a method that maximizes accuracy, for a risk estimation method. We proposed an AUC-based method instead of accuracy and presented important features of AUC, such as insensitivity to class distribution and error cost, as being statistically more consistent and discriminating. Then, we summarized the different methods proposed so far designed to maximize AUC.

Aiming to maximize AUC, we proposed a risk estimation method called REMARC. We have shown that for a categorical feature there is only one ordering that gives the maximum AUC. Then we showed the sufficient and necessary condition for a risk function to achieve this ordering. As a result, we proposed a risk function that finds the maximum possible AUC on one categorical feature. Aiming to maximize AUC, we handled the continuous

features using the MAD method, as it can discretize a continuous variable. Then we used these AUC values as weights in computing the risk scores as weighted averages of feature value risks. With this simple heuristic we averaged all feature risk values in order to achieve maximum AUC over the whole dataset.

We present the characteristics of the REMARC risk prediction model and how it should be interpreted. REMARC's prediction model is easy to understand and interpret by domain experts.

After supporting the theoretical background, we compared REMARC with 26 different algorithms. According to empirical evaluation, REMARC significantly outperformed 15 algorithms on an AUC basis and 13 algorithms on a time basis. It also outperformed all algorithms on the average AUC and 17 of them on an average time basis.

Cardiovascular surgery domain is selected as a test domain for REMARC. The data required are gathered from TurkoSCORE database. REMARC is compared one of the most popular cardiac surgery risk evaluation system, called EuroSCORE. Before this comparison, since EuroSCORE model is based on European cardiac patient population, demographic differences between the European and Turkish patients are investigated. Then, the validation of EuroSCORE model is performed on Turkish patient population. The calibration and discrimination of EuroSCORE model on Turkish cardiac patients are researched and it is shown that EuroSCORE model is not proper for Turkish patient population. Finally, EuroSCORE model and REMARC is compared. REMARC outperformed EuroSCORE on AUC basis by using the EuroSCORE risk factors on Turkish patient population. Then, the predictive performance of REMARC by using TurkoSCORE features is investigated. A REMARC based risk estimation system is proposed.

As a future work, REMARC can be compared with other risk methods and methods designed to maximize AUC. In order to improve the performance of REMARC, ensembling methods can be employed. Since there exists no validation dataset on TurkoSCORE database, the validation of the model is left as a future work. Since REMARC outperformed EuroSCORE even with EuroSCORE risk factors, the application of the REMARC method to European patient dataset, which is used in EuroSCORE project, can be an interesting future direction.

To conclude, a fast and highly predictive risk estimation method is proposed in this thesis. A simple yet effective predictive model, it is understandable by domain experts and will be useful for the machine learning community. The properties of this method are shown by applying it to the cardiovascular surgery domain.

BIBLIOGRAPHY

1 J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*; 143, 29-36, 1982.

2 B.O. Bradley and M.S. Taqqu. Handbook of Heavy-Tailed Distributions in Finance. In: Rachev, S.T. (Ed.), *Financial risk and heavy tails* (pp. 35–103). Rotterdam: Elsevier, 2003.

3 R. M. Conroy, K. Pyörälä, A. P. Fitzgerald *et al*. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *European Heart Journal*; 11, 987–1003, 2003.

4 R.B. D'Agostino, S.V. Ramachandran, J. Pencina, *et al*. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*; 17, 743–753, 2008.

5 K. Dowd and D. Blake. After VaR: The Theory, Estimation, and Insurance Applications of Quantile-Based Risk Measures. *The Journal of Risk and Insurance*, 73(2); 193-229, 2006

6 A. Giddens. Risk and Responsibility. *Modern Law Review*; doi:10.1111/1468-2230.00188, 1999

7 M. Sebag, J. Aze, N. Lucas. ROC-based evolutionary learning: Application to medical data mining. *Lecture Notes in Computer Sciences*; doi:10.1007/b100704, 2004.

8 D.J.M. Tax, R.P.W. Duin, Y. Arzhaeva. Linear model combining by optimizing the area under the roc curve. *Proceedings of the 18th IEEE International Conference on Pattern Recognition*; 119–122, 2006.

9 T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*; 27, 861–874, 2006.

10 A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*; 30(7), 1145-1159, 1997.

11 J Huang and C. X. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*; 17(3), 299–310, 2005.

12 K.A. Toh, J. Kim and S. Lee. Maximizing area under ROC curve for biometric scores fusion, *Pattern Recognition*; 41, 3373–3392, 2008.

13 A. Rakotomamonjy. Optimizing area under roc curve with svms. In *Workshop on ROC Analysis in Artificial Intelligence*; 71-80, 2004.

14 J S. Nashef, F. Roques, and P. Michel. European system for cardiac operative risk evaluation (euroscore). *European Journal of Cardio-Thoracic Surgery*; 16:9-13., 1999.

15 J. Pons, A. Granados, and J. Espinas. Assessing open heart surgery mortality in catalonia (spain) through a predictive risk model. *European Journal of Cardio-Thoracic Surgery*; 11:415-23., 1997.

16 F. Roques, S. Nashef, and P. Michel. Risk factors and outcome in European cardiac surgery: Analysis of the EuroSCORE multinational database of patients. *European Journal of Cardio-Thoracic Surgery*; 15:816-23., 1999.

17 J. Tu, S. Jaglal, and C. Naylor. Multicenter validation of a risk index for mortality, intensive care unit stay, and overall hospital length of stay after cardiac surgery. steering committee of the provincial adult cardiac care network of ontario. *Circulation*; 91:677-684., 1995.

18 C. Yap, C. Reid, M. Yii, M. A. Rowland, M. Mohajeri, P. D. Skillington, et al. Validation of the EuroSCORE model in Australia. *European Journal of Cardio-Thoracic Surgery*; 29:441-446., 2006.

19 Z. Zheng, Y. Li, S. Zhang, S. Hu on behalf of the Chinese CABG Registry Study. The Chinese Coronary Artery Bypass Grafting Registry Study: how well does the EuroSCORE predict operative risk for Chinese population?. *European Journal of Cardio-Thoracic Surgery*; doi:10.1016/j.ejcts.2008.08.001, 2009.

20 S. O. Hansson. Risk. Stanford Encyclopedia of Philosophy; Accessed 30 May. 2010, 2007.

21 I. T. Cameron and R. Raman. Process Systems Risk Management, Vol. 6. *Risk- Estimation, Presentation and Perception* (pp. 37-65) San Diego: Elsevier Science & Technology Books, 2005.

22 V. M. Shishkin and S. V. Savkov. The Method of Interval Estimation in Risk-Analysis System. *Proceedings of the 2nd international conference on Security of information and networks*; doi:10.1145/1626195.1626199, 2009.

23 A. Y. Huang. An optimization process in Value-at-Risk estimation. *Review of Financial Economics*; doi:10.1016/j.rfe.2010.03.001, In press, 2010.

24 S. V. Stoyanov, B. Racheva-Iotova, S. T. Rachev and F. J. Fabozzi. Stochastic Models for Risk Estimation in Volatile Markets: a Survey. *Annals of Operations Research*; doi: 10.1007/s10479-008-0468-1, 2008.

25 D. Ferrari and S. Paterlini. The maximum Lq-likelihood method: An application to extreme quantile estimation in finance. *Methodology and Computing in Applied Probability*; 11, 3–19, 2007.

26 F. Roques, P. Michel, A. Goldstone, and S. Nashef. The logistic EuroSCORE, *European Heart Journal*; doi: 10.1016/S0195-668X(02)00799-6, 2003.

27 E. Hannan, C. Wu, E. Bennett, R. Carlson, A. Culliford, et al. Risk stratification of in-hospital mortality for coronary artery bypass graft surgery.

Journal of the American Collage of Cardiology; doi:10.1016/j.jacc.2005.10.057, 2006.

28 D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(B), 187-220, 1972.

29 I. Colombet, A. Ruelland, G. Chatellier, F. Gueyffier, et al. Models to predict cardiovascular risk: comparison of cart, multilayer perceptron and logistic regression. *Proccedings of American Medical Informatics Association Symposium*; 156-160, 2000.

30 B. Biagioli, S. Scolletta, G. Cevenini, E. Barbini, et al. A multivariate bayesian model for assessing morbidity after coronary artery surgery. *Critical Care*; doi:10.1186/cc4951, 2006.

31 D. Gamberger, G. Krstacic and T. Smuc. Medical expert evaluation of machine learning results for a coronary heart disease database. *Lecture Notes on Computer Science*; 1933, 159-168, 2000.

32 J. Galindo and P. Tamayo. Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics*; doi:10.1023/A:1008699112516, 2000.

33 K-J. Kim. Financial time series forecasting using support vector machines. *Neurocomputing*; doi:10.1016/S0925-2312(03)00372-2, 2003.

34 J. H. Min and Y-C Lee. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*; doi:10.1016/j.eswa.2004.12.008, 2005.

35 W.J. Krzanowski, and D.J. Hand. *ROC curves for continuous data*; CRC Press, Taylor and Francis Group, 2009.

36 D.M. Green and J.A Swets. *Signal Detection Theory and Psychophysics*; Wiley, New York, 1966.

37 M.H. Zweig and G. Campbell. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*; 39(8), 561–577, 1993.

38 M.S. Pepe. The statistical evaluation of medical tests for classification and prediction. Oxford, New York, 2003.

39 K.A. Spackman. Signal detection theory: Valuable tools for evaluating inductive learning. In *Proceedings of the Sixth International Workshop on Machine Learning*; 160–163, 1989.

40 F. Provost and T. Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Knowledge Discovery and Data Mining*; 43–48, 1997.

41 F. Provost, T. Fawcett and R. Kohavi. The Case Against Accuracy Estimation for Comparing Induction Algorithms. In: J. Shavlik (Ed.). *Proceedings of the Fifteenth International Conference on Machine Learning*; 445–453, 1998.

42 P. Domingos. MetaCost: A general method for making classifiers cost-sensitive. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 155–164, 1999.

43 P. Flach and S. Wu. Repairing concavities in ROC curves. In *Procedings UK Workshop on Computational Intelligence*; 38–44, 2003.

44 S. Rosset. Model selection via the AUC. In: *Procedings of International Conference on Machine Learning*; 69, 89-96, 2004.

45 B. Mac Namee, P. Cunningham, S. Byrne, and O. Corrigan. The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*; 24, 51-70, 2002.

46 T. Fawcett and, F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*; 1, 291-316, 1997.

47 M.A. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In: Procedings International Conference on Machine Learning. Workshop on Learning from Imbalanced Data Sets II, 2003.

48 F. Provost, and T. Fawcett. Robust Classification for Imprecise Environments. *Machine Learning;* 42(3), 203–231, 2001.

49 C. Cortes, and M. Mohri AUC optimization vs. error rate minimization. *Neural Information Processing Systems (NIPS)*; 16, 313-320, 2003.

50 L. Yan, R. Dodier, M.C. Mozer and R, Wolniewicz. Optimizing Classifier Performance Via the Wilcoxon-Mann-Whitney Statistics. In: *Proceedings of the Twentieth International Conference on Machine Learning*; 848–855, 2003.

51 M. C. Mozer, R. Dodier, M. D. Colagrosso, C. Guerra-Salcedo, R. Wolniewicz. Prodding the ROC curve: Constrained optimization of classifier performance. In: *Advances in Neural Information Processing Systems*; 14, 1409–1415, 2002.

52 A. Herschtal and B. Raskutti. Optimising the area under the ROC curve using gradient descent. In: *Proccedings of International Conference on Machine Learning*; 49–56, 2004.

53 C. Ferri, P. Flach, J. Hernandez. Learning decision trees using the area under the ROC curve. In C. Sammut, and A. Hoffmann (Eds.), *Proceedings of the 19th International Conference on Machine Learning* (ICML-02); 139–146, 2002.

54 H. Boström. Maximizing the area under the ROC curve using incremental reduced error pruning. In: *ICML 2005 Workshop on ROC Analysis in Machine Learning,* 2005.

55 R. Prati and P. Flach. Roccer: A roc convex hull rule learning algorithm. In *Proccedings of the ECML/PKDD Workshop on Advances in Inductive Rule Learning*; 144-153, 2004.

56 T. Fawcett. Using Rule Sets to Maximize ROC Performance. *In: Proceedings of the IEEE International Conference on Data Mining* (ICDM-2001); 131–138, 2001.

57 C. Marrocco, M. Molinara, F. Tortorella. Exploiting AUC for optimal linear combinations of dichotomizers. *Pattern Recognition Letters*; 27(8), 900–907, 2006.

58 C. Marrocco, R.P.W. Duin, and, F. Tortorella. Maximizing the area under the ROC curve by pairwise feature combination. *Pattern Recognition*; 41, 1961–1974, 2008.

59 Y. Freund, R. Iyer, R. E. Schapire, and, Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*; 4, 933-969, 2003.

60 K. Ataman, W. N. Street and Y. Zhang. Learning to rank by maximizing AUC with linear programming. In *IEEE International Joint Conference on Neural Networks* (IJCNN); 123–129, 2006.

61 C.L. Ling and H. Zhang. Toward Bayesian classifiers with accurate probabilities. *Advances in Knowledge Discovery and Data Mining*; LNAI 2336, 123-134, 2002.

62 T. Calders, and S. Jaroszewicz. Efficient AUC Optimization for Classification. In *Knowledge Discovery in Databases*: PKDD; 42-53, 2007.

63 G. Han and C. Zhao. AUC maximization linear classifier based on active learning and its application. *Neurocomputing*; doi:10.1016/j.neucom.2010.01.001, 2010.

64 H. Liu, F. Hussain, C.L. Tan, M. Dash. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*; 6(4):393–423, 2002.

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

66 J.R. Quinlan. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

67 J.R. Quinlan. Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence and Research*; 4:77-90, 1996.

68 J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

69 J. Catlett. On changing continuous attributes into ordered discrete attributes. In *Proceedings of the Fifth European Working Session on Learning*. Berlin: Springer-Verlag; 164–177, 1991.

70 U. Fayyad and K. Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*; 8:87–102, 1992.

71 T. Van de Merckt. Decision trees in numerical attribute spaces. *Machine Learning*; 1016–1021, 1990.

72 J, Cerquides and R.L. Mantaras. Proposal and empirical comparison of a parallelizable distance-based discretization method. In *KDD97: Third International Conference on Knowledge Discovery and Data Mining*; 139–142, 1997.

73 K.M. Ho and P.D. Scott. Zeta: A global method for discretization of continuous variables. In KDD97: *3rd International Conference of Knowledge Discovery and Data Mining*. Newport Beach; 191–194, 1997.

74 C.C. Chan, C. Batur, A. Srinivasan. Determination of quantization intervals in rule based model for dynamic. In *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*. Charlottesvile, Virginia; 1719–1723, 1991.

75 R. Kerber. Chimerge: Discretization of numeric attributes. In Proc. AAAI-92, *Ninth National Confrerence Articial Intelligence*. AAAI Press/The MIT Press; 123–128, 1992.

76 H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence*. Herndon, Virginia; 388–391, 1995.

77 K. Wang and B. Liu. Concurrent discretization of multiple attributes. In *Pacific Rim International Conference on AI*; 250–259, 1998.

78 M. Kurtcephe, and H.A. Guvenir. A Discretization Method based on Maximizing the Area Under ROC Curve. Technical Report. No: BU-CE-1001. Bilkent University. http://www.cs.bilkent.edu.tr/tech-reports/2010/BU-CE-1001.pdf. Accessed 14 June 2010, 2010.

79 W.W. Cohen, R.E. Schapire and Y. Singer. Learning to order things. In *Advances in Neural Information Processing Systems*;10, 243-270, 1998.

80 A. Frank and A. Asuncion. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science; http://archive.ics.uci.edu/ml, 2010.

81 M. Hall *et al*. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*; Volume 11, Issue 1, 2009.

82 C.-C. Chang and C.-C. Lin. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

83 A. Tunca. Predicting Risk of Mortality in Patients Undergoing Cardiovascular Surgery. Thesis. Bilkent University Computer Engineering Department. 2008. http://www.cs.bilkent.edu.tr/tech-reports/2008/BU-CE-0807.pdf. Accessed 01 July 2010

84 F. Roques, S.A.M. Nashef, P. Michel, E. Gauducheau, C. de Vincentiis, E. Baudet, J. Cortina, M. David, A. Faichney, F. Gavrielle, E. Gams, A. Harjula, M.T. Jones, P. Pinna Pintor, R. Salamon, L. Thulin. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. *European Journal of Cardio-Thoracic Surgery*; 15:816-822., 1999.

Appendix A

EuroSCORE

Risk Factor in EuroSCORE	Approximation in TurkoSCORE
Age	Yaş
Sex	Cinsiyet
Chronic pulmonary disease	Kronik Obstrüktif Akciğer Hastalığı (KOAH) (Orta KOAH, Ciddi KOAH)
Extracardiac arteriopathy	Periferik Arter Hastalığı
Neurological dysfunction	Serebrovasküler Hastalık (CVA, Non-invaziv karotis incelemesinde çapta %70'den fazla daralma)
Previous cardiac surgery	Operasyon insidansı (Redo kardiyak cerrahi)
Serum creatinine	Son Preoperatif Kreatinin Düzeyi (Kreatinin 1.5 - 2.26 mg/dL, Kreatinin > 2.26 mg/dL)
Active endocarditis	İnfektif Endokardit (Pozitif kan kültürü ile infektif endokardit öntanısı, Ekokardiyografide vejetasyon veya görüntüleme yöntemleri ile endokardit öntanısı, Prostetik kapak endokarditi)
Critical preoperative state	Kritik Preoperatif Durum (VT / VF,Preoperatif Resüsitasyon, Preoperatif IABP, Preoperatif respirator, Preoperatif akut renal yetmezlik, Preoperatif inotrop gereksinimi)
Unstable angina	Unstabil Angina (CCS4C, CCS4D) veya Ameliyat öncesi stabil Angina Pektoris (CCS4)
LV dysfunction	Moderate: Sol Ventrikül Ejeksiyon fraksiyonu kategorik (Grade II)
	Poor: Sol Ventrikül Ejeksiyon fraksiyonu kategorik (Grade III veya Grade IV)
Recent myocardial infarct	Operasyon öncesi geçirilmiş MI (< 3 ay)
Pulmonary hypertension	Sistolik pulmoner arter basıncı değeri (mmHg) > 60
Emergency	Operasyon Önceliği (Acil veya Salvaj)
Other than isolated CABG	Koroner arter bypass cerrahisi dışında veya alternatif olarak yapılan ameliyatlar: Kapak Cerrahisi, Kalp Nakli Kardiyak Tümör, Sol Ventrikül Anevrizma Onarımı, Batista operasyonu, Sol Ventrikül Restorasyon, Kök Hücre İmplantasyonu, Transmiyokardiyal laser revaskülarizasyon Atrial septal defekt (ASD) Onarımı, Konjenital diğer defektlerin onarımı Aritmi cerrahisi Radyo-Frekans veya microwave Ablasyon, Kardiyak kist hidatik eksizyonu
Surgery on thoracic aorta	Aort Cerrahisi
Postinfarct septal rupture	Ventriküler septal defekt/rüptür Onarımı

Table A.1 TurkoSCORE approximations

Appendix B

TurkoSCORE

EuroScore Risk Factor	Feature AUC
Age	0.674
Sex	0.553
Chronic pulmonary disease	0.581
Extracardiac arteriopathy	0.542
Neurological dysfunction	0.558
Previous cardiac surgery	0.554
Serum creatinine	0.606
Active endocarditis	0.542
Critical preoperative state	0.640
Unstable angina	0.578
LV dysfunction	0.607
Recent myocardial infarct	0.617
Pulmonary hypertension	0.740
Emergency	0.640
Other than isolated CABG	0.507
Surgery on thoracic aorta	0.543
Postinfarct septal rupture	0.503

Table B.1 AUC values of EuroSCORE risk factors

#	Risk Factor	AUC of Risk Factor	Weight
1	Ameliyat öncesi dispne (NYHA klasifikasyonuna göre)	0.688	0.376
2	Yaş	0.668	0.336
3	Operasyon Önceliği	0.634	0.268
4	Sol Ventrikül Ejeksiyon fraksiyonu kategorik	0.612	0.224
5	Son Preoperatif Kreatinin Düzeyi	0.610	0.220
6	Ameliyat öncesi stabil Angina Pektoris	0.599	0.198
7	BMI	0.588	0.176
8	Konjestif Kalp Yetmezliği	0.580	0.160
9	Kronik Obstrüktif Akciğer Hastalığı (KOAH GOLD Sınıflaması)	0.578	0.156
10	Diabetes Mellitüs	0.562	0.124
11	Hemodinamik Status	0.561	0.122
12	Preoperatif inotrop gereksinimi bulunması	0.558	0.116
13	Serebrovasküler Hastalık	0.558	0.116
14	Cinsiyet	0.555	0.110
15	Operasyon insidansı (Redo kardiyak cerrahi)	0.555	0.110
16	Geçirilmiş MI Sayısı	0.548	0.096
17	Ritm statusu	0.548	0.096
18	Aort Cerrahisi olacak	0.546	0.092
19	Renal Yetmezlik	0.540	0.080
20	Kapak Cerrahisi olacak	0.540	0.080
21	Periferik Arter Hastalığı (Serebrovasküler hastalık hariç)	0.539	0.078
22	Preoperatif akut renal yetmezlik (anürü veya oligüri, 10 ml/saat)	0.530	0.060
23	Koroner Bypass Cerrahisi Olacak	0.526	0.052
24	Preoperatif IABP takılmış olması	0.524	0.048
25	Aritmi cerrahisi Radyo-Frekans veya microwave Ablasyon olacak	0.522	0.044
26	Sol Ventrikül Anevrizma Onarımı olacak	0.514	0.028
27	Hipertansiyon kategorik	0.514	0.028
28	Karotis Cerrahisi olacak	0.511	0.022

Table B.2 TurkoSCORE selected features and AUC values of each feature. AUC values are calculated by using ten-fold cross validation

#	Risk Factor	AUC of Risk Factor	Weight
1	AmeliyatOncesiDispne	0.729	0.457
2	Yas	0.683	0.366
3	Onceligi	0.637	0.273
4	Ameliyat öncesi stabil Angina Pektoris	0.626	0.252
5	Sol Ventrikül Ejeksiyon fraksiyonu kategorik	0.617	0.234
6	Kronik Obstrüktif Akciğer Hastalığı (KOAH GOLD Sınıflaması)	0.610	0.220
7	BMI	0.605	0.210
8	Konjestif Kalp Yetmezliği	0.604	0.209
9	Son Preoperatif Kreatinin Düzeyi	0.597	0.193
10	Preoperatif inotrop gereksinimi bulunması	0.582	0.163
11	Serebrovasküler Hastalık	0.570	0.139
12	Ritm statusu	0.560	0.120
13	Operasyon insidansı (Redo kardiyak cerrahi)	0.557	0.114
14	Diabetes Mellitüs	0.554	0.107
15	Aort Cerrahisi olacak	0.552	0.104
16	Hemodinamik Status	0.550	0.099
17	Periferik Arter Hastalığı (Serebrovasküler hastalık hariç)	0.545	0.089
18	Kapak Cerrahisi olacak	0.545	0.089
19	Cinsiyet	0.540	0.079
20	Renal Yetmezlik	0.533	0.067
21	Geçirilmiş MI Sayısı	0.530	0.060
22	Koroner Bypass Cerrahisi Olacak	0.530	0.059
23	Preoperatif IABP takılmış olması	0.526	0.051
24	Preoperatif akut renal yetmezlik (anürü veya oligüri, 10 ml/saat)	0.525	0.051
25	Aritmi cerrahisi Radyo-Frekans veya microwave Ablasyon olacak	0.525	0.050
26	Hipertansiyon kategorik	0.515	0.030
27	Karotis Cerrahisi olacak	0.513	0.026
28	Sol Ventrikül Anevrizma Onarımı olacak	0.512	0.025

Table B.3 TurkoSCORE selected features and AUC values of each feature. AUC values are calculated by using whole dataset as training set (6871 patients)

Knowledge Learned By REMARC:

Cinsiyet: AUC=0,540 Weight=0,079
K: Risk=0,0236, #cases=1953
E: Risk=0,0165, #cases=4914

Yas: AUC=0,683 Weight=0,366
88.5..90.0: Risk=1,0000, #cases=1
1.0..15.0: Risk=0,0769, #cases=26
<1.0: Risk=0,0769, #cases=13
79.5..88.5: Risk=0,0638, #cases=94
78.5..79.5: Risk=0,0408, #cases=49
76.5..78.5: Risk=0,0376, #cases=133
68.5..76.5: Risk=0,0348, #cases=1292
67.5..68.5: Risk=0,0242, #cases=207
66.5..67.5: Risk=0,0186, #cases=215
63.5..66.5: Risk=0,0170, #cases=706
58.5..63.5: Risk=0,0166, #cases=1148
55.5..58.5: Risk=0,0100, #cases=698
15.0..55.5: Risk=0,0071, #cases=2264
90.0<: Risk=0,0000, #cases=3

BMI: AUC=0,605 Weight=0,210

15.41631..16.014544: Risk=1,0000, #cases=3

16.014544..20.173252: Risk=0,0375, #cases=160

20.173252..23.120625: Risk=0,0347, #cases=663

37.912354..53.550346: Risk=0,0336, #cases=119

23.120625..25.23634: Risk=0,0172, #cases=1049

25.23634..26.511805: Risk=0,0166, #cases=845

26.511805..28.3771: Risk=0,0162, #cases=1295

28.3771..29.6875: Risk=0,0148, #cases=743

29.6875..37.912354: Risk=0,0121, #cases=1811

53.550346<: Risk=0,0000, #cases=23

<15.41631: Risk=0,0000, #cases=11

AmeliyatOncesiAnjinaPektoris: AUC=0,626 Weight=0,252

4: Risk=0,0365, #cases=631

0: Risk=0,0326, #cases=429

1: Risk=0,0226, #cases=1415

3: Risk=0,0164, #cases=1521

2: Risk=0,0105, #cases=2844

AmeliyatOncesiDispne: AUC=0,729 Weight=0,457

4: Risk=0,1190, #cases=168

3: Risk=0,0448, #cases=1005

2: Risk=0,0166, #cases=2654

1: Risk=0,0061, #cases=2951

KojestifKalpYetmezligi: AUC=0,604 Weight=0,209

1: Risk=0,0851, #cases=388

0: Risk=0,0145, #cases=6394

HemodinamikStatus: AUC=0,550 Weight=0,099

3: Risk=0,2500, #cases=16

2: Risk=0,1778, #cases=45

1: Risk=0,0175, #cases=5657

PreStatus3: AUC=0,526 Weight=0,051

TRUE: Risk=0,2059, #cases=34

FALSE: Risk=0,0176, #cases=6837

PreStatus5: AUC=0,525 Weight=0,051

TRUE: Risk=0,0889, #cases=90

FALSE: Risk=0,0175, #cases=6781

PreStatus6: AUC=0,582 Weight=0,163

TRUE: Risk=0,0618, #cases=469

FALSE: Risk=0,0153, #cases=6402

DM: AUC=0,554 Weight=0,107

5: Risk=0,1333, #cases=15

4: Risk=0,0257, #cases=505

2: Risk=0,0220, #cases=318

0: Risk=0,0189, #cases=4187

3: Risk=0,0123, #cases=1221

1: Risk=0,0000, #cases=6

HipertansiyonHikayesi: AUC=0,515 Weight=0,030

2: Risk=0,1538, #cases=13

1: Risk=0,0188, #cases=3246

0: Risk=0,0176, #cases=3577

KOAH: AUC=0,610 Weight=0,220

5: Risk=0,2857, #cases=7

2: Risk=0,0938, #cases=64

4: Risk=0,0714, #cases=14

1: Risk=0,0577, #cases=104

3: Risk=0,0356, #cases=872

0: Risk=0,0140, #cases=5787

RenalYetmezlik: AUC=0,533 Weight=0,067

2: Risk=0,1333, #cases=45

1: Risk=0,1000, #cases=30

0: Risk=0,0163, #cases=6668

3: Risk=0,0000, #cases=1

SonPreopKreatinin: AUC=0,597 Weight=0,193

3: Risk=0,1136, #cases=88

2: Risk=0,0549, #cases=164

1: Risk=0,0404, #cases=371

0: Risk=0,0157, #cases=5336

PeriferikArterHastalik: AUC=0,545 Weight=0,089

4: Risk=0,1667, #cases=6

6: Risk=0,0480, #cases=333

1: Risk=0,0174, #cases=115

0: Risk=0,0169, #cases=6285

3: Risk=0,0000, #cases=1

5: Risk=0,0000, #cases=5

2: Risk=0,0000, #cases=3

SerebrovaskulerHastalik: AUC=0,570 Weight=0,139

4: Risk=0,1667, #cases=12

6: Risk=0,1538, #cases=26

1: Risk=0,1111, #cases=9

2: Risk=0,0474, #cases=274

3: Risk=0,0333, #cases=60

7: Risk=0,0280, #cases=143

0: Risk=0,0157, #cases=6230

5: Risk=0,0000, #cases=3

MISayisi: AUC=0,530 Weight=0,060

3: Risk=0,0545, #cases=55

2: Risk=0,0222, #cases=405

1: Risk=0,0199, #cases=2061

0: Risk=0,0170, #cases=4175

RitmStatus: AUC=0,560 Weight=0,120

6: Risk=0,3333, #cases=3

2: Risk=0,3333, #cases=3

1: Risk=0,0428, #cases=421

0: Risk=0,0153, #cases=5821

3: Risk=0,0000, #cases=3

7: Risk=0,0000, #cases=5

4: Risk=0,0000, #cases=1

5: Risk=0,0000, #cases=1

SolVentrikulEjeksiyonFraksiyonu: AUC=0,617 Weight=0,234

4: Risk=0,2000, #cases=5

3: Risk=0,0472, #cases=318

0: Risk=0,0460, #cases=239

2: Risk=0,0205, #cases=1950

1: Risk=0,0123, #cases=4143

Onceligi: AUC=0,637 Weight=0,273

3: Risk=0,1667, #cases=36

2: Risk=0,0664, #cases=256

1: Risk=0,0459, #cases=567

0: Risk=0,0130, #cases=6005

Insidans: AUC=0,557 Weight=0,114

5: Risk=0,5000, #cases=2

3: Risk=0,0959, #cases=73

2: Risk=0,0521, #cases=211

1: Risk=0,0164, #cases=6534

4: Risk=0,0000, #cases=14

OperasyonGrup0: AUC=0,530 Weight=0,059

TRUE: Risk=0,0294, #cases=681

FALSE: Risk=0,0173, #cases=6190

OperasyonGrup1: AUC=0,545 Weight=0,089

TRUE: Risk=0,0606, #cases=264

FALSE: Risk=0,0168, #cases=6607

OperasyonGrup3: AUC=0,552 Weight=0,104

TRUE: Risk=0,0667, #cases=270

FALSE: Risk=0,0165, #cases=6601

OperasyonGrup4: AUC=0,513 Weight=0,026

TRUE: Risk=0,1000, #cases=40

FALSE: Risk=0,0180, #cases=6831

KardiakProsedur1: AUC=0,512 Weight=0,025

TRUE: Risk=0,0476, #cases=105

FALSE: Risk=0,0180, #cases=6766

KardiakProsedur9: AUC=0,525 Weight=0,050

TRUE: Risk=0,0842, #cases=95

FALSE: Risk=0,0176, #cases=6776

Table B.4 Knowledge learned by using REMARC on TurkoSCORE dataset

Publications Produced in This Thesis

1 M. Kurtcephe, and H.A. Guvenir. A Discretization Method based on Maximizing the Area Under ROC Curve. Technical Report. No: BU-CE-1001. Bilkent University. http://www.cs.bilkent.edu.tr/tech-reports/2010/BU-CE-1001.pdf. Accessed 14 June 2010, 2010. Submitted to *Knowledge-Based Systems*.

2 M. Kurtcephe, and H.A. Guvenir. Risk Estimation by Maximizing the Area under ROC Curve. No: BU-CE-1003. Bilkent University. http://www.cs.bilkent.edu.tr/tech-reports/2010/BU-CE-1003.pdf. Accessed 12 July 2010, 2010. Submitted to *Machine Learning*.

3 A. R. Akar, M. Kurtcephe, S. Durdu, E. Sener, C. Alhan, A. Kunt, H. A. Güvenir, The Working Group of the Turkish Society of Cardiovascular Surgery. Validation of the EuroSCORE Risk Model in Turkish Adult Cardiac Surgery. Prepared for the *European Journal of Cardio-Thoracic Surgery*.