STATISTICAL COMPARISON OF CLASSIFIERS USING RECEIVER OPERATING CHARACTERISTICS INFORMATION

by

Özlem Aslan

Bachelor of Science, Computer Science and Engineering, Işık University, 2007

Bachelor of Science, Industrial Engineering, Işık University, 2007

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2009

ACKNOWLEDGEMENTS

I'm so grateful to my thesis supervisor Ethem Alpaydın for his valuable guidance, for answering my questions with patience during the thesis and also for the background that I've earned during his courses and during the thesis period. I feel so lucky to have a chance to work with him who will be a role model for me in my future academical study.

I'd like to thank to Olcay Taner Yıldız for his contributions to my thesis, for using his Machine Learning Library Software ISELL and for being my thesis examiner. I'd like to thank Fikret Gürgen for being my thesis examiner.

For the efficient and nice working environment in PIlab, I'd like to thank my friends, Oya Aran, İsmail Arı, Onur Dikmen, Mehmet Gönen and Mehmet Aydın Ulaş. I'm also thankful to my friend Levent Ünver for printing and submitting my thesis.

I'd like to thank to Cesim Erten who made me join a research project in my undergraduate thesis. Before graduation, I became certain about my decision to have an academical career after such a good experience.

I'm so grateful to my spiritual sister Neslihan for being by my side, understanding me and encouraging me all the time. Because I know that my dear parents deserve much more than any words that I could say, it is so hard to express my gratitude and my feelings to them but without a doubt, any of my accomplishments, if I have, are also their accomplishments.

ABSTRACT

STATISTICAL COMPARISON OF CLASSIFIERS USING RECEIVER OPERATING CHARACTERISTICS INFORMATION

Statistical tests in the literature mainly use error rate for comparison and assume equal loss for false positives and negatives. Receiver Operating Characteristics (ROC) curves and/or the Area Under the ROC Curve (AUC) can also be used for comparing classifier performances under a spectrum of loss values. A ROC curve and hence an AUC value is typically calculated from one training/test pair and to average over randomness in folds, we propose to use k-fold cross-validation to generate a set of ROC curves and AUC values to which we can fit a distribution and test hypotheses on. Experiment results on 15 datasets using 5 different classification algorithms show that our proposed test using AUC values is to be preferred over the usual paired t test on error rate because it can detect equivalences and differences which the error test cannot.

The approach we use for ROC curves can also be applied to Precision-Recall curves, used mostly in information retrieval by applying k-fold cross-validated test on the area under the Precision-Recall curve.

When multiple classifiers are to be compared over one dataset or multiple datasets, we can use Analysis of Variance (ANOVA). When we use more than one performance metric, we use the multivariate ANOVA, that is, MANOVA. Performance metrics of ANOVA is error or AUC. Performance metrics of MANOVA are true positive, false positive, true negative and false negative rates. We also perform the nonparametric version of ANOVA which is called Friedman test. We apply Sign test when we compare multiple classifiers over multiple datasets. We observe that using more than one per-

formance metric includes their correlation in the statistical test and therefore produces more accurate results.

ÖZET

SINIFLANDIRICILARIN ROC BİLGİSİ KULLANARAK İSTATİSTİKSEL KARŞILAŞTIRILMASI

Literatürdeki istatistiksel testler genelde hata oranını kullanırlar ve yanlış pozitif and yanlış negatiflerin maliyetlerinin aynı olduğunu varsayarlar. ROC eğrileri ve/veya ROC Eğrilerinin Altındaki Alan (AUC), çeşitli maliyet değerlerine göre sınıflandırıcıların performanslarını karşılaştırmak için kullanılabilir. Bir ROC eğrisi ve bir ROC eğrisinin altındaki alan genellikle bir öğrenme/sınama çiftinden hesaplanır ve verideki rastsallığın ortalamasını almak için ve dağılım oturtabileceğimiz ve üzerinde hipotez testi yapabileceğimiz bir ROC eğrileri kümesi ve AUC değerleri oluşturmayı öneriyoruz. 15 veri kümesi üzerinde 5 farklı sınıflandırma algoritması kullanılarak bulduğumuz deneysel sonuçlar gösteriyor ki bizim önerdiğimiz AUC testi hata oranını kullanan eşli t testine göre daha üstündür çünkü AUC testi hata testinin fark edemeyeceği eşitlik ve farklılıkları fark edebiliyor. ROC eğrileri için kullandığımız yaklaşım, Doğruluk-Anımsama eğrilerinin altında kalan alana k-kat çapraz-geçerleme uygulayarak da kullanılabilir.

Birden çok sınıflandırıcıyı bir veri kümesi veya birden çok veri kümesi üzerinde karşılaştırımak için Varyans Analizi (ANOVA) kullanabiliriz. Birden çok performans metriği üzerinden karşılaştırma yapmak için, çok değişkenli ANOVA, MANOVA, kullanırız. ANOVA'nın performans metrikleri hata veya AUC olabilir. MANOVA'nın performans metrikleri doğru pozitif, yanlış pozitif, doğru negatif ve yanlış negatif değerleridir. ANOVA'nın parametrik olmayan versiyonu olan Friedman testini de yapıyoruz. Çoklu sınıflandırıcıları çoklu veri kümeleri üzerinden karşılaştırırken İşaret testi uyguluyoruz. Birden çok performans metriği kullanmanın onların korelasyonlarını içerdiğini ve bu yüzden daha güvenilir sonuçlar ürettiğini gözlemliyoruz.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Comparing the performances of classifiers is a critical problem in machine learning. In the literature, to compare the generalization error of learning algorithms, statistical tests have been proposed [1], [2]. In choosing between two learning algorithms, one can use a pairwise test to compare their generalization error and select the one that has lower error. Typically, cross-validation is used to generate a set of training, validation folds and one compare the expected error on the validation folds after training on the training folds. Examples of such tests are parametric tests, such as k-fold paired t test, 5×2 cv t test [1], 5×2 cv F test [3], nonparametric tests, such as the sign test and Friedman's test, or range tests, such as Wilcoxon signed rank test [2].

If we define the class labels of the two-class classification problem as positive and negative, the confusion matrix shown in Table 1.1 contains the following items:

- True positive (TP): If both the class label and the predicted class are positive.
- False negative (FN): If the class label is positive and the predicted class is negative.
- False positive (FP): If the class label is negative and the predicted class is positive.
- True negative (TN): If both the class label and the predicted class are negative.

Different metrics calculated from these values are used in the literature are:

Table 1.1. Confusion matrix

	Predicted Class		
True Class	Positive	Negative	Total
Positive	TP	FN	P
Negative	FP	TN	N

$$\textit{hit rate} = \frac{TP}{TP+FN} \tag{1.1}$$

$$\textit{false alarm rate} = \frac{FP}{TN+FP} \tag{1.2}$$

$$\textit{error rate} = \frac{FP+FN}{TP+TN+FP+FN} \tag{1.3}$$

$$\textit{precision} = \frac{TP}{TP+FP} \tag{1.4}$$

$$\textit{accuracy} = \frac{TP+TN}{P+N} \tag{1.5}$$

$$\textit{recall} = \frac{TP}{P} \tag{1.6}$$

where P is the total number of positives and N is the total number of negatives.

Generally, error rate (or accuracy) is the most frequently used performance metric for classifiers. However, it is a poor performance metric since it assumes equal misclassification cost all classes. In real world problems, different misclassifications have different costs and generally, the most critical classes have fewer instances. For example, a false positive and false negative in a patient classification in case of a terminal illness should not be assumed to have equal cost. In signal detection, people have been using *Receiver Operating Characteristic* (ROC) curves to observe the relationship between the hit rate and the false alarm rate [4]. ROC curves do not make any assumption on class priors or misclassification costs and this makes them interesting in the field of machine learning [5].

Although ROC curves are very useful for visualizing error rates for different misclassification costs, one may need to summarize the ROC curve for comparing performance of classifiers by a single number and for this *Area Under the ROC curve* (AUC) is used. ROC curve is constructed for a single test set and the trapezoidal area under the ROC curve is calculated. AUC has been related to the *Wilcoxon* statistic and it has been defined as an estimate of the 'true' area under the ROC curve, that is, the area constructed from an infinite sample [6]. However, ROC and AUC use a single training and testing pair [7], [8], [9] and this makes the AUC dependent on the test test that it is used. In this thesis, we extend this idea and use k-fold cross-validation to generate k ROC curves hence k AUC values. Then we fit a distribution to the set of AUC values and test our hypothesis on these distributions. Fitting distribution to AUC values is also used by Bravo et al [10]. However, they do not compare it with error metric, they just use it to evaluate their results. Confidence intervals for AUC are also proposed [11], [12]. The effect of class distribution on error and AUC is also experimented in [13].

Similar to ROC curves, *Precision-Recall curves* are also used [14], mostly in Information Retrieval [15], and they are preferred to ROC curves when there is a large skew of class distribution [16], [17], [18], [19]. Precision is the y-axis and recall is the x-axis. We also calculate the area under the Precision-Recall curve like the area under the ROC curve, using the trapezoidal area method. The *area under Precision-Recall curve* will be called as AUC-PR.

When more than two classifiers are to be compared, the paired t test is not applicable. Analysis of Variance (ANOVA) is the appropriate method for this problem [20]. ANOVA is a parametric test that has many assumptions; Friedman test is the nonparametric version of ANOVA that can be used when the assumptions of ANOVA are not met [2]. We also want to compare classifiers using the values of true positives, false positives,true negatives and false negatives. Since the data is four dimensional instead one, we propose to apply Multivariate ANOVA (MANOVA) [21]. If the number of wins, losses and ties of classifier pairs are known, Binomial Sign test can be used to test whether there is a significant difference between their performances or not [2],

[27].

This thesis is organized as follows: ROC curves and Precision-Recall curves are introduced in Chapter 2. Our proposed paired t test on AUC values and on AUC-PR values, experiments and results are given in Chapter 3. We introduce the methods for the comparison of multiple classifiers in Chapter 4. We give experimental setup and report results for comparison of multiple classifiers over one dataset in Chapter 5, for comparison of univariate comparison of multiple classifiers over multiple datasets in Chapter 6, for comparison of multivariate comparison of multiple classifiers over multiple datasets for different threshold points in Chapter 7. We conclude and discuss future work in Chapter 8.

2. THE ROC AND PRECISION RECALL CURVES AND THE AREA UNDER THEM

2.1. RECEIVER OPERATING CHARACTERISTICS

The hit rate or true positive rate (TPR) defines the y axis and the false alarm rate or the false positive rate (FPR) defines the x axis in a ROC space. The point (0, 0) in the ROC space corresponds to a classifier that labels all instances as negative. The point (1, 1) means the opposite, that is, the classifier that labels all instances as positive. The point (0, 1) means 100 percent classification accuracy, whereas (1, 0) points means 0 percent classification accuracy. Being in the lower-left region in the ROC space means being 'conservative', that is, the classifier has less tendency for assigning the positive class label. This corresponds to a classifier that has low false alarm rate and low hit rate. Conversely, being in the upper-right region means being 'liberal', that is, the classifier has much more tendency for assigning the positive class label. This is a classifier with high false alarm rate and high hit rate.

The ROC curve is used in signal processing to plot the trade-off between the hit rate and the false alarm rate. It allows visualization of performance for a set of conditions instead of just the misclassification error. A classifier is good if it has a high hit rate and a low false alarm rate, that is, if the curve is closer to the upper left corner. Two examples for the ROC curves are shown in Figure 2.1. The diagonal line indicates the curve for random prediction.

In a two-class problem, if the posterior probability of the positive class is greater than the posterior probability of the negative class, the classifier predicts the class label of a test instance as positive, otherwise it predicts the class label as negative. This is equivalent to checking if the posterior probability of the positive class is greater than the *threshold value* of $\theta = 0.5$. For a given test set, a ROC curve is constructed in ROC space by plotting the hit rates on y axis and the false alarm rates in x axis for different threshold values. The ROC curve construction algorithm is given in Figure 2.2 [4]. In

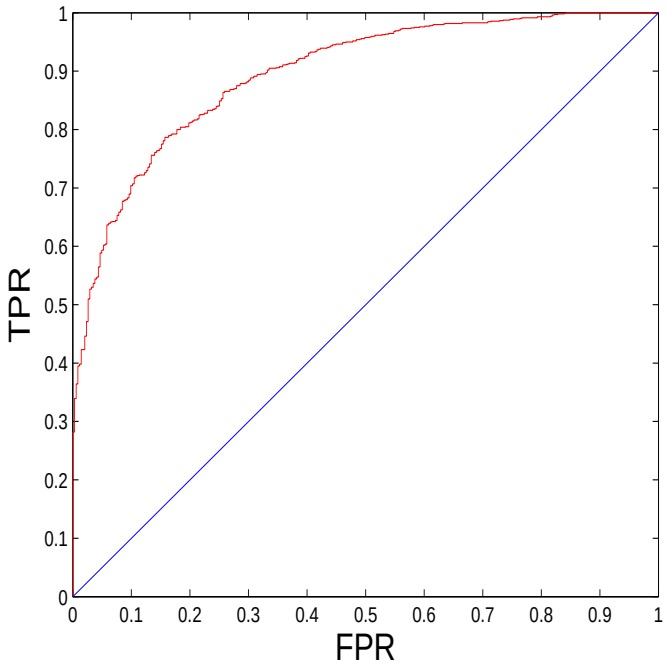


Figure 2.1. Example ROC curve

Line 4, the test set is sorted in decreasing order of posterior probabilities of the positive class. In Line 5, FP and TP are initialized to zero. In Line 6, the list of ROC points R is set to empty list. In Line 9, the loop that traverses all the instances in the test set starts. At each iteration, the posterior probability of the instance is taken as a threshold point. Therefore, the 'if' statement in Line 10 checks the condition of equal posterior probabilities to prevent taking equal posterior probabilities as threshold points. If the posterior probability of the instance is not equal to the posterior probability of the previous instance in the list, then in Line 11, the TP and FP points are added to the list R and in Line 12, the current posterior probability $f(i)$ is assigned to f_{prev}. In Line 14, the 'if' statement checks whether the class of the current instance is positive or not. If the class of the instance is positive, then TP is incremented in Line 15 since the posterior probabilities of the positive instances in the beginning part of the list is greater than the current threshold point and they are correctly classified. If the class of the instance is negative, then the FP is incremented in Line 17 since the posterior probabilities of the negative instances in the beginning part of the list is greater than the current threshold point and they are classified incorrectly. Hence, for a given test set, the algorithm takes the posterior probabilities of the test set as different threshold

1: **Input:** L, the set of test examples; f(i), the posterior probability that example i is positive; P and N, the number of positive and negative examples.
2: **Output:** R, a list of ROC points in increasing order of fp rate.
3: **Require:** $P > 0$ and $N > 0$.
4: $L_{sorted} \leftarrow L$ sort in decreasing order of f values
5: $FP \leftarrow TP \leftarrow 0$
6: $R \leftarrow \langle\rangle$
7: $f_{prev} \leftarrow -\infty$
8: $i \leftarrow 1$
9: **while** $i \leq |L_{sorted}|$ **do**
10: **if** $f(i) \neq f_{prev}$ **then**
11: push$\left(\frac{FP}{N}, \frac{TP}{P}\right)$ onto R
12: $f_{prev} \leftarrow f(i)$
13: **end if**
14: **if** $L_{sorted}[i]$ is a positive example **then**
15: TP $\leftarrow$ TP+1
16: **else**
17: FP $\leftarrow$ FP+1
18: **end if**
19: $i = i + 1$
20: **end while**
21: push $\left(\frac{FP}{N}, \frac{TP}{P}\right)$ onto R

Figure 2.2. ROC Curve construction [4]

points rather than sweeping the same threshold points for all test sets. It also prevents traversing threshold points needlessly. The time complexity of the algorithm is $O(n \log n)$ for a test set with size n [4].

Let us define the positive class as class C_1 and the negative class as C_2. For a test instance x, given the loss matrix in Table 2.1, the risk of choosing C_1 is:

Table 2.1. Loss matrix

	C_1	C_2
C_1	0	λ
C_2	1	0

$$\begin{aligned} R(C_1|x) &= \lambda_{11}P(C_1|x) + \lambda_{12}P(C_2|x) && (2.1)\\ &= \lambda P(C_2|x) && (2.2) \end{aligned}$$

and the risk of choosing C_2 is:

$$R(C_2|x) = P(C_1|x) \tag{2.3}$$

Then we choose C_1 if

$$R(C_1|x) < R(C_2|x) \tag{2.4}$$

$$\lambda P(C_2|x) < P(C_1|x) \tag{2.5}$$

that is, if

$$\frac{P(C_1|x)}{P(C_2|x)} > \lambda \tag{2.6}$$

Since $P(C_1|x) + P(C_2|x) = 1$, this gives

$$\frac{P(C_1|x)}{1 - P(C_1|x)} > \lambda \tag{2.7}$$

$$P(C_1|x) > \frac{\lambda}{1+\lambda} \tag{2.8}$$

We see that the threshold of 0.5 which we use to calculate misclassification error corresponds to $\lambda = 1$, where a false positive and a false negative has equal loss. We get a variety of thresholds when we vary λ:

$$\theta = \frac{\lambda}{1+\lambda} \tag{2.9}$$

$$\lambda = 0.5 \quad \rightarrow \quad \theta = 1/3 \tag{2.10}$$

$$\lambda = 1 \quad \rightarrow \quad \theta = 1/2 \tag{2.11}$$

$$\lambda = 2 \quad \rightarrow \quad \theta = 2/3 \tag{2.12}$$

That is, the threshold points on the ROC curve indicate the λ values in the risk calculation. This is the reason why using the ROC curve (or the AUC value, as we will see) is better than using the misclassification error because the latter gives equal emphasis and makes no distinction between false positives and false negatives and thus may not be the best measure for many applications; the ROC curve (and the AUC) takes a set of possible loss proportions into account and this makes them a more robust measure.

If the ROC curve of the first classifier is always over the ROC curve of the second classifier, we can easily say that the first classifier is better than the second classifier. But this case does not always happen. In some cases, the ROC curve of the first classifier may be over the ROC curve of the second classifier in one part, whereas the second classifier's curve is over the ROC curve of the first one in some other part; this implies that the two classifiers are to be preferred under different loss conditions.

2.2. AREA UNDER THE ROC CURVE

ROC is a curve; one may reduce it to a single value using the Area Under ROC Curve (AUC). If a ROC curve is closer to the upper left corner, its AUC value gets closer to 1. The AUC value of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive and negative instance correctly. A classifier with a greater AUC is said to be better than a classifier with a smaller AUC.

1: **Input:** L, the set of test examples; f(i), the posterior probability that example i is positive; P and N, the number of positive and negative examples.
2: **Output:** A, the area under the ROC curve .
3: **Require:** $P > 0$ and $N > 0$.
4: $L_{sorted} \leftarrow L$ sort in decreasing order of f values
5: $FP \leftarrow TP \leftarrow 0$
6: $FP_{prev} \leftarrow TP_{prev} \leftarrow 0$
7: $A \leftarrow 0$
8: $f_{prev} \leftarrow -\infty$
9: $i \leftarrow 1$
10: **while** $i \leq |L_{sorted}|$ **do**
11: **if** $f(i) \neq f_{prev}$ **then**
12: $A \leftarrow A + AREA(FP, FP_{prev}, TP, TP_{prev})$
13: $f_{prev} \leftarrow f(i)$
14: $FP_{prev} \leftarrow FP$
15: $TP_{prev} \leftarrow TP$
16: **end if**
17: **if** $L_{sorted}[i]$ is a positive example **then**
18: TP $\leftarrow$ TP+1
19: **else**
20: FP $\leftarrow$ FP+1
21: **end if**
22: $i = i + 1$
23: **end while**
24: $A \leftarrow A + AREA(N, FP_{prev}, N, TP_{prev})$
25: $A \leftarrow A/(P \times N)$
26: FUNCTION $AREA(X1, X2, Y1, Y2)$
27: $Base \leftarrow |X1 - X2|$
28: $Height_{avg} \leftarrow (Y1 + Y2)/2$
29: RETURN $Base \times Height_{avg}$

Figure 2.3. AUC calculation [4]

The Wilcoxon statistic is calculated to be the estimate of the 'true' area under the ROC curve for the infinite sample. Let N_1 be the number of examples that belong to the positive class C_1 and N_2 be the number of examples that belong to the negative class C_2. Then Wilcoxon statistic is defined as [6]

$$W = \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} S(P(C_1|x_i), P(C_2|x_j) \tag{2.13}$$

where

$$S(P(C_1|x_i), P(C_2|x_j)) = \begin{cases} 1 & \text{if } P(C_1|x_i) > P(C_2|x_j) \\ 1/2 & \text{if } P(C_1|x_i) = P(C_2|x_j) \\ 0 & \text{if } P(C_1|x_i) < P(C_2|x_j) \end{cases} \tag{2.14}$$

The area under the ROC curve can be estimated by summing the tropezoidal areas formed by successive points on the ROC curve. The AUC calculation algorithm is given in Figure 2.3 [4]. It is very similar to the algorithm in Figure 2.2. Now, the previous TP and FP points are stored for calculating the trapezoidal area. Therefore, TP, FP, the previous TP and FP points are initialized to zero in Line 5 and Line 6. In Line 7, the area under the ROC curve is initialized to zero since it is calculated cumulatively at each iteration. If the condition in Line 11 is met, the area of the current trapezoid is calculated in Line 12 by calling the $AREA$ function that takes parameters of TP, FP, the previous TP and FP points and is added to the area A. In Line 14 and Line 15, current FP and TP points are assigned to the FP_{prev} and TP_{prev}, respectively. In Line 25, A is normalized to obtain an AUC value between 0 and 1.

2.3. THE PRECISION-RECALL CURVE

An alternative to the ROC curve for making a comparison for different threshold values is the Precision-Recall curve. It is mostly used in Information Retrieval [19]. The precision and the recall in each threshold point are calculated and the Precision-Recall curve is constructed by joining these values. Nonparametric estimate of Precision-Recall curves are also another issue that is worked on [18]. They are compared and

preferred to ROC curves in case of large skew in class distribution [16]. An example Precision-Recall curve is given in Figure 2.4.

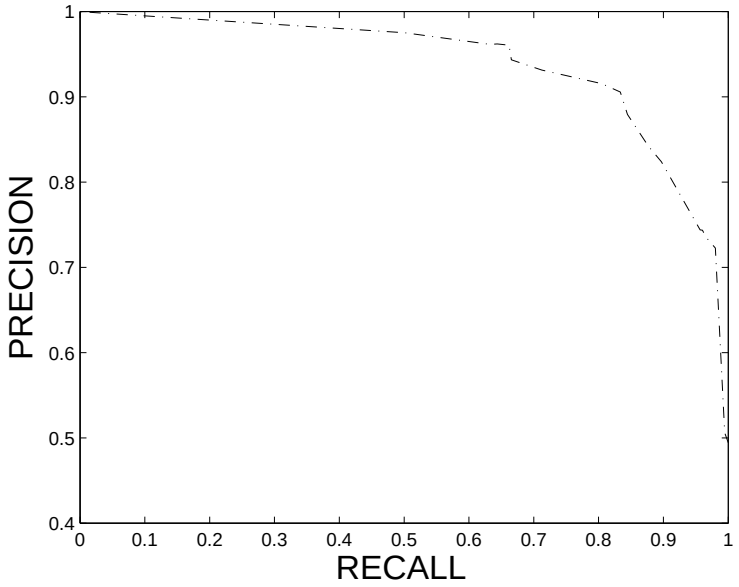


Figure 2.4. Example Precision-Recall curve

For summarizing the ROC curve, the area under the Precision-Recall curve (AUC-PR) is used. It is estimated by the same trepozoidal area calculation method described for AUC in Section 2.2.

2.4. DIFFERENCE BETWEEN THE ROC CURVE AND THE PRECISION-RECALL CURVE

There are three main differences between ROC and Precision-Recall curves [16]. First, their visiualization is different. In ROC curves, TPR is the y-axis and FPR is the x axis. In Precision-Recall curves, Precision is the y-axis and Recall is the x axis. In ROC space, the curve that is upper left of the ROC space is better. In Precision-Recall space, the curve that is upper right of the Precision-Recall space is better.

Second, Precision Recall curves are very sensitive to the class skewness. However, ROC curves are not sensitive to the class skewness. Therefore, Precision Recall curves are preferred when the dataset has a high class skew. It can be understood by looking at confusion matrix in Table 1.1. Since the class skew is the proportion of P to N,

when the class skew changes, the proportion of the first row to second row of Table 1.1 changes. The metrics that use the values of the elements from both rows are effected by the increased class skewness, others are not effected. Since Precision uses values from both rows it changes, however TPR and FPR do not change since they use values from only one row [4].

Third, the AUC and AUC-PR give different comparison results. A classifier that is better in terms of AUC does not have to be better in terms of AUC-PR; a classifier that is better in terms of AUC-PR does not have to be better in terms of AUC.

A one-to-one correspondence between a ROC curve and a PR curve have been proven. It also have been proven that one ROC curve dominates the other ROC curve if and only if the corresponding Precision Recall curve domites the other [16]. Despite the dominance relationship between ROC and PR curves, Davis and Goadrich [16] state that if AUC of one curve is greater than the second one, AUC-PR of the first curve can be less than the second one, therefore optimizing AUC does not mean optimizing also AUC-PR. It is intuitive, we believe that comparison of area under the curves can give different results for ROC and PR curves according to the problem. Corresponding points in curves can dominate each other in paralel in ROC and PR curves, however the magnitude of these differences determines the area differences. Consequently, since the metrics are different, the area between the curves are different. They also argue that the convex hull in ROC space can be converted to achievable PR curve and AUC-PR can be calculated.

3. COMPARISON OF TWO CLASSIFIERS

3.1. PROPOSED AUC TEST

A classifier is first trained using a training set, then, using a test set, one constructs the ROC curve and calculates the AUC value only once. To average over randomness in the training and testing split, one can use more than one training and testing pair, which results in multiple ROC curves and AUC values. The main idea of this part of the thesis is to fit a distribution to these values and test hypotheses on such distributions.

We use k-fold cross-validation to generate k training sets and train k classifiers whose ROC curves and AUC values we calculate over a single test set. At the end, for each classification algorithm we have k AUC values. When we compare multiple algorithms, to have a paired test, we should use the same training and test sets for all algorithms. Afterwards, for example, two classification algorithms can be compared by applying the paired t test with the null hypothesis that two classifiers have the same AUC mean versus the alternative hypothesis that the two AUC means are different.

The k-fold cross-validated paired t test on AUC values will be called as *AUC test* in the rest of this thesis. Each classification algorithm is trained on the training set T_i, $i = 1, \ldots, k$ and the posterior probabilities are calculated on the test set D_i , $i = 1, \ldots, k$. Using these posterior probabilities, the AUC of each classifier is calculated as A_i^1 and A_i^2. The AUC difference is calculated in each fold as $A_i = A_i^1 - A_i^2$ for $i = 1, \ldots, k$. The distribution of differences is normal since the A_i^1 and A_i^2 distributions are approximately normal (we extend the normality assumption for errors to AUC values). Then, if the mean of this distribution can be said to be equal to zero, we can say that classifiers have equal AUC's:

$$H_0 : \mu = 0 \tag{3.1}$$

$$H_1 : \mu \neq 0 \tag{3.2}$$

Then $m = \sum_{i=1}^{k} \frac{A_i}{k}$, $S^2 = \sum_{i=1}^{k} \frac{(A_i - m)^2}{k-1}$. The test statistic is calculated as $\frac{\sqrt{k}.m}{S}$ which is t distributed with $(k-1)$ degrees of freedom. The null hypothesis is accepted at significance level α, if the test statistic is in the interval $(-t_{\alpha/2,k-1}, t_{\alpha/2,k-1})$.

3.2. EXPERIMENTAL SETUP

3.2.1. DATASETS

We use a total of 15 datasets where 11 (*aibocolor, chess, connect-4, mushroom, nursery, pageblock, report, shuttle, spambase, thyroid, wave*) are from UCI and 4 (*ada, caravan, gina, sylva*) are from IJCNN 2007 [22]. The datasets with the number of instances greater than 3000 or approximately 3000 are selected to decrease the dependency between the folds of 30-fold cross validation. Since two-class classification is applied, the datasets with more than two classes are converted to two-class datasets by selecting the two classes that are most confused by looking at the confusion matrix (We first use 1-nearest neighbor over all classes to choose these two).

3.2.2. LEARNING ALGORITHMS

We use five learning algorithms from ISELL Machine Learning Software [23]:

- *C4.5*: C4.5 decision tree algorithm [24].
- *LP*: Linear perceptron with softmax outputs trained by gradient-descent to minimize cross-entropy.
- *k-NN*: k-nearest neighbor. For the optimization of k, values of 1, 3, 5, 7, 11, 21 are tried and the one with minimum validation error is selected.
- *NB*: Naive Bayes which is a parametric discriminator assuming independent inputs.
- *Ripper*: Rule learning algorithm with two optimization steps [25].

3.2.3. DIVISION OF TRAINING, VALIDATION AND TEST SETS

Our methodology is as follows: A data set is first divided into two parts, with 1/3 as the test set, *test*, and 2/3 as the training set, *train-all*. The training set, *train-all*, is then resampled using 30 times cross-validation to generate $tra_i, i = 1, \ldots, 30$, which are used to train the classifiers and the tests are run on the test set.

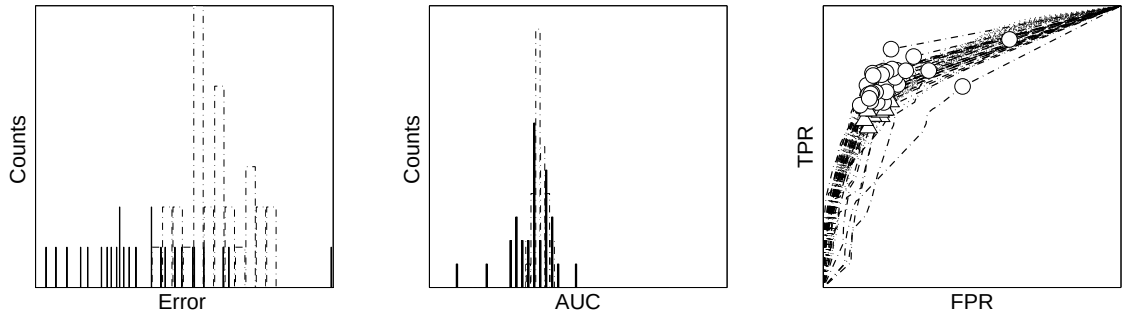
3.3. EXPERIMENTAL RESULTS

We compare 5 algorithms in a pairwise manner on 15 datasets using k-fold cv paired t test on errors (error test) and k-fold cv paired t test on AUC values (AUC test) at the significance level of 0.05, which makes a total of 150 comparisons. The null hypothesis of both the error test and the AUC test are that the two populations have the same mean. There are four possible cases:

- Both the error test and the AUC test accept the null hypothesis. This case occured only 1 time.
- The error test accepts and the AUC test rejects the null hypothesis. This case occured 10 times.
- The error test rejects and the AUC test accepts the null hypothesis. This case occured 9 times.
- Both the error test and the AUC test reject the null hypothesis. This case occured 130 times.

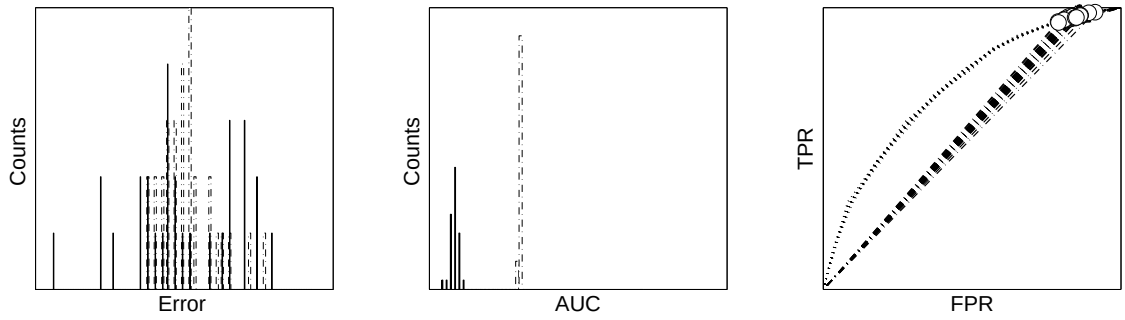
We now discuss some examples of these cases: In Figure 3.1, we show the results on *chess* dataset for *C4.5* and *Ripper* algorithms for the case where both the error test and the AUC test accept the null hypothesis that the two populations have the same mean. We show in (a) the error distributions of two algorithms with white and black histograms. It can be seen that the error distribution of the two algorithms overlap and this supports the decision of error test. We show in (b) the AUC distribution of the two algorithms with white and black histograms. It can be seen that the AUC distribution of the two algorithms also overlap and this supports the decision of AUC test. ROC

curves that can be seen in (c) as white and black curves supports the agreement. Since the ROC curves of the algorithms overlap, it is consistent with the result of the AUC test. The 0.5 threshold points are marked on the ROC curves (by circle and triangle for the two algorithms). The marked points also overlap which is consistent with the result of the error test.



(a) Error distribution (b) AUC distribution (c) ROC curves

Figure 3.1. An example for case 1 where both the error test and the AUC test accept the null hypothesis



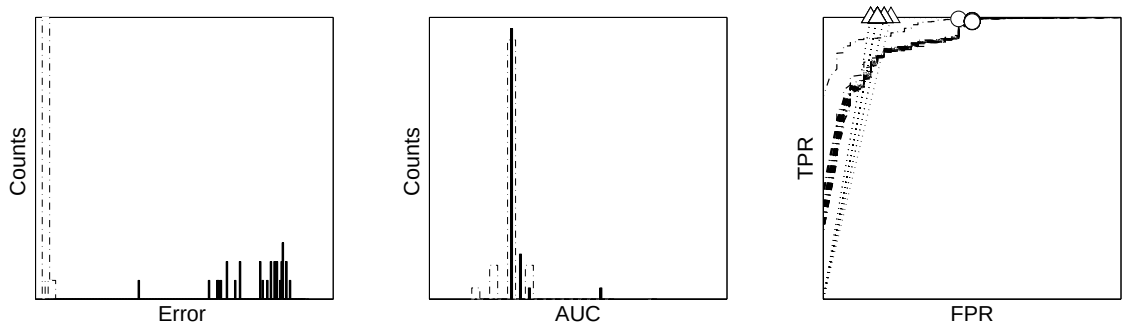
(a) Error distribution (b) AUC distribution (c) ROC curves

Figure 3.2. An example for case 2 where the error test accepts and the AUC test reject the null hypothesis

Figure 3.2 shows the second case where the error test accepts and our AUC test rejects the null hypothesis. In Figure 3.2(a), it can be seen that the error distributions of the *k-NN* (white) and *Ripper* (black) on the *report* dataset overlap and this supports the decision of the error test. In Figure 3.2(b), it can be seen that the AUC distributions are significantly separated which is consistent with the result of the AUC test. In Figure

3.2(c), we see why; the ROC curves of *k-NN* (white) are above the ROC curves of the *Ripper* (black); although the marked threshold points overlap, if we look overall, we see that the algorithms have indeed different behavior over all possible thresholds. We see that the AUC test is able to detect differences that the error test cannot and that is why, we can say that the AUC test has higher power.

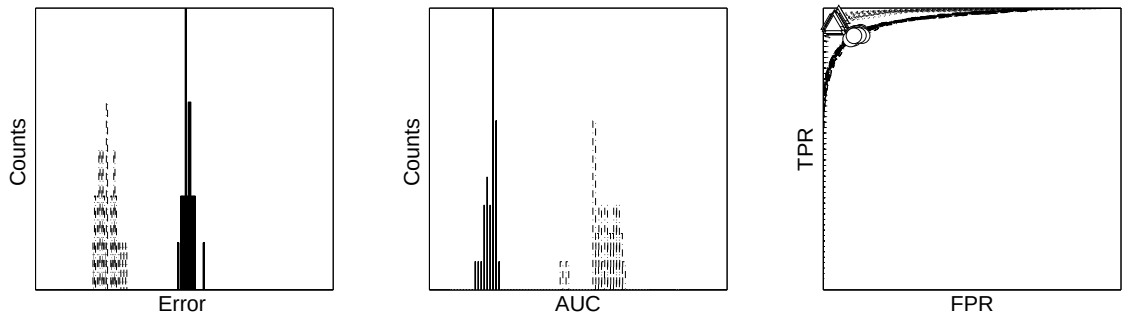
In Figure 3.3, we show the third case where the error test rejects and our AUC test accepts the null hypothesis that the algorithms have equal expected performance. If we look at Figure 3.3(a), we see that there is a significant difference in error distributions of *k-NN* (white) and *NB* (black) on the *shuttle* dataset. Looking at Figure 3.3(b), it can be seen that there is not a significant difference in the AUC distributions. In Figure 3.3(c), the ROC curves intersect; to the left of the intersection, *NB* (black) is better and to the right, *k-NN* (white) is better. Though, the error test says that they are different, if we average over all possible losses (as AUC does), we see that there is no significant difference. The AUC test does not reject such cases and can therefore be said to have lower type I error.



(a) Error distribution (b) AUC distribution (c) ROC curves

Figure 3.3. An example for case 3 where the error test rejects and the AUC test accept the null hypothesis

Figure 3.4 is an example of the fourth case where both the error test and our AUC test reject the null hypothesis. In Figure 3.4(a) and 3.4(b), the error and area distributions of *C4.5* (white) and *LP* (black) on *nursery* dataset are well-separated.



(a) Error distribution (b) AUC distribution (c) AUC distribution

Figure 3.4. An example for case 4 where both the error test and the AUC test reject the null hypothesis

Figure 3.4(c) also supports this claim, the ROC curves of *LP* (black) are over the ROC curves of *C4.5* (white) and the threshold marks are also quite well-separated.

3.4. RELATED WORK

Hanley and McNeil [6] stated that Wilcoxon statistic is an estimate of 'true' area under the ROC curve, the area constructed from an infinite sample. They have also given a standard error formula which takes five parameters: the probability that two randomly chosen abnormal images will both be ranked higher than a randomly chosen normal image, the probability that one randomly chosen positive example will be ranked higher than two randomly chosen negative examples, the number of positive examples, the number of negative examples and the estimated area under the ROC curve. However, for calculating the standard error of the estimated AUC, the distributions of the positive and negative examples should also be known. Using the probabilities defined in the calculation of standard error, they have also given a formula that finds the required number of positive and negative examples for detecting the difference of two AUC's depending on the specified type I and type II error rates (It also requires specific distributions for the values of positive and negative samples).

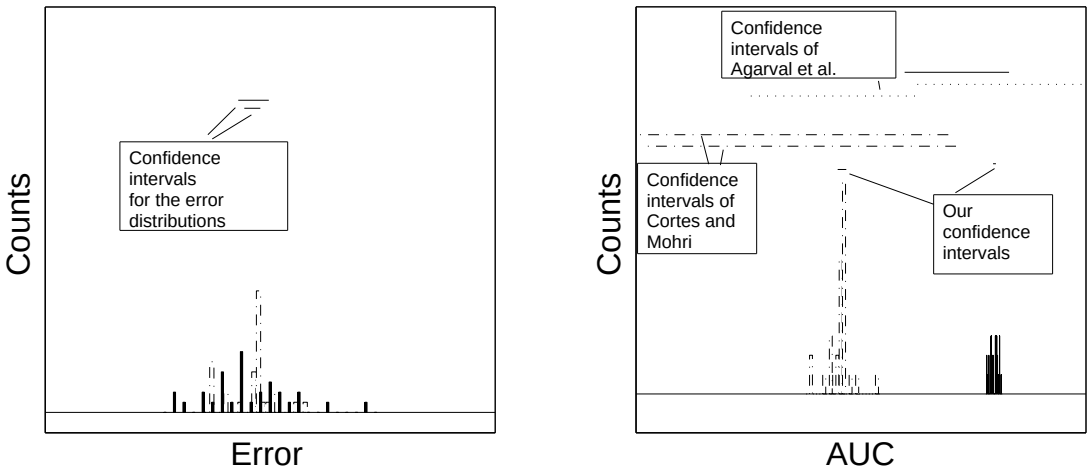
Hanley and McNeil [26] have argued that comparing different ROC curves with

a single dataset limits their usefulness. They state that there is a correlation between AUC's calculated from the same dataset, where correlation is included in the calculation of the standard error of difference in AUC's. They have noticed that a paired test can be used for comparing two algorithms and therefore included the correlation in the statistical test for applying the behaviour of paired t test. A z test statistic is constructed using this standard error and the null hypothesis that 'true' AUC's are equal. They state that they make a correction for pairing like t test. However, we directly use the paired t test, by applying cross-validation to dataset. Therefore, their motivation supports our work.

Paired t test is applied to AUC results, but it is not compared with an error test. It is only used for evaluating the results[10].

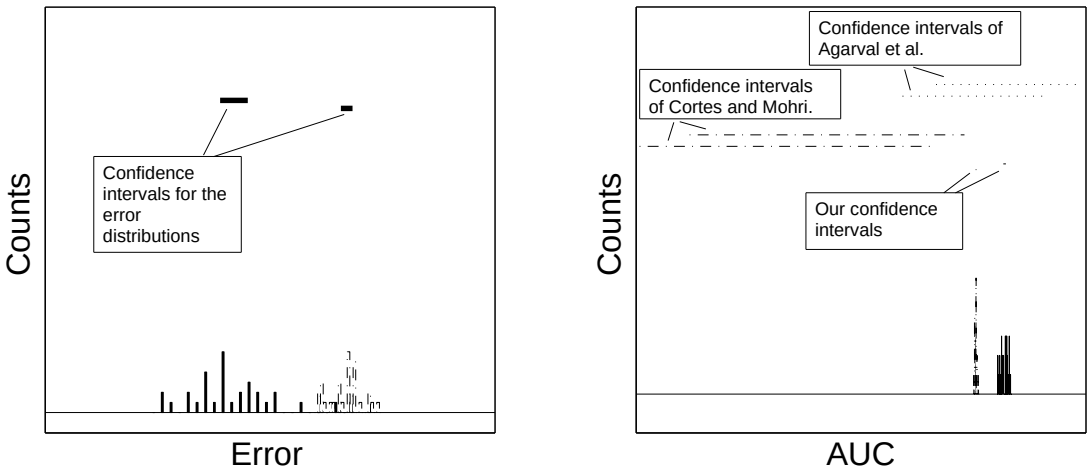
Cortes and Mohri [11] have also proposed to calculate confidence intervals for AUC. A confidence interval for AUC has been derived from the confidence interval of error. First, they define expectation and variance of AUC in terms of the expected error, the number of negative instances and the number of positive instances by using the Wilcoxon-Mann-Whitney statistic. Using these values, the confidence intervals are constructed without any assumption on the distribution for AUC. For large values of the sample size, they make a normal distribution assumption for error.

We argue that there are two weaknesses in their work. First, using error for deriving a confidence interval for AUC is not a good idea, because as we show below, in some cases, AUC intervals can be significantly different although the error intervals are not significantly different. However, their confidence interval formulations give the same AUC interval for the same error value. For comparing our results with their results, we trained and tested the classification algorithms without cross-validation and substituted the error results in their formulations since they use one error value. In Figure 3.5 (a), the error distributions of the classifiers *Ripper* and *LP* on dataset *ada* are shown, they overlap indicating the equality of their means and the error test can not reject the null hypothesis that the means of these error distributions are equal. However, in Figure 3.5(b), it can be seen that the corresponding AUC distributions



(a) Error distribution (b) AUC distribution

Figure 3.5. Confidence intervals for error and AUC for the case where the error test accepts and the AUC test rejects the null hypothesis



(a) Error distribution (b) AUC distribution

Figure 3.6. Confidence intervals for error and AUC for the case where the both error test and the AUC test reject the null hypothesis

are separated despite the overlapping of error distributions and our AUC test rejects the null hypothesis that the means of these AUC distributions are equal. The dashed-dotted lines above the distributions in Figure 3.5(b) are the AUC confidence intervals found by the method of Cortes and Mohri [11]. Their confidence intervals do not show a good fit to the empirical AUC distributions since AUC confidence intervals can be significantly different although error confidence intervals are not. Their confidence intervals also fail when the error results are different. In Figure 3.6(a), the error distributions of the classifiers *k-NN* and *LP* on dataset *ada* are shown, they do not overlap and the error test can not reject the null hypothesis that the means of these AUC distributions are equal. In Figure 3.6(b), the AUC distributions do not overlap and our AUC test can not reject the null hypothesis that the means of these AUC distributions are equal. The confidence intervals of Cortes and Mohri do not fit to the distributions. Another point to note is that, as seen in Figures 3.5(b) and 3.6(b), their confidence intervals are too large because their approach is nonparametric. However, they are inefficient when the sufficient conditions for the distribution assumptions are met.

Another approach for finding the confidence intervals for AUC has been proposed in [12]. Agarwal et al. give a large deviation bound for the distribution independent case. In Figures 3.5(b) and 3.6(b), their confidence intervals are shown with dotted lines above the AUC distributions. The figures support their claim that confidence intervals are too large since no distribution assumption is made. They state that the AUC value follows an asymptotically normal distribution and for large N, the normal approximation can be used to obtain a tighter bound (as we do for deriving the parametric t test). They also state that one can estimate the actual variance of AUC directly from data for obtaining tighter intervals, for example, one can use resampling methods to approximate it that they can be useful in practice despite being approximate. This is similar to what we have done in our proposed test. They also criticize the AUC definition of Cortes and Mohri because of the same reason that we have stated above. They argue that AUC and error are different metrics, therefore different analyses should be done for them.

AUC values have been used to compare classifiers over multiple datasets [2]. However, in our work, we try to gain an insight to the difference in the behavior of the error and AUC tests. J. Demsar compares two classifiers with paired t test over multiple datasets. They state that this test makes normality assumption on the difference of random variables and for this, the dataset size should be approximately 30. They also use the Wilcoxon signed-ranks test since it is nonparametric compared to the paired t test. They calculate AUC values by applying 5-fold internal cross validation and take the average of them, thus they do not apply test on these values like us. They compare AUC of different C4.5 algorithms over 14 datasets. They state that commensurability of differences over datasets can be assumed and no distribution assumption is done in this nonparametric test compared to the paired t test. They compare AUC's of C4.5 algorithms with 5-fold internal cross validation over 14 datasets using the Friedman test which is a nonparametric version of ANOVA.

The ROC curves are preferred when there is class skewness and different misclassification costs. The effect of class distribution on error and AUC is also experimented in [13]. We experiment the effect of imbalanced cost in error and AUC.

Precision Recall curves are mostly used in Information Retrieval. For instance, it is used in [15] for identifying the user profiles, 10-fold cross validation is applied to the dataset and the Precision and Recall metrics used commonly in Information Retrieval and accuracy used in Machine Learning are calculated and the average of each metric is calculated and the values of each metric for each classifier-problem pair are compared. However, the variance information is lost when the average is taken.

AUC-PR is compared with performance metrics of AUC, *F-Measure* which is derived from the Precision Recall Curve and Kolmogorov-Smirnov (KS) statistic by Folleco et al [14]. They apply different sampling methods to the datasets with imbalanced classes. They use cross-validation on data and in each fold, in testing step, they calculate the value of each performance metric for a classifier. They take the average of each metric and compare the average values of these four metrics for each sampling and classifier. They claim that they do not only take average of the values of the met-

rics that they calculate in each fold, but also use statistical tests and also they claim that statistical tests support their results, however they do not give the names of the statistical tests and their results.

Trepozoidal area under the Precision Recall Curve is used to compare classifiers by applying a three-way ANOVA with factors of thresholds, priors, classification models [17].

3.5. PROPOSED AUC-PR TEST

For making a comparison that takes into account the different threshold values, that is, different cost values, we also use AUC-PR values. The k-fold cross-validated paired t test on AUC values of PR curves will be called the *AUC-PR test* in the rest of this thesis. The test procedure is the same as the procedure in Section 3.1. The experimental setup is the same as the setup in Section 3.2, except that AUC-PR values are calculated instead of AUC values.

3.6. EXPERIMENTAL RESULTS OF AUC-PR TEST

There are four possible cases:

- Both the error test and the AUC-PR test accept the null hypothesis. This case occured only 1 time.
- The error test accepts and the AUC-PR test rejects the null hypothesis. This case occured 10 times.
- The error test rejects and the AUC-PR test accepts the null hypothesis. This case occured 10 times.
- Both the error test and the AUC-PR test reject the null hypothesis. This case occured 129 times.

The distributions of error values and AUC-PR values and Precision-Recall curves of examples for the four cases are drawn. In Figure 3.7, we show the results on *chess*

dataset for *C4.5* and *Ripper* algorithms for the case where both the error test and the AUC-PR test accept the null hypothesis that the two populations have the same mean. Figure 3.8 shows the second case where the error test accepts and our AUC-PR test rejects the null hypothesis for the results on *ada* dataset for *Ripper* and *LP* algorithms. In Figure 3.9, we show the results on *gina* dataset for *k-NN* and *Ripper* algorithms for the case where the error test rejects the null hypothesis and the AUC-PR test accepts the null hypothesis. In Figure 3.10, we show the results on *chess* dataset for *C4.5* and *LP* algorithms for the case where both the error test rejects the null hypothesis and the AUC-PR test rejects the null hypothesis. As seen in these figures, the statistical tests that use error metric result in local decisions by assuming equal costs and the statistical tests that use AUC metric results in global decisions by sweeping over different cost conditions.

The results are slightly different from the AUC test. The different decisions occured in the third and fourth cases. There are 11 different decisions among all the comparisons. We show the examples of four cases of decisions of AUC and AUC-PR: In Figure 3.11, the case that both AUC test and AUC-PR test accept the null hypothesis that two algorithms have equal perfomance is given. In Figure 3.11 (a) Precision-Recall curves of the algorithms *C4.5* white and *Ripper* black on the dataset *chess* overlap. In Figure 3.11 (c), the distributions of AUC-PR of the algorithms also close to eachother. In Figure 3.11 (b) ROC curves of the algorithms overlap. In Figure 3.11 (d), the distributions of AUC of the algorithms also close to eachother. In Figure 3.15, the distributions of errors of the algorithms are close to each other, therefore the error test accepts the null hypothesis.

In Figure 3.12, the case that AUC test rejects the null hypothesis and AUC-PR test accepts the null hypothesis is given. In Figure 3.12 (a) Precision-Recall curves of the algorithms *LP* white and *NB* black on the dataset *chess* overlap. In Figure 3.12 (c), the distributions of AUC-PR of the algorithms are also close to eachother. In Figure 3.12 (b) ROC curves of the algorithms seems overlapping but do not completely overlap. In Figure 3.12 (d), the distributions of AUC of the algorithms are significantly different. In Figure 3.16, the distributions of errors of the algorithms are not close to

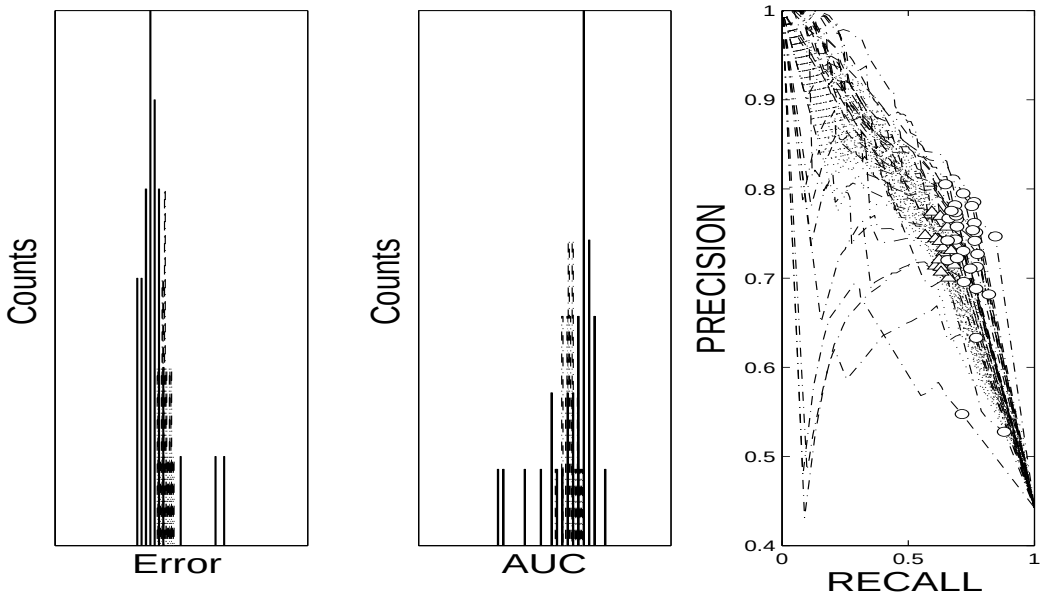


Figure 3.7. An example for case 1 where both the error test and the AUC-PR test accept the null hypothesis

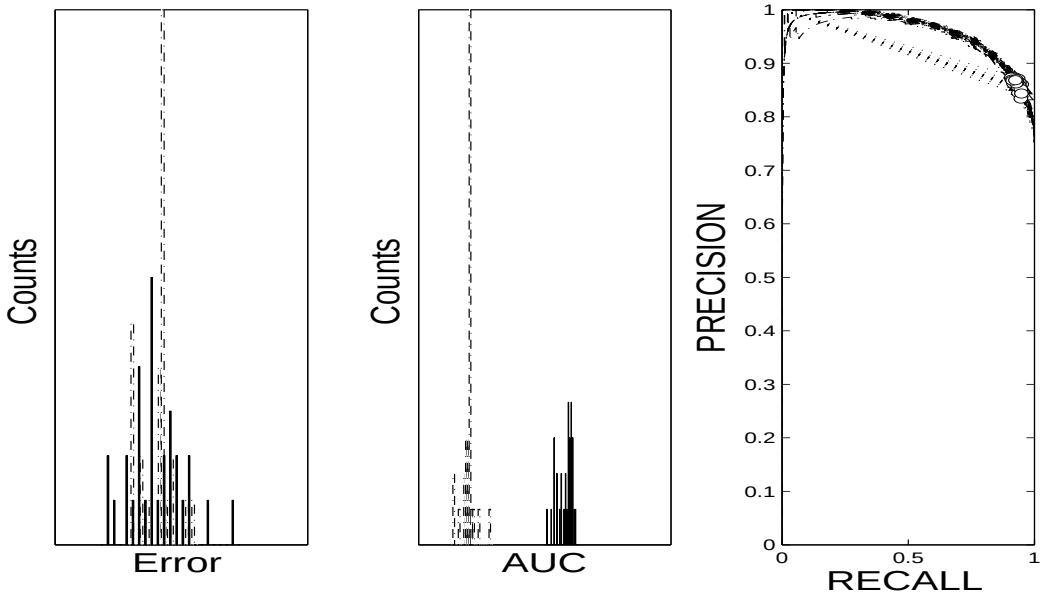


Figure 3.8. An example for case 2 where the error test accepts and the AUC-PR test reject the null hypothesis

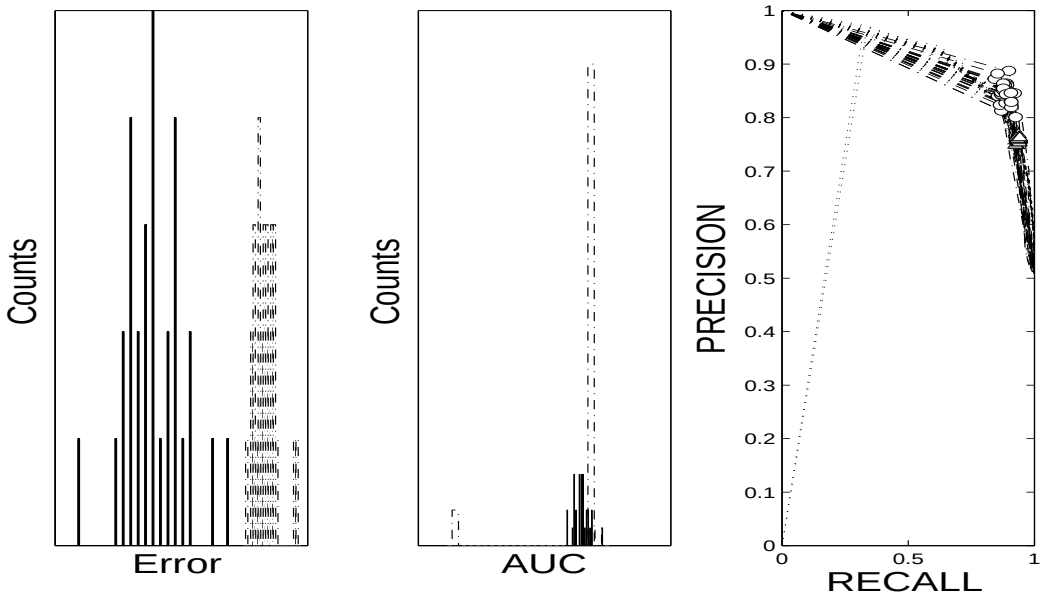


Figure 3.9. An example for case 3 where the error test rejects and the AUC-PR test accept the null hypothesis

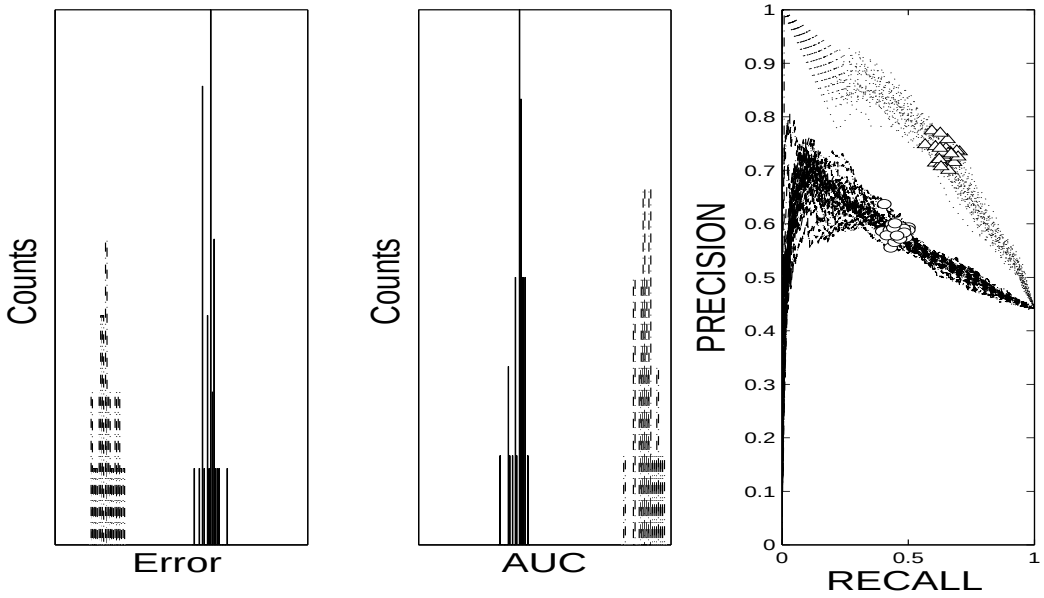
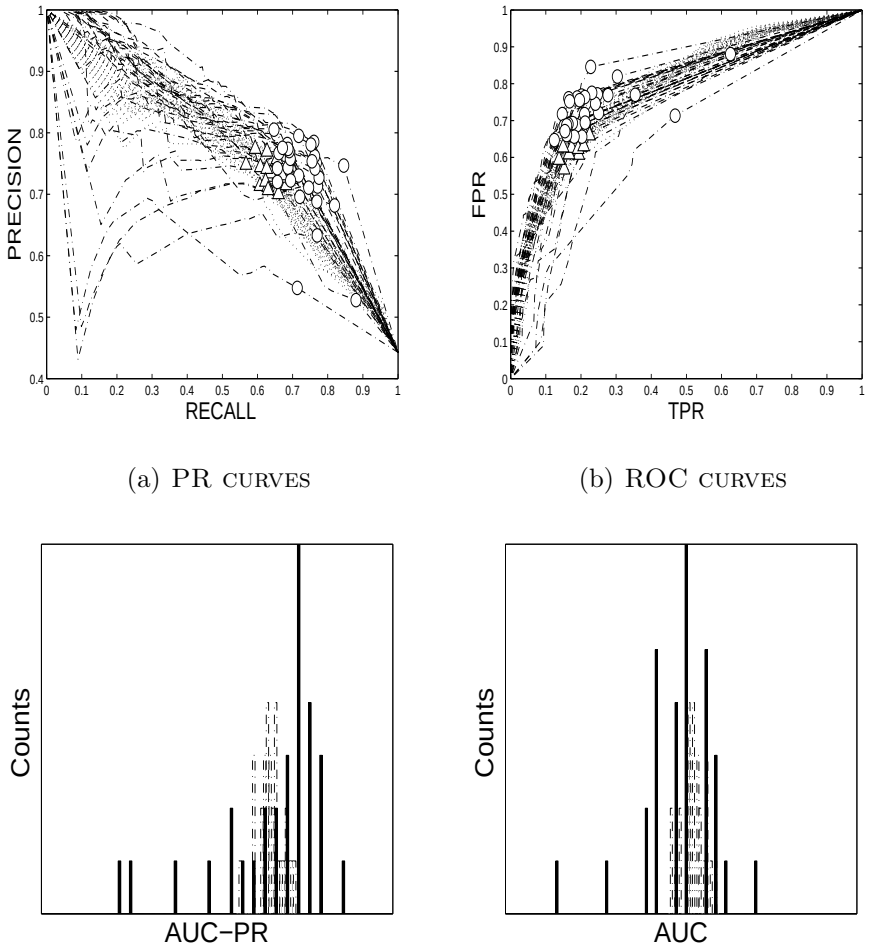


Figure 3.10. An example for case 4 where both the error test and the AUC-PR test reject the null hypothesis

eachother, therefore the error test rejects the null hypothesis.

In Figure 3.13, the case that AUC test accepts the null hypothesis and AUC-PR test rejects the null hypothesis is given. In Figure 3.13 (a) Precision-Recall curves of the algorithms *C4.5* white and *Ripper* black on the dataset *gina* do not overlap; the Precision-Recall curves of *Ripper* dominates the Precision-Recall curves of *C4.5* . In Figure 3.13 (c), the distribution of *Ripper* is on the right hand side of the distribution of AUC-PR of *C4.5* . In Figure 3.13 (b) ROC curves of the algorithms do not completely overlap, but in half of the curves, the ROC curves of *Ripper* dominates the ROC curves of *C4.5* and in the other half, the ROC curves of *C4.5* dominates the ROC curves of *Ripper* , therefore it results in an insignificant difference.. In Figure 3.13 (d), the distributions of AUC of the algorithms are not significantly different. In Figure 3.17, the distributions of errors of the algorithms are not close to eachother, therefore the error test rejects the null hypothesis.

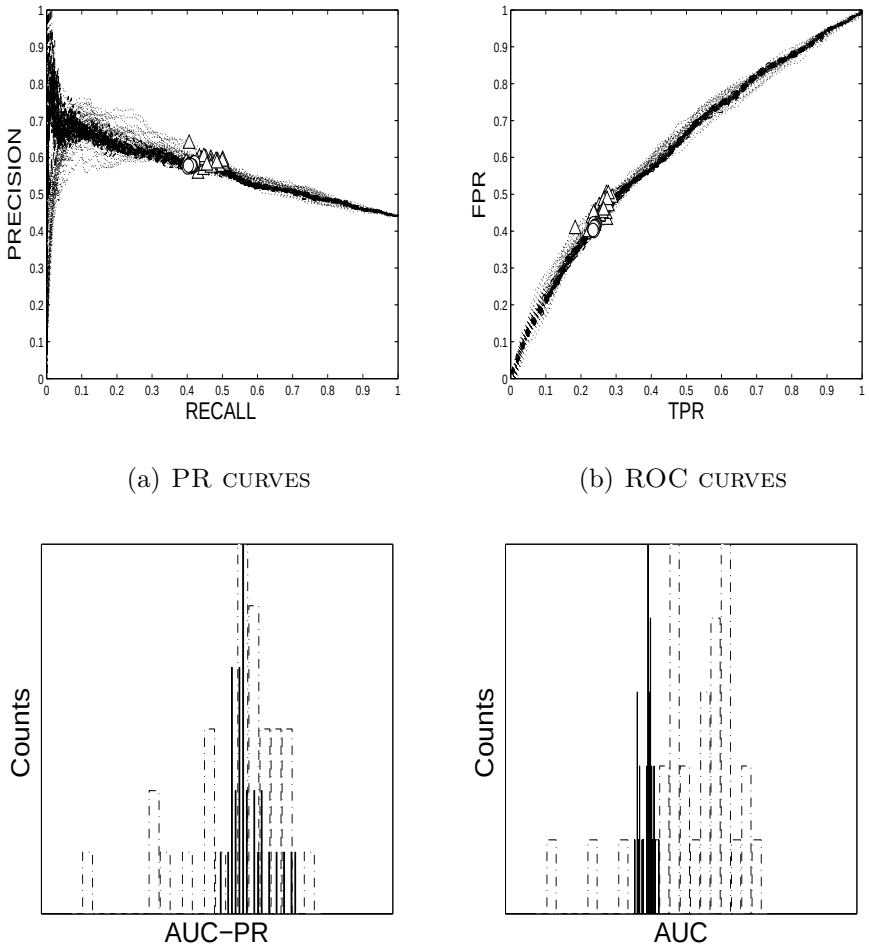
In Figure 3.14, the case that AUC test rejects the null hypothesis and AUC-PR test rejects the null hypothesis is given. In Figure 3.14 (a) the Precision-Recall curves of *C4.5* white dominate the Precision-Recall curves of *LP* black on the dataset *chess*. In Figure 3.14 (c), the distribution of AUC-PR of *C4.5* is on the right hand side of the distribution of AUC-PR of *LP* . In Figure 3.14 (b) ROC curves of the algorithms the ROC curves of *C4.5* white dominate the ROC curves of *LP* black on the dataset. In Figure 3.14 (d), the distributions of AUC of the algorithms are significantly different. In Figure 3.18, the distributions of errors of the algorithms are not close to eachother, therefore the error test rejects the null hypothesis.



(a) PR curves (b) ROC curves

(c) AUC-PR distribution (d) AUC distribution

Figure 3.11. The graphics for *C4.5* and *Ripper* on the dataset *chess*



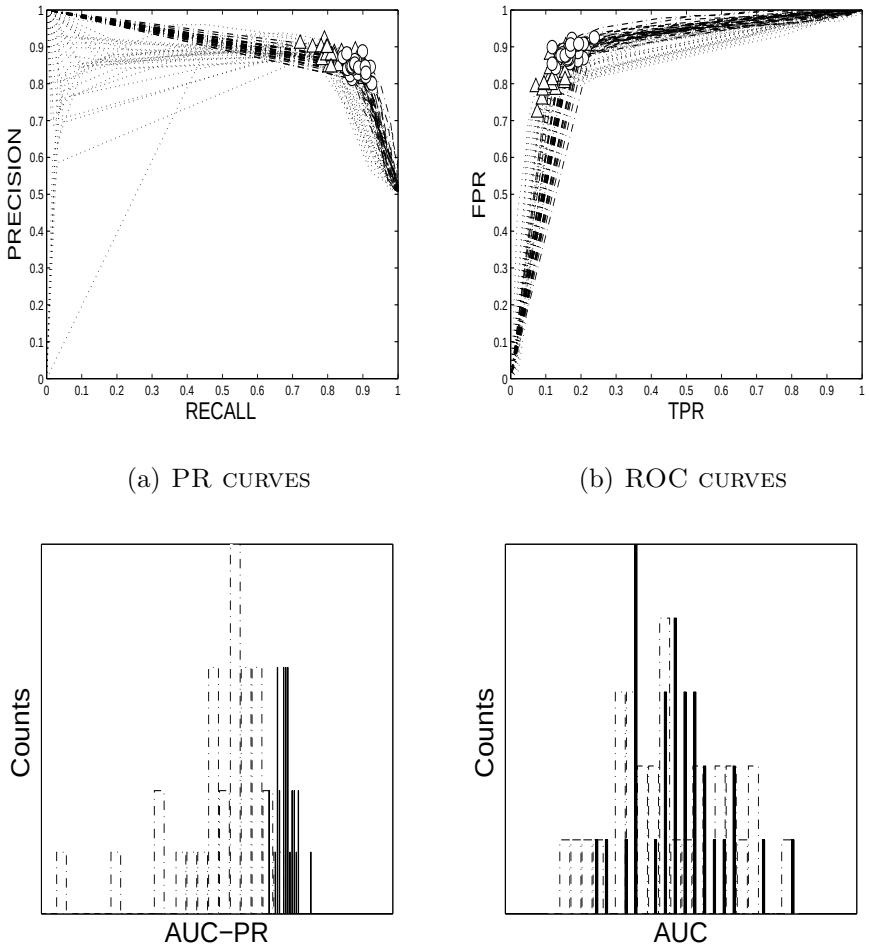
(a) PR curves

(b) ROC curves

(c) AUC-PR distribution

(d) AUC distribution

Figure 3.12. The graphics for *LP* and *NB* on the dataset *chess*



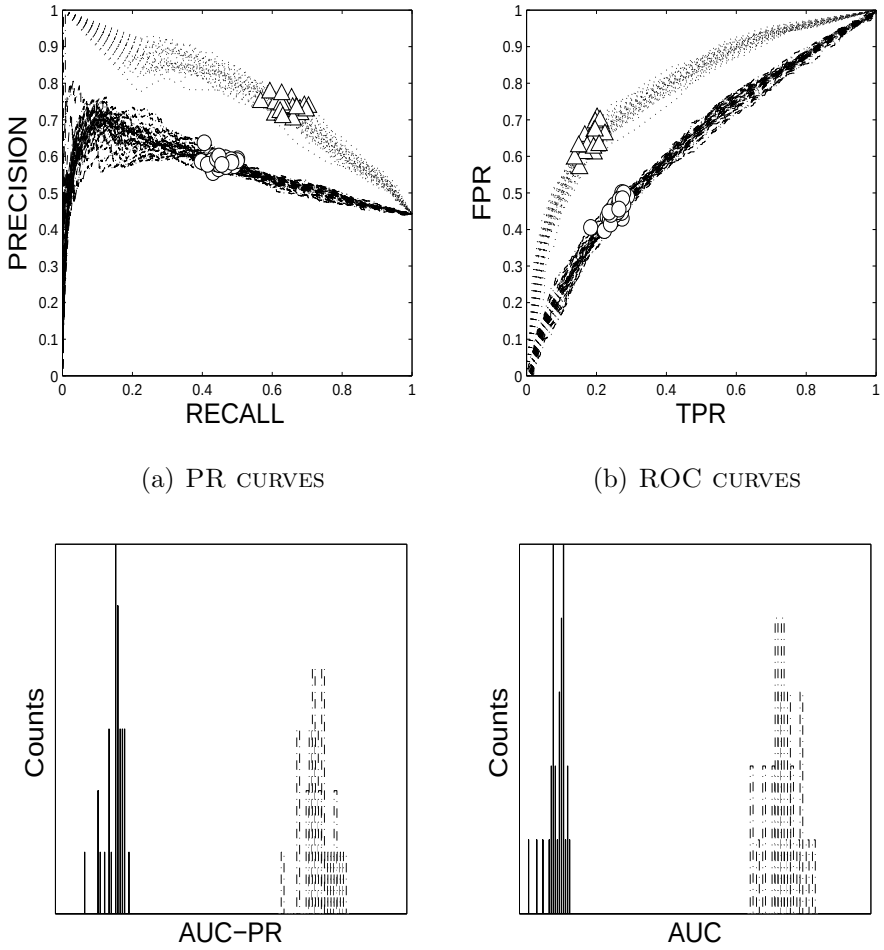
(a) PR curves

(b) ROC curves

(c) AUC-PR distribution

(d) AUC distribution

Figure 3.13. The graphics for *C4.5* and *Ripper* on the dataset *gina*



(a) PR curves

(b) ROC curves

(c) AUC-PR distribution

(d) AUC distribution

Figure 3.14. The graphics for *C4.5* and *LP* on the dataset *chess*

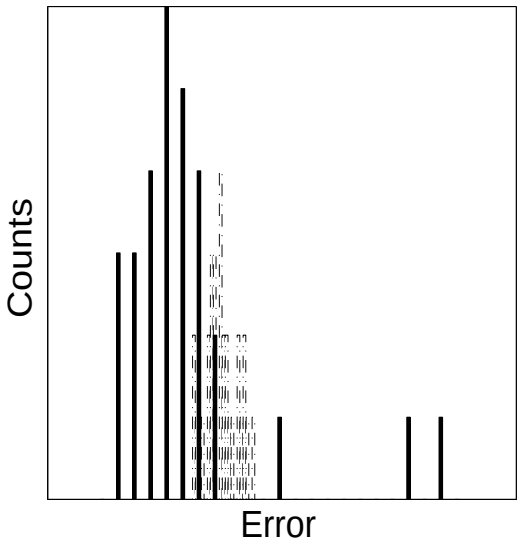


Figure 3.15. The error distribution for *C4.5* and *Ripper* on the dataset *chess*

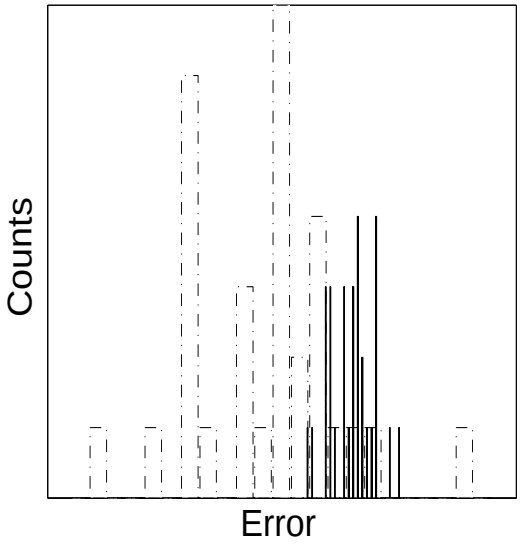


Figure 3.16. The error distribution for *LP* and *NB* on the dataset *chess*

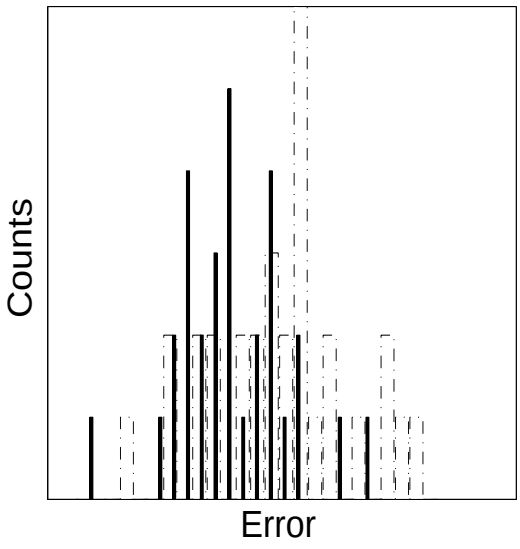


Figure 3.17. The error distribution for *C4.5* and *Ripper* on the dataset *gina*

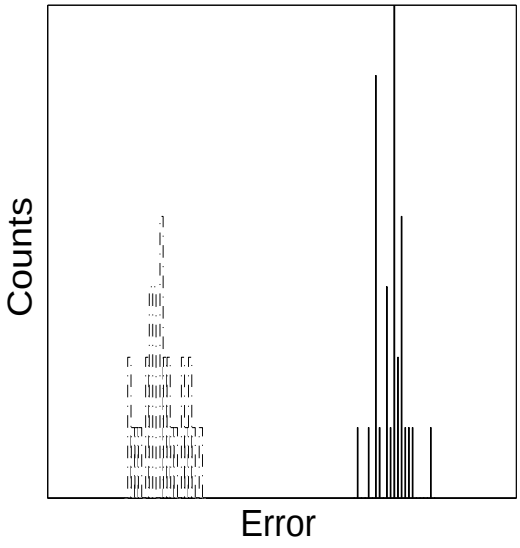


Figure 3.18. The error distribution for *C4.5* and *LP* on the dataset *chess*

4. COMPARISON OF MULTIPLE CLASSIFIERS

4.1. ANALYSIS OF VARIANCE

We have applied k-fold cross-validated paired t test to error and AUC results for comparing two classifiers. However, in the case of multiple classifiers, the paired t test is not applicable and the *analysis of variance (ANOVA)* can be used. ANOVA provides statistical comparison of the means of more than two groups. It is a parametric test that makes normality assumption on each group like the t test. The following theoretical background of ANOVA is taken from [20].

The data that will be used in ANOVA is shown in Table 4.1. There are a *treatments* (groups) and n observations (replications) for each treatment. In our case, treatments correspond to classification algorithms and observations correspond to performance values in folds. The notation is as follows: i in observation y_{ij} is the treatment index and j is the observation index. $y_{i.}$ is the sum of the observations that belong to treatment i, $\bar{y}_{i.}$ is the average of the observations that belong to treatment i, $y_{..}$ is the grand sum of all the the observations in the data and $\bar{y}_{..}$ is the grand average of all the observations in the data:

$$y_{i.} = \sum_{j=1}^{n} y_{ij} \text{ and } \qquad \bar{y}_{i.} = y_{i.}/n, \qquad i = 1, 2, \ldots, a \tag{4.1}$$

$$y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij} \text{ and } \quad \bar{y}_{..} = y_{..}/N \text{ where } \quad N = an \tag{4.2}$$

Each observation y_{ij} is modeled as:

$$y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases} \tag{4.3}$$

where μ_i is the mean of treatment i and ϵ_{ij} is the random error. This model is called

Table 4.1. ANOVA data

Treatments	Observations				Totals	Averages
1	y_{11}	y_{12}	...	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	...	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
3	y_{31}	y_{32}	...	y_{3n}	$y_{3.}$	$\bar{y}_{3.}$
.	.	.	...	.	.	.
.	.	.	...	.	.	.
.	.	.	...	.	.	.
a	y_{a1}	y_{a2}	...	y_{an}	$y_{a.}$	$\bar{y}_{a.}$

the *means model*. Another model is the *effects model* that is defined as:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases} \tag{4.4}$$

where μ is the *overall mean* that is the mean of all treatments and τ_i is the ith treatment effect. Both models are *linear statistical models* since the response variable is linearly dependent on the independent variables. The effects model is more intuitive for interpreting the effect of treatments. This ANOVA model investigates one factor which represents the treatments. Therefore, it is called as *one-way* or *single-factor* analysis of variance model.

Additionally, the model requires a *completely randomized design* where the observations are obtained in a random order to make the experiment conditions effects approximately uniform. Random errors are assumed to be $NID(0, \sigma^2)$, that is, the observations y_{ij} are assumed to be $N(\mu + \tau_i, \sigma^2)$ and mutually independent. Since we want to test whether the means of the treatments are equal or not, the hypothesis test is constructed as

$$H_0 : \qquad \mu_1 = \mu_2 = \cdots = \mu_a \tag{4.5}$$

$$H_1 : \quad \mu_i \neq \mu_j \text{ for at least one pair } (i, j) \tag{4.6}$$

which tests the equality of treatment means. Using the effects model, the hypothesis test can be written as:

$$H_0: \quad \tau_1 = \tau_2 = \cdots = \tau_a = 0 \tag{4.7}$$

$$H_1: \quad \tau_i \neq 0 \text{ for at least one } i \tag{4.8}$$

which tests the equality of treatment effects. Partitioning of total variability is the basis of ANOVA. The total sum of squares is

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 \tag{4.9}$$

It can also be written as

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}\left[(\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})\right]^2 \tag{4.10}$$

then,

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 \tag{4.11}$$

$$+ \; 2\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.}) \tag{4.12}$$

The last term is zero, since

$$\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.}) = y_{i.} - n\bar{y}_{i.} = y_{i.} - n(y_{i.}/n) = 0 \tag{4.13}$$

Then, the total sum of squares can be partitioned as

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 \tag{4.14}$$

The two terms in equation 4.14 can be interpreted as *between* and *within* sum of squares.

The between sum of squares measures the difference between the treatment average and grand average that indicates the difference between treatment means. The within sum of squares measures the difference between the observation and treatment average that indicates the random error. Equation 4.14 can be written as

$$SS_T = SS_{Treatments} + SS_E \tag{4.15}$$

where SS_T has $N-1$ degrees of freedom, $SS_{Treatments}$ has $a-1$ degrees of freedom and SS_E has $N-a$ degrees of freedom. Then, $MS_{Treatments}$ and MS_E are defined as :

$$MS_{Treatments} = \frac{SS_{Treatments}}{a-1} \tag{4.16}$$

$$MS_E = \frac{SS_E}{N-a} \tag{4.17}$$

The expected value of MS_E is equal to the total variability σ^2:

$$E\left[MS_E\right] = E\left[\frac{SS_E}{N-a}\right] = \frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2\right] \tag{4.18}$$

$$= \frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}^2 - 2y_{ij}\bar{y}_{i.} + \bar{y}_{i.}^2)\right] \tag{4.19}$$

$$= \frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 - 2n\sum_{i=1}^{a}\bar{y}_{i.}^2 + n\sum_{i=1}^{a}\bar{y}_{i.}^2\right] \tag{4.20}$$

$$= \frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 - \frac{1}{n}\sum_{i=1}^{a}y_{i.}^2\right] \tag{4.21}$$

It can also be written by substituting the means model:

$$E\left[MS_E\right] = \frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu+\tau_i+\epsilon_{ij})^2 - \frac{1}{n}\sum_{i=1}^{a}\left(\sum_{j=1}^{n}(\mu+\tau_i+\epsilon_{ij}\right)^2\right] \tag{4.22}$$

Since $E\left[\epsilon_{ij}\right] = 0$, it equals to:

$$E\left[MS_E\right] = \frac{1}{N-a}E\left[N\mu^2 + n\sum_{i=1}^{a}\tau_i^2 + N\sigma^2 - N\mu^2 - n\sum_{i=1}^{a}\tau_i^2 - a\sigma^2\right] \tag{4.23}$$

$$E\left[MS_E\right] = \sigma^2 \tag{4.24}$$

It can also be shown that

$$E\left[MS_{Treatments}\right] = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1} \tag{4.25}$$

Therefore, if there is no difference between the treatment means, $MS_{Treatments}$ estimates σ^2. If there is any difference in treatment means, $E\left[MS_{Treatments}\right]$ is greater than σ^2. $MS_{Treatments}$ is chi-square distributed with $a-1$ degrees of freedom and MS_E is chi-square distributed with $N-a$ degrees of freedom. Thus, using Cochran's theorem, the test statistic for the equality of means of treatments can be tested by looking at the ratio of $MS_{Treatments}$ to MS_E:

$$F_0 = \frac{MS_{Treatments}}{MS_E} \tag{4.26}$$

where F_0 is F distributed with $a-1$ and $N-a$ degrees of freedom. The null hypothesis should be rejected if

$$F_0 > F_{\alpha,a-1,N-a} \tag{4.27}$$

If the null hypothesis of the ANOVA test is rejected, then pairwise *post-hoc* tests are applied for making pairwise comparisons. *Tukey's test* is one of them which has the null hypothesis and alternative hypothesis:

$$H_0 : \mu_i = \mu_j \tag{4.28}$$

$$H_1 : \mu_i \neq \mu_j \tag{4.29}$$

for all $i \neq j$. For equal sample sizes, the overall significance level is α. For each pairwise comparison of two samples, the null hypothesis that their means are equal is rejected at significance level α if the absolute value of the difference of their sample mean is

greater than

$$T_\alpha = q_\alpha(a, f)\sqrt{\frac{MS_E}{n}} \tag{4.30}$$

where a is the number of treatments and f is the degrees of freedom related to MS_E. Then, $100(1-\alpha)$ confidence intervals for each sample pair $i \neq j$ can be written as

$$\bar{y}_{i.} - \bar{y}_{j.} - q_\alpha(a, f)\sqrt{\frac{MS_E}{n}} \leq \mu_i - \mu_j \leq \bar{y}_{i.} - \bar{y}_{j.} + q_\alpha(a, f)\sqrt{\frac{MS_E}{n}} \tag{4.31}$$

The distribution of *Studentized range statistic* is

$$q = \frac{\bar{y}_{max} - \bar{y}_{min}}{\sqrt{MS_E}/n} \tag{4.32}$$

where $\bar{y}_{max}$ is the maximum sample mean and $\bar{y}_{min}$ is the minimum sample mean.

4.2. ANALYSIS OF VARIANCE WITH BLOCKING

The effect of variability caused by the nuisance factor on the results can be eliminated by including this factor in the model if it is known and controlable; this is called *blocking*. The experimental setup is now called as *randomized complete block design*. Randomization is applied to each block, the order of treatments is random in each block and this is called as restriction in randomization. *Complete* means that there is a complete set of treatments in each block. The effects model can now be written as

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases} \tag{4.33}$$

where β_j is the block effect. It is assumed that

$$\sum_{i=1}^{a} \tau_i = 0 \tag{4.34}$$

$$\sum_{i=1}^{a} \beta_i = 0 \tag{4.35}$$

The hypothesis test is

$$H_0: \qquad \mu_1 = \mu_2 = \cdots = \mu_a \tag{4.36}$$

$$H_1: \quad \mu_i \neq \mu_j \text{ for at least one } (i,j) \text{ pair} \tag{4.37}$$

Since $\mu_i = (1/n)\sum_{j=1}^{n}(\mu + \tau_i + \beta_j) = \mu + \tau_i$ for treatment i, the hypothesis test can also be written as

$$H_0: \qquad \tau_1 = \tau_2 = \cdots = \tau_a \tag{4.38}$$

$$H_1: \quad \tau_i = 0 \text{ for at least one } i \tag{4.39}$$

The partitioning procedure can be rearranged according to the blocking effect. The sum of squares are

$$y_{i.} = \sum_{j=1}^{n} y_{ij} \quad i = 1, 2, \ldots, a \tag{4.40}$$

$$y_{.j} = \sum_{i=1}^{a} y_{ij} \quad j = 1, 2, \ldots, n \tag{4.41}$$

$$y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} = \sum_{i=1}^{a} y_{i.} = \sum_{j=1}^{n} y_{.j} \tag{4.42}$$

The averages are

$$\bar{y}_{i.} = y_{i.}/n \quad \bar{y}_{.j} = y_{.j}/a \quad \bar{y}_{..} = y_{..}/N \tag{4.43}$$

where $N = an$. The total sum of squares can be written as

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}\left[(\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})\right]^2 \quad (4.44)$$

It can be rearranged as

$$\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 &= n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + a\sum_{j=1}^{n}(\bar{y}_{.j} - \bar{y}_{..})^2 && (4.45)\\
&+ \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 + 2\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{.j} - \bar{y}_{..}) && (4.46)\\
&+ 2\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{.j} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) && (4.47)\\
&+ 2\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) && (4.48)
\end{aligned}$$

The equality of three cross products to zero can be shown, and then

$$\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 &= n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + a\sum_{j=1}^{n}(\bar{y}_{.j} - \bar{y}_{..})^2 && (4.49)\\
&+ \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{.j} - \bar{y}_{i.} + \bar{y}_{..})^2 && (4.50)
\end{aligned}$$

The first term on the right hand side is the sum of squares of differences between the treatment averages and the grand average, the second term is the difference between the block averages and the grand average and the last term is the sum of squares of error:

$$SS_T = SS_{Treatments} + SS_{Blocks} + SS_E \quad (4.51)$$

where SS_T has $N-1$ degrees of freedom, $SS_{Treatments}$ has $a-1$ degrees of freedom, SS_{Blocks} has $b-1$ degrees of freedom. SS_E has $N-1-(a-1)-(b-1) = (a-1)(b-1)$ degrees of freedom. The mean squares can be obtained by dividing the sum of squares

by their degrees of freedom. The expected values of the mean squares are:

$$E\left[MS_{Treatments}\right] = \sigma^2 + n\frac{\sum_{i=1}^{a}\tau_i^2}{a-1} \tag{4.52}$$

$$E\left[MS_{Blocks}\right] = \sigma^2 + a\frac{\sum_{j=1}^{n}\beta_j^2}{n-1} \tag{4.53}$$

$$E\left[MS_E\right] = \sigma^2 \tag{4.54}$$

The test statistic is F distributed with $a-1$ and $(a-1)(b-1)$ degrees of freedom and it can be calculated by dividing mean squares of treatment by mean squares of error:

$$F_0 = \frac{MS_{Treatments}}{MS_E} \tag{4.55}$$

The null hypothesis that the means of all treatments are equal can be rejected if $F_0 > F_{\alpha,a-1,(a-1)(b-1)}$.

4.3. FRIEDMAN TEST

Friedman test is a nonparametric version of ANOVA which is a parametric statistical test. Instead of fitting a normal distribution like ANOVA, it ranks the classifiers on each dataset according to their performances. Following formulations are taken from [2]. Let r_j^i be the rank of classifier i on the dataset j and average rank of the classifier i over L datasets be

$$R_i = \frac{1}{L}\sum_j r_j^i \tag{4.56}$$

Instead of comparing means of performance metrics like ANOVA, Friedman test compares average ranks of the classifiers. Thus, the hypothesis test can be contructed as:

$$H_0: \qquad R_1 = R_2 = \cdots = R_a \tag{4.57}$$

$$H_1: \quad R_i \neq R_j \text{ for at least one pair } \quad (i,j) \tag{4.58}$$

The Friedman statistic is defined as

$$\chi_F^2 = \frac{12L}{a(a+1)} \left[\sum_i R_i^2 - \frac{a(a+1)^2}{4} \right] \tag{4.59}$$

is distributed as chi-square distribution with $a-1$ degrees of freedom. A better test statistic which is distributed as F distribution with $a-1$ and $(a-1)(L-1)$ degrees of freedom is proposed since it is too conservative:

$$F_F^2 = \frac{(L-1)\chi_F^2}{L(a-1) - \chi_F^2} \tag{4.60}$$

The post-hoc test performed when Friedman test rejects the null hypothesis is Nemenyi test. Two classifiers have significantly different performances at significance level α if the difference of their average ranks is greater than or equal to the critical difference

$$CD = q_\alpha \sqrt{\frac{a(a+1)}{6L}} \tag{4.61}$$

4.4. MULTIVARIATE ANALYSIS OF VARIANCE

Multivariate analysis of variance (MANOVA) is the multivariate case of univariate analysis of variance. The following theoretical background is taken from [21]. Let us represent the replication i of treatment j as $\mathbf{x}_{ij}$ and since it is p dimensional, it is a $(p \times 1)$ vector. The effect model can be written as

$$\mathbf{x}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_j + \boldsymbol{\varepsilon}_{ij} \tag{4.62}$$

where $\boldsymbol{\varepsilon}_{ij}$ is distribued as $NID(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ is the overall mean and $\boldsymbol{\tau}_j$ is the effect of treatment j where the mean of treatment j is written as

$$\boldsymbol{\mu}_j = \boldsymbol{\mu} + \boldsymbol{\tau}_j \tag{4.63}$$

The equality of means is tested by

$$H_0 : \qquad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \cdots = \boldsymbol{\mu}_a \tag{4.64}$$

$$H_1 : \quad \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \text{ for at least one (i,j)} \tag{4.65}$$

The sum of squares in the univariate case is now in matrix form. There are three SSP (squares and products) matrices *total* SSP matrix, *within-samples* SSP matrix and *between-samples* SSP matrix : $\mathbf{T}, \mathbf{W}, \mathbf{B}$:

$$\mathbf{W} = \sum_{j=1}^{a} \sum_{i=1}^{n} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_j)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_j)^T \tag{4.66}$$

$$\mathbf{B} = \sum_{j=1}^{a} n(\overline{\mathbf{x}}_j - \overline{\mathbf{x}})(\overline{\mathbf{x}}_j - \overline{\mathbf{x}})^T \tag{4.67}$$

$$\mathbf{T} = \sum_{j=1}^{a} \sum_{i=1}^{n} (\mathbf{x}_{ij} - \overline{\mathbf{x}})(\mathbf{x}_{ij} - \overline{\mathbf{x}})^T \tag{4.68}$$

The null hypothesis is tested using the likelihood test, also known as the *Wilks* Λ *test*:

$$\Lambda \;=\; |\mathbf{W}|/|\mathbf{T}| \tag{4.69}$$

$$\Lambda \;=\; |\mathbf{W}|/|\mathbf{W} + \mathbf{B}| \tag{4.70}$$

where $|W|$ is the determinant of matrix W. The null hypothesis is rejected for small values of Λ which can also be written as

$$\Lambda = \prod_{j=1}^{p} (1 + \lambda_j)^{-1} \tag{4.71}$$

where $\lambda_1, \ldots \lambda_p$ are the eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$. Then instead of using only Wilks lambda, the test of dimensionality is applied to obtain the actual dimensionality of data if the null hypothesis is rejected. If the null hypothesis is true, the actual dimensionality is 0. If the null hypothesis is rejected, the dimensionality problem is to find the actual dimension r where $r = 0, 1, \ldots, t$, and $r \leq \min(p, a-1) = t$. Then, the hypothesis test

is constructed as

$$H_0: \quad \boldsymbol{\mu} \text{ lie in an } r \text{ dimensional hyperplane} \tag{4.72}$$

$$H_1: \quad \boldsymbol{\mu}_i \text{ are unrestricted for } i = 1, \ldots, a \tag{4.73}$$

This test is a LR (likelihood ratio) test with known variance. Then, because $\bar{\mathbf{x}}_\mathbf{i} \sim N_p(\boldsymbol{\mu}_\mathbf{i}, \boldsymbol{\Sigma}/n)$, the log likelihood function can be expressed as

$$l(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_a) = c - \frac{1}{2} \sum_{i=1}^{a} n(\boldsymbol{x}_i - \boldsymbol{\mu}_i)^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_i) \tag{4.74}$$

where c is a constant. Under the alternative hypothesis, $\boldsymbol{\mu}_i = \bar{\mathbf{x}}_i$ and

$$\max_{H_1} l(\bar{\boldsymbol{\mu}}_1, \ldots, \bar{\boldsymbol{\mu}}_a) = c \tag{4.75}$$

Under the null hypothesis, the maximum of function l above can be found by using the theorem:

$$\max_{H_0} l(\bar{\boldsymbol{\mu}}_1, \ldots, \bar{\boldsymbol{\mu}}_a) = c - \frac{1}{2}(\gamma_{r+1} + \cdots + \gamma_p) \tag{4.76}$$

where $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_p$ are the eigenvalues of $|B - \gamma\Sigma| = 0$. Then, λ is the difference between Equations 4.76 and 4.75:

$$-2 \log \lambda = \gamma_{r+1} + \cdots + \gamma_p \tag{4.77}$$

If the values of Equation 4.77 are large, then the null hypothesis is rejected. Equation 4.77 is distributed as chi-square with f degrees of freedom for large values of n. Thus,

$$\gamma_{r+1} + \cdots + \gamma_p \sim \chi^2_f \quad f = pa - p(r+1) - (a - r - 1)r = (p - r)(a - r - 1) \tag{4.78}$$

The alternative test is the LR (likelihood ratio) test with unknown variance, estimated by an unbiased estimate $\mathbf{W}/(n - a)$ and the LR test with known variance is applied

with this unbiased estimate. Then, Equation 4.78 becomes

$$(n-a)(\lambda_{r+1}+\cdots+\lambda_p) \sim \chi^2_f \text{ where } \quad f=(p-r)(a-r-1) \tag{4.79}$$

and $\lambda_{r+1} \ldots \lambda_p$ are the roots of

$$|\mathbf{B}-\lambda\mathbf{W}|=0 \tag{4.80}$$

Equation 4.79 is improved with another statistic:

$$D^2_r=(n-1-\frac{1}{2}(p+a))\sum_{j=r+1}^{p}\log(1+\lambda_j) \sim \chi^2_f \tag{4.81}$$

Then, the dimensionality is tested for $r=0,1,\ldots,t$. First, $H_0 : r=0$ is tested. If D^2_0 is significant, we continue with other values for r. When we continue, if the test statistic is significant until r but not significant for r, the dimensionality is taken as r. $r=0$ means that we have a point and we can not reject the hypothesis that the group means are equal. $r=1$ means that the dimensionality is 1 and we can reject the hypothesis that group means are equal and we can not reject the hypothesis that group means are on a line (plane if $r=2$).

The two-sample Hotelling's T^2 test is applied for paired post-hoc comparisons. The Mahalanobis distance D between two populations with means $\boldsymbol{\bar{\mu}_1}$and $\boldsymbol{\bar{\mu}_2}$ using a common covariance matrix $\boldsymbol{\Sigma}$ is

$$D^2=(\bar{\boldsymbol{\mu}}_1-\bar{\boldsymbol{\mu}}_2)^T\boldsymbol{\Sigma^{-1}}(\bar{\boldsymbol{\mu}}_1-\bar{\boldsymbol{\mu}}_2) \tag{4.82}$$

The population parameters can be estimated with the unbiased sample parameters. Two samples with sizes n_1 and n_2 and total size $n_1+n_2=n$ can be calculated with a common sample covariance $\boldsymbol{S}_u=(n_1\boldsymbol{S}_1+n_2\boldsymbol{S}_2)/(n-2)$ by

$$D^2=(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\boldsymbol{S}_u^{-1} \tag{4.83}$$

We represent each sample by a data matrix $\boldsymbol{X}_i$ $(i = 1, 2)$ where the rows of the matrix are the instances. The following theorem offers a distribution for this distance.

If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent and the rows of $\boldsymbol{X}_i$ are *i.i.d.* and $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for $i = 1, 2$ and $\bar{\boldsymbol{\mu}}_1 = \bar{\boldsymbol{\mu}}_2$ and $\bar{\boldsymbol{\Sigma}}_1 = \bar{\boldsymbol{\Sigma}}_2$, $(n_1 n_2)/n$ equals to $T^2(p, n-2)$. This statistic is called *Hotelling's two-sample T^2 statistic* and it can be transformed to an F statistic by

$$\frac{n_1 n_2 (n-p-1)}{n(n-2)p} D^2 \sim F_{p,n-p-1} \tag{4.84}$$

When multiple pairwise comparisons are performed, Bonferroni correction should be applied to obtain a target significance level.

4.5. MULTIVARIATE ANALYSIS OF VARIANCE WITH BLOCKING

The nuisance factor can also be included in MANOVA. Because the blocking matrix **BL** is added now, there are four SSP (squares and products) matrices that are *total* SSP matrix, *within-samples* SSP matrix and *between-samples* SSP matrix , *blocking* SSP matrix: $\mathbf{T}, \mathbf{W}, \mathbf{B}, \mathbf{BL}$:

$$\mathbf{T} = \sum_{j=1}^{a} \sum_{i=1}^{n} (\mathbf{x}_{ij} - \overline{\mathbf{x}})(\mathbf{x}_{ij} - \overline{\mathbf{x}})' \tag{4.85}$$

$$\mathbf{B} = \sum_{j=1}^{a} n(\overline{\mathbf{x}}_j - \overline{\mathbf{x}})(\overline{\mathbf{x}}_j - \overline{\mathbf{x}})' \tag{4.86}$$

$$\mathbf{BL} = a \sum_{j=1}^{a} \sum_{i=1}^{n} (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})' \tag{4.87}$$

$$\mathbf{W} = \mathbf{T} - \mathbf{B} - \mathbf{BL} \tag{4.88}$$

4.6. BINOMIAL SIGN TEST

Following procedure is taken from [27]. Let one classifier be C_1 and other be C_2. Let π_+ be the proportion that C_1 wins C_2 in the classifier population. Since the equality of the performance of two classifiers are tested, null hypothesis states that π_+ equals to 0.5:

$$H_0 : \pi_+ = 0.5 \tag{4.89}$$

$$H_1 : \pi_+ \neq 0.5 \tag{4.90}$$

Let the number of times that C_1 wins C_2 is x. Since we have L datasets, the probability of having a value that is equal to or more extreme than x can be calculated as:

$$P(\geq x) = \sum_{r-x}^{L} \binom{n}{x} (\pi_+)^x (1 - \pi_+)^{L-x} \tag{4.91}$$

If the probability calculated using Equation 4.91 is equal to or less than significance level $\alpha/2$, null hypothesis can be rejected. If proportion of wins are greater than 0.5, the more extreme value means that the values greater than the value x. Conversely, if proportion of wins are less than 0.5, the more extreme value means that the values less than the value x. In case of ties, they are equally seperated among wins and losses. Odd number of them result in a decrease in N.

5. COMPARISON OF MULTIPLE CLASSIFIERS OVER ONE DATASET

5.1. EXPERIMENTAL SETUP

We compare the misclassification error of 5 classifiers for each dataset by applying ANOVA with blocking effect. We add blocking effect since each instance in the dataset comes from one fold. Therefore, blocks are the folds and we eliminate the nuisance factor. If the null hypothesis that the means of errors of 5 classifiers are equal is rejected, we apply Tukey's test as a post-hoc test for making pairwise comparisons.

We calculate TP, FP, TN and FN from the posterior probabilities for each classifier and dataset. Then we apply MANOVA with blocking effect. Again, because of the folds, blocking is included. If the null hypothesis that the means of error of 5 classifiers are equal is rejected, we apply two-sample Hotelling's T^2 test as a post-hoc test for making pairwise comparisons. We use Bonferroni correction in pairwise tests to have an overall significance α.

5.2. EXPERIMENTAL RESULTS

When we apply post-hoc tests after ANOVA and MANOVA, we obtain cases where they do not agree. We first examine the test statistics to observe their behavior. When *k-NN* and *Ripper* is compared, ANOVA that uses error metric accepts and MANOVA that uses TP, FP, TN and FN rejects on *report* dataset. The means and the difference of means of TP, FP, TN and FN of *k-NN* and *Ripper* are given in Table 5.1 and the unbiased pooled covariance matrix estimate is given in Table 5.2. The means, differences in means and variances of errors of *k-NN* and *Ripper* are given in Table 5.3. When the variance of the error decreases, the lower bound for rejection of the null hypothesis of Tukey test decreases, so it rejects more easily. One can observe that the difference of error is not significant but since the error is sum of FP and FN and

they take plus and minus values, MANOVA results in a significant difference. Another disagreement is that ANOVA rejects and MANOVA accepts the null hypothesis for *C4.5* and *Ripper* and on dataset *Pageblock*. The means and the difference of means of TP, FP, TN and FN are given in Table 5.4. The unbiased pooled covariance matrix estimate is given in Table 5.5. Means, difference in means and variances of errors are given in Table 5.6. FP and FN both have plus signs. Therefore, error is divided to these metrics a decrease occurs resulting in an insignificant difference in MANOVA and significant difference in ANOVA.

These two cases give an intuitive difference of the tests. However, we should also take into account the unbiased pooled covariance matrix estimate in both cases, because covariance can change the test statistic. Since we want to visualize the correlation and the difference at the same time, we draw two dimensional data for FP and FN using the unbiased pooled covariance matrix. The examples for the acceptance case of two-sample Hotelling's T^2 test are shown in Figure 5.1. The classifiers are shown with asterisk and cross markers with the common covariance matrix. They are too close to each other. The examples for the rejection case of two-sample Hotelling's T^2 test are shown in Figure 5.2. Classifiers are too far away to each other which results in a greater Mahalanobis distance, so a greater F statistic. We conclude that the reason of this disagreement is that the multivariate version takes into account the false positives and false negatives separately and also their correlation instead of only error which is the sum of these values.

Table 5.1. Means of TP, FP, TN and FN for *k-NN* and *ripper*

	TP	FP	TN	FN
k-NN	78.2768	16.6226	2.7047	2.3959
ripper	77.8237	16.1667	3.1606	2.8490
difference	0.4531	0.4558	-0.4558	-0.4531

Table 5.2. Unbiased pooled covariance matrix estimate for *k-NN* and *Ripper* on *report* dataset

TP	FP	TN	FN	
TP	0.2062	0.1554	-0.1554	-0.2062
FP	0.1554	0.1388	-0.1388	-0.1554
TN	-0.1554	-0.1388	0.1388	0.1554
FN	-0.2062	-0.1554	0.1554	0.2062

Table 5.3. Test statistics calculated with Tukey test

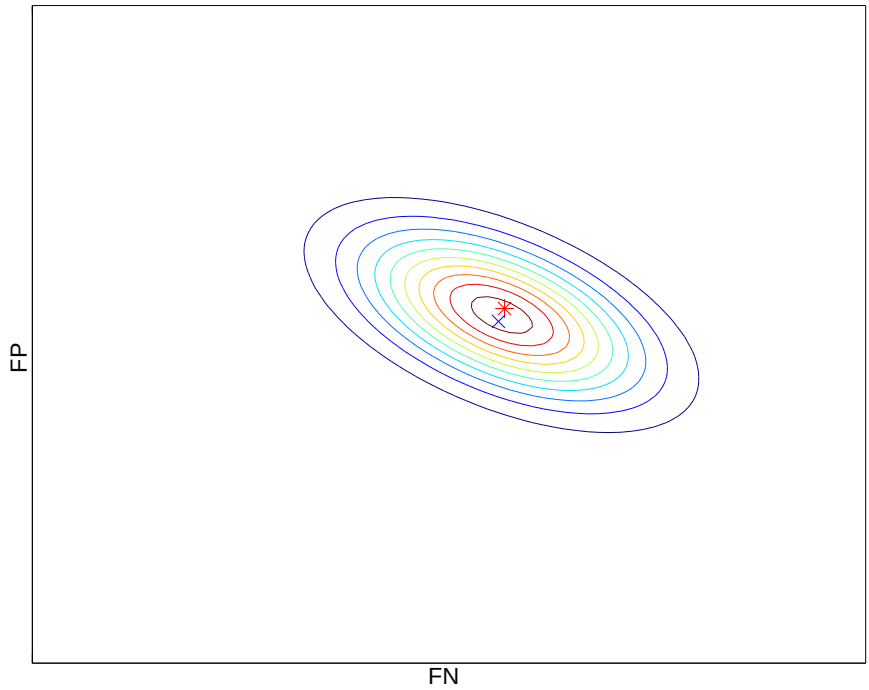
k-NN mean	19.0185
ripper mean	19.0158
mean difference	0.0027
k-NN variance	0.0130
ripper variance	0.0554

Table 5.4. Means of TP, FP, TN and FN for *C4.5* and *Ripper* on *Pageblock* dataset

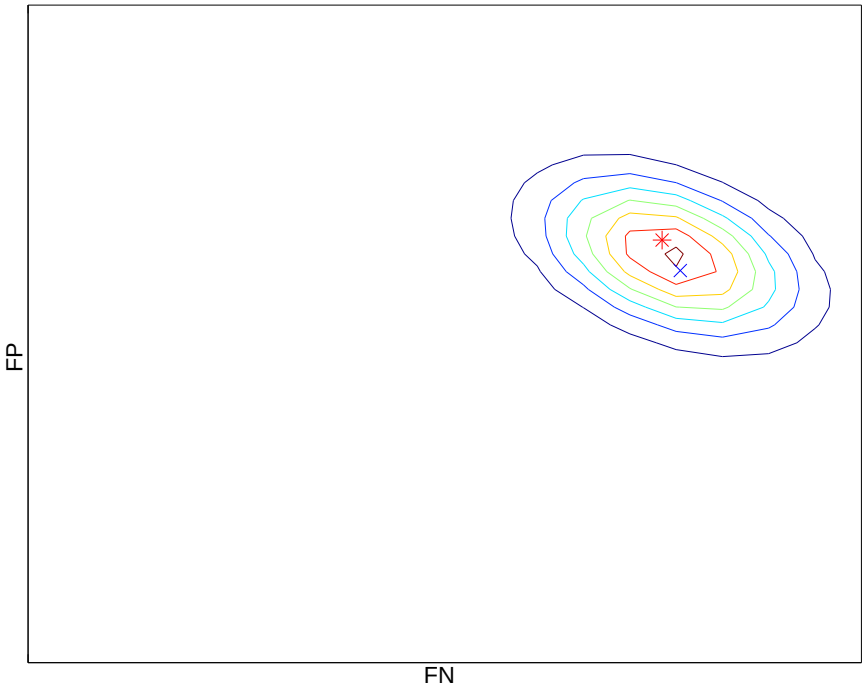
	TP	FP	TN	FN
C4.5	92.8108	0.8047	5.4882	0.8963
Ripper	92.9996	0.7685	5.5244	0.7075
difference	-0.1888	0.0362	-0.0362	0.1888

Table 5.5. Unbiased pooled covariance matrix estimate for *C4.5* and *Ripper* on *Pageblock* dataset

0.1074	0.0395	-0.0395	-0.1074
0.0395	0.0399	-0.0399	-0.0395
-0.0395	-0.0399	0.0399	0.0395
-0.1074	-0.0395	0.0395	0.1074

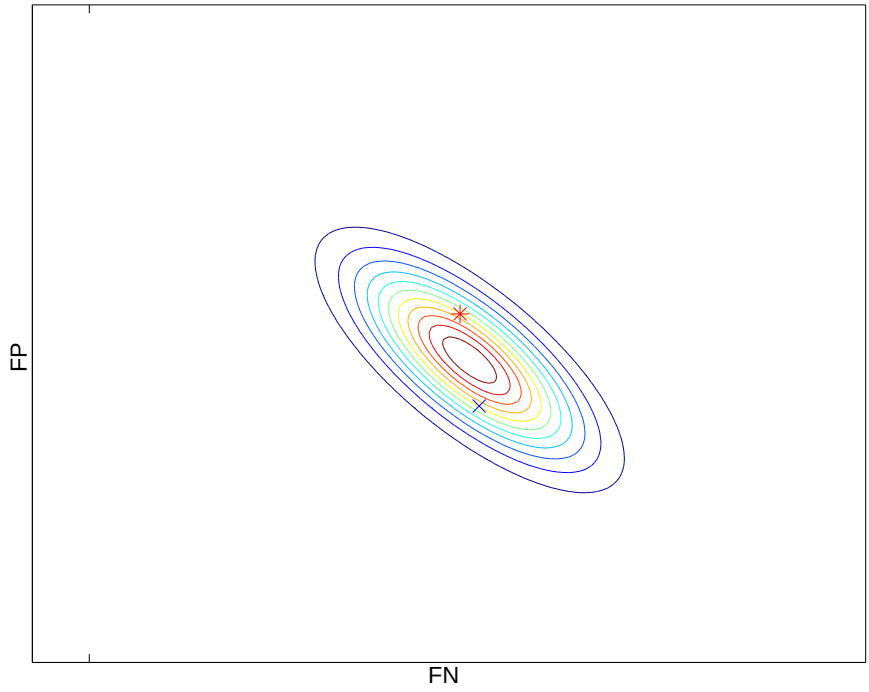


(a) The covariance graphic of *C4.5* and *Ripper* on dataset *spambase*

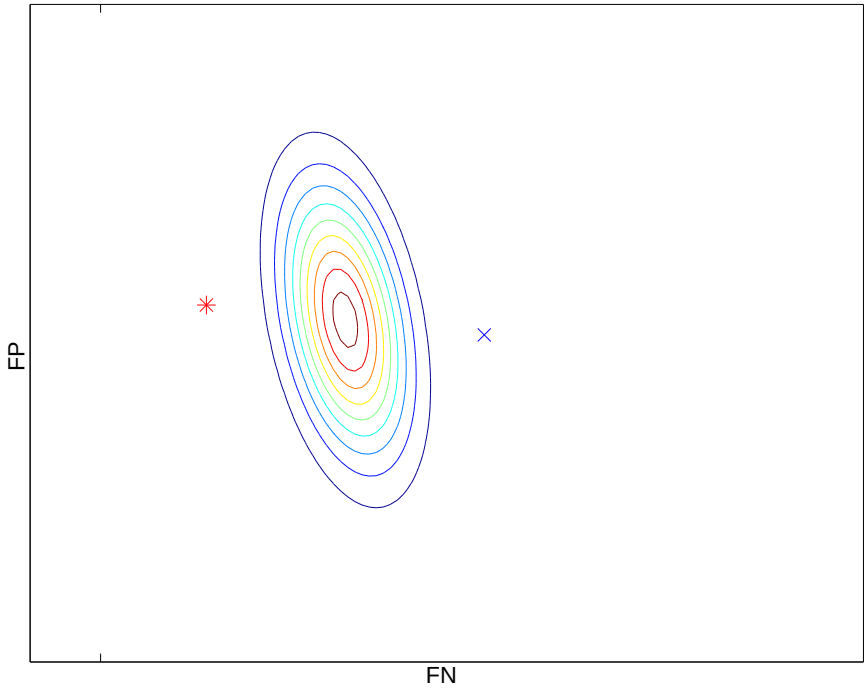


(b) The covariance graphic of *C4.5* and *Ripper* on dataset *thyroid*

Figure 5.1. Examples for the acceptance of null hypothesis of MANOVA



(a) The covariance graphic of *C4.5* and *LP* on dataset *gina*



(b) The covariance graphic of *Ripper* and *LP* on dataset *nursery*

Figure 5.2. Examples for the rejection of null hypothesis of MANOVA

Table 5.6. Test statistics calculated with Tukey test

C4.5 mean	1.7010
Ripper mean	1.4759
mean difference	0.2251
C4.5 variance	0.0997
Ripper variance	0.0368

6. UNIVARIATE COMPARISON OF MULTIPLE CLASSIFIERS OVER MULTIPLE DATASETS

6.1. EXPERIMENTAL SETUP

In previous chapters, we compare multiple classifiers on a single dataset. We now compare 5 classifiers over 15 datasets using ANOVA. We use error and AUC as univariate metrics. For each dataset, we apply ANOVA with two factors that are classifiers and blocking effect because of cross-validation folds. After ANOVA, we apply Tukey test as a post-hoc test. In multiple comparisons, we calculate the wins for each classifier pair. When we use error, if Tukey test rejects the null hypothesis that the two classifiers are equal, the classifier with smaller error wins. When we use AUC, if Tukey test rejects the null hypothesis that the two classifiers are equal, the classifier with greater AUC wins. Thus, we obtain 15 pairwise comparisons of 5 classifiers. We sum up the number of wins over 15 datasets. We also use k-fold cross validated t test with Bonferroni correction for comparing the results with the Tukey test. After obtaining the number of wins of classifier pairs over 15 datasets, we apply Sign test to the number of wins. For each classifier, we take the union of wins gives us an intuitive order. ANOVA is a parametric test which is built on some assumptions. For the sake of completeness, we also apply the nonparametric version of ANOVA test to check if there are any differences in the results. Nemenyi test is used as the post hoc test. We calculate the average of error over 30 folds and the average of AUC over 30 folds and take datasets as a blocking factor. In multiple comparisons, we also calculate the wins for each classifier pair. In pairwise comparisons, if Nemenyi test rejects the null hypothesis, the classifier with smaller average rank wins over the other classifier.

6.2. EXPERIMENTAL RESULTS

Table 6.3 shows the number of wins using Tukey test after ANOVA for the performance metric of error. The total number of wins for each classifier is shown in the

column named as *total*. Therefore we can a ranking based on this column: *k-NN* is better than *Ripper* and *LP*, *C4.5* and *NB* which have equal performances. However, we may want to test the significance of this ranking. Thus, we perform the Sign test on data given in Table 6.3. The result of the Sign test is shown in Table 6.4 and in Figure 6.5 where the classifier pairs that do not have a significant difference in ranking are combined with lines and the average ranks are also shown. As we can predict from the total wins, there are significant differences between *NB* and the other classifiers.

Table 6.5 shows the number of wins of classifiers using Tukey test after ANOVA for the performance metric of AUC. We can obtain the ranking based on total number of wins: *k-NN* is better than *LP*, *LP* is better than *C4.5*, *C4.5* is better than *NB* and *NB* is better than *Ripper*. We perform the Sign test on data in Table 6.5. The result of the Sign test is shown in Table 6.6 and in Figure 6.6. There is a significant difference between only *NB* and *LP*.

We repeat the same procedure for k-fold cross validated t test with Bonferroni correction. The number of wins calculated using t test performed on error results are shown in Table 6.7, an order can be found: *k-NN* has equal performance with *LP*, they are better than *C4.5*, *C4.5* is better than *Ripper*, *Ripper* is better than *NB*. Pairwise comparisons using Sign test is shown in Table 6.8 and in Figure 6.3. The number of wins calculated using t test performed on AUC results are shown in Table 6.9; an order can be found: *k-NN* has equal performance with *LP*, they are better than *C4.5*, *C4.5* has equal performance with *NB*, *NB* is better than *Ripper*. Pairwise comparisons using Sign test is shown in Table 6.10 and in Figure 6.4 . It can be seen that results of paired t test and Tukey test mostly agree.

We also use Friedman test with the post-hoc test of Nemenyi test. The result of Nemenyi test using error is shown in Table 6.1 and Figure 6.2 . Result of Nemenyi test using AUC is shown in Table 6.2 and Figure 6.1.

In all the comparisons that use error as a performance metric, there are only significance differences between *NB* and all the other classifiers. When AUC is used as

a performance metric in paired t test and Tukey test, there are no difference between any classifier pair that do not include *NB* again, but most of the differences between *NB* and all the other classifiers do not exists now. When AUC is used as a performance metric in Nemenyi test, the pairwise comparison results changes totally. It can be seen in Figure 6.1 that Nemenyi results in small differences between classifiers, the reason that Nemenyi test has low power. Binomial Sign test has also a low power since it requires too much number of wins for detecting a significant difference.

We also use AUC-PR metric. The results of Nemenyi test are given in Table 6.11 and they are different from the results of error and AUC. The number of wins using ANOVA and post-hoc test of Tukey are shown in Table 6.12 and the results of Sign test applied to the number of wins are given in Table 6.13. We can see in Table 6.13 that there is no significant difference between any classifier pairs. The number of wins using ANOVA and post-hoc test of t test are shown in Table 6.14 and the results of Sign test applied to the number of wins are given in Table 6.15. We can see in Table 6.15 that there is no significant difference between any classifier pairs. It can be seen that the number of wins are distributed for AUC-PR in a more homogeneous manner than error and AUC. However, the weakness of Sign test prevents detecting differences that could be significant using a more powerful test.

Table 6.1. Results of Nemenyi test using error

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	0	0	1
C4.5	0	0	0	0	1
Ripper	0	0	0	0	1
LP	0	0	0	0	1
NB	1	1	1	1	0

Table 6.2. Results of Nemenyi test using AUC

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	1	0	0
C4.5	0	0	0	1	0
Ripper	1	0	0	1	0
LP	0	1	1	0	0
NB	0	0	0	0	0

Table 6.3. Number of wins obtained from Tukey test after ANOVA with blocking using error

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*	total
k-NN	0	7	5	7	15	15
C4.5	3	0	3	6	14	14
Ripper	3	8	0	6	14	14
LP	7	8	7	0	14	14
NB	0	0	0	0	0	0

Table 6.4. Results of Sign test using Tukey test and error

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	0	0	1
C4.5	0	0	0	0	1
Ripper	0	0	0	0	1
LP	0	0	0	0	1
NB	1	1	1	1	0

Table 6.5. Number of wins obtained from of Tukey test after ANOVA with blocking using AUC

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*	total
k-NN	0	12	13	6	9	14
C4.5	2	0	8	3	5	11
Ripper	2	2	0	2	5	7
LP	6	11	12	0	11	13
NB	3	8	9	1	0	10

Table 6.6. Results of Sign test using Tukey test and AUC

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	0	0	0
C4.5	0	0	0	0	0
Ripper	0	0	0	0	0
LP	0	0	0	0	0
NB	0	0	0	0	0

Table 6.7. Number of wins obtained from paired t test after ANOVA with blocking using error

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*	total
k-NN	0	7	7	7	15	15
C4.5	4	0	4	5	14	14
Ripper	5	6	0	6	13	13
LP	7	9	7	0	15	15
NB	0	0	0	0	0	0

Table 6.8. Results of Sign test using paired t test and error

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	0	0	1
C4.5	0	0	0	0	1
Ripper	0	0	0	0	1
LP	0	0	0	0	1
NB	1	1	1	1	0

Table 6.9. Number of wins obtained from paired t test after ANOVA with blocking using AUC

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*	total
k-NN	0	13	13	7	11	14
C4.5	2	0	6	3	5	9
Ripper	2	2	0	2	5	7
LP	7	12	12	0	14	14
NB	3	8	8	1	0	9

Table 6.10. Results of Sign test using paired t test and AUC

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	0	0	0
C4.5	0	0	0	0	0
Ripper	0	0	0	0	0
LP	0	0	0	0	1
NB	0	0	0	1	0

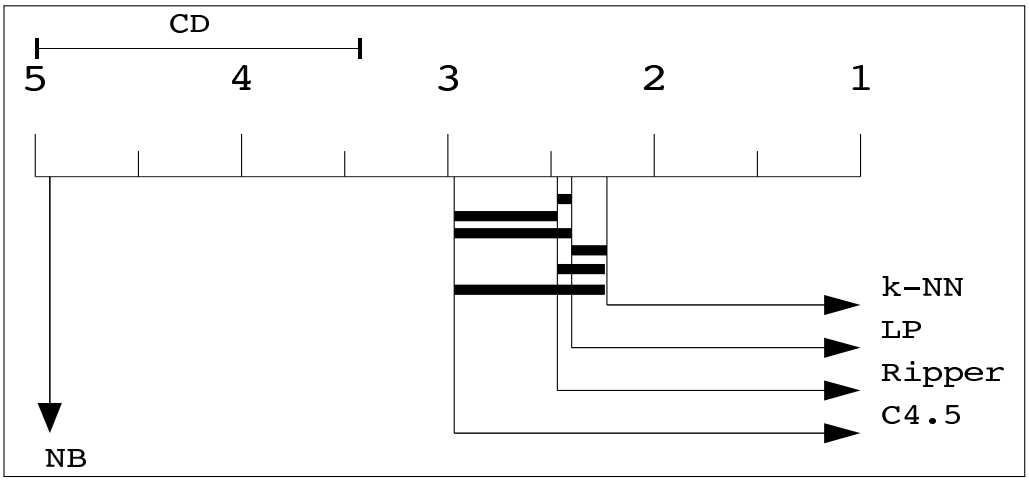


Figure 6.1. Pairwise comparisons of Nemenyi test using error

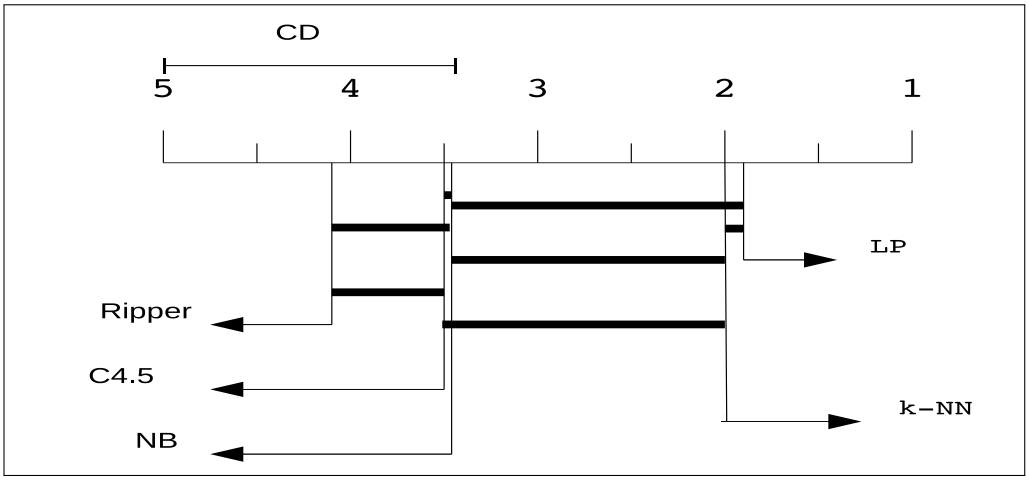


Figure 6.2. Pairwise comparisons of Nemenyi test using AUC

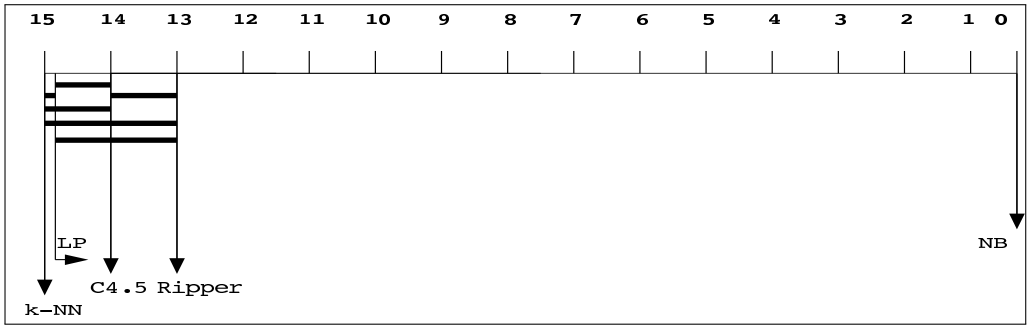


Figure 6.3. Pairwise comparisons of Sign test using paired t test using error

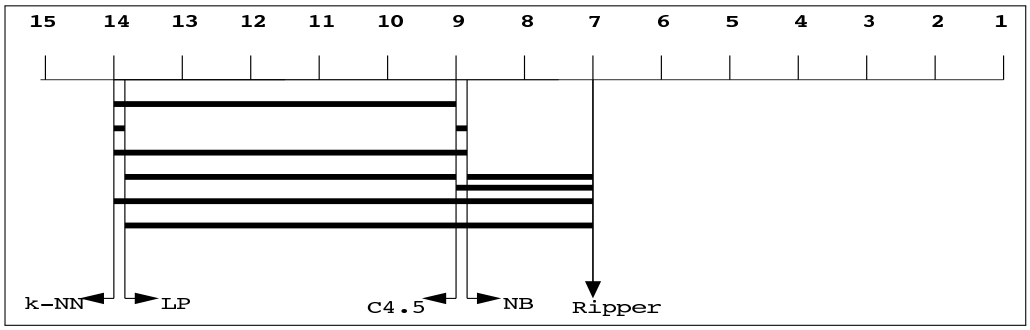


Figure 6.4. Pairwise comparisons of Sign test using paired t test using AUC

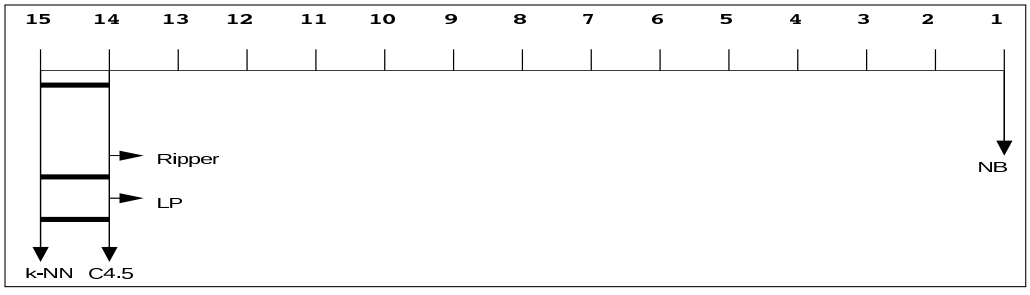


Figure 6.5. Pairwise comparisons of Sign test using Tukey test using error

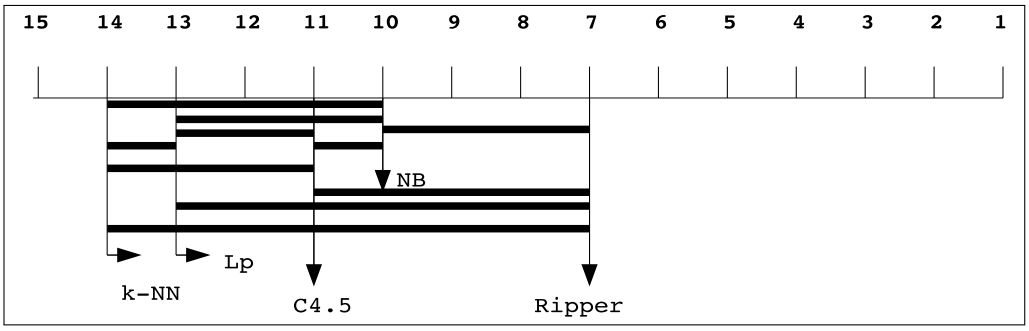


Figure 6.6. Pairwise comparisons of Sign test using Tukey test using AUC

Table 6.11. Results of Nemenyi test using AUC-PR

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	1	0	0	0
C4.5	1	0	0	1	0
Ripper	0	0	0	0	0
LP	0	1	0	0	0
NB	0	0	0	0	0

Table 6.12. Number of wins obtained from Tukey test after ANOVA with blocking using AUC-PR

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*	total
k-NN	0	11	8	4	7	14
C4.5	2	0	3	3	4	6
Ripper	3	8	0	3	5	13
LP	5	11	10	0	7	13
NB	3	10	6	1	0	12

Table 6.13. Results of Sign test using Tukey test and AUC-PR

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	0	0	0
C4.5	0	0	0	0	0
Ripper	0	0	0	0	0
LP	0	0	0	0	0
NB	0	0	0	0	0

Table 6.14. Number of wins obtained from paired t test after ANOVA with blocking using AUC-PR

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*	total
k-NN	0	11	10	5	11	14
C4.5	3	0	2	3	5	6
Ripper	4	5	0	3	7	10
LP	8	11	12	0	12	13
NB	4	8	6	1	0	11

Table 6.15. Results of Sign test using paired t test and AUC-PR

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	0	0	0
C4.5	0	0	0	0	0
Ripper	0	0	0	0	0
LP	0	0	0	0	0
NB	0	0	0	0	0

7. MULTIVARIATE COMPARISON OF MULTIPLE CLASSIFIERS OVER MULTIPLE DATASETS FOR DIFFERENT THRESHOLD POINTS

7.1. EXPERIMENTAL SETUP

We use 4 dimensional multivariate variable that are TP, FP, TN and FN for comparing multiple classifiers over multiple datasets. Since the metric is multivariate, MANOVA is preferred. We include the nuisance factor in the statistical test by taking cross-validation folds as blocking effects. For establishing a comparison that examines all the threshold points like a ROC curve, we perform MANOVA using different values of TP, FP, TN and FN that are calculated from different threshold points. Therefore, a classifier wins if the total risk of the classifier is less than the other for a threshold point. We calculate the risk of a classifier by multiplying TP, FP, TN and FN with their corresponding costs and summing up these multiplied costs. For a dataset, we repeat this procedure for different thresholds from 0 to 1 and obtain a grid of number of wins for each threshold value. Then for each dataset, if a classifier wins over the other for at least one threshold, then we count it as a win.

First, we start with a threshold point of 0.5. We apply MANOVA to each dataset. Because of the value of the threshold point, the costs of the FP and FN are equal. Then, the risk can be calculated by summing up FP and FN which is equal to error. If the post-hoc test rejects the null hypothesis that performances of two classifiers are the same, one classifier wins if sum of its FP and FN is less than the other.

After the threshold point of 0.5, we traverse 21 threshold points: $\theta = 0, 0.05, 0.1, \ldots, 1$. For each dataset, we sum up the 21 different grids calculated from 21 different thresholds, then we apply right-tailed Binomial Sign test to the total grid to test if there is a significant difference over all the threshold points, that is, the ROC curves. Then for each dataset, we have the grids showing the classifier pairs for wins, then we

sum up these 15 grids. We apply two-tailed Binomial Sign test to test if these is a significant difference between the classifiers over 15 datasets.

7.2. EXPERIMENTAL RESULTS

For threshold point of 0.5, dimension results of MANOVA test for each dataset are shown in Table 7.1. MANOVA test can not reject the null hypothesis for only one dataset which has dimension result of zero in Table 7.1. The number of wins obtained using post-hoc test are shown in Figure 7.2. The total performance of each classifier is calculated by taking union of datasets on each row and is shown in the column with the name total. The results of two-tailed Binomial Sign test are shown in Table 7.3. It can be seen that *NB* is the worst classifier and there are not any signicant differences between other classifiers.

For 21 different threshold points, total number of wins calculated by summing up the 15 grid results of right-tailed Binomial test are shown in Table 7.4. The two-tailed Binomial Sign test is applied to the number of wins and results that show equality with zeros and inequality with ones are given in Table 7.5. According to Table 7.4, *NB* seems worst, however it can be seen in Table 7.5 there are no significant difference between any classifiers. Therefore, we have obtained differences in results of MANOVA with threshold point of 0.5 and MANOVA with 21 different threshold points. The reason behind these differences is that results of comparisons change when one traverse different risk points instead of one risk point which is the property of ROC curves.Another reason is the weakness of sign test.

We also use Precision and Recall metrics for analyzing the Precision and Recall Operating Characteristics. In Table 7.6 the dimension results of MANOVA are given for threshold point of 0.5. In Table 7.7, number of wins after post-hoc test is given. It can be seen that *NB* is worst. Results of two-tailed Binomial Sign test are shown in Table 7.3. There are significant differences between *NB* and others as we expected. The results of number of wins for 21 threshold points are shown in Table 7.9. However, it can be seen from the results of two-tailed Binomial Sign test given in Table 7.10 that there

is only a significant difference between *NB* and *Ripper*. Again global decision property of using different threshold points gives different results than using one threshold point.

Table 7.1. Dimension decision of MANOVA for threshold point of 0.5

dataset	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
dimension	4	2	1	2	2	3	0	3	3	4	3	2	3	3	4

Table 7.2. Number of wins calculated using MANOVA for the threshold point of 0.5

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*	total
k-NN	0	8	7	7	14	14
C4.5	5	0	3	6	13	14
Ripper	6	6	0	7	14	14
LP	7	8	7	0	14	14
NB	0	1	0	0	0	1

Table 7.3. Result of two-tailed Binomial Sign test for the threshold point of 0.5

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	0	0	1
C4.5	0	0	0	0	1
Ripper	0	0	0	0	1
LP	0	0	0	0	1
NB	1	1	1	1	0

Table 7.4. Number of wins calculated using MANOVA and Binomial Sign test for 21 threshold points

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	0	2	3
C4.5	2	0	1	4	6
Ripper	3	1	0	4	5
LP	2	1	0	0	5
NB	0	0	0	0	0

Table 7.5. Result of two-tailed Binomial Sign test for 21 threshold points

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	0	0	0
C4.5	0	0	0	0	0
Ripper	0	0	0	0	0
LP	0	0	0	0	0
NB	0	0	0	0	0

Table 7.6. Dimension decision of MANOVA using Precision and Recall for threshold point of 0.5

dataset	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
dimension	2	2	1	2	2	2	1	2	2	2	2	2	2	2	2

Table 7.7. Number of wins calculated using MANOVA using Precision and Recall for the threshold point of 0.5

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*	total
k-NN	0	7	9	7	15	15
C4.5	7	0	4	7	14	15
Ripper	5	5	0	7	14	15
LP	7	8	8	0	14	15
NB	0	1	1	1	0	11

Table 7.8. Result of two-tailed Binomial Sign test using Precision and Recall for the threshold point of 0.5

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	0	0	1
C4.5	0	0	0	0	1
Ripper	0	0	0	0	1
LP	0	0	0	0	1
NB	1	1	1	1	0

Table 7.9. Number of wins calculated using MANOVA and Binomial Sign test using Precision and Recall for 21 threshold points

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	5	5	5	11
C4.5	3	0	2	6	11
Ripper	5	4	0	6	13
LP	5	5	1	0	11
NB	0	1	0	1	0

Table 7.10. Result of two-tailed Binomial Sign test using Precision and Recall for 21 threshold points

	k-NN	*C4.5*	*Ripper*	*LP*	*NB*
k-NN	0	0	0	0	0
C4.5	0	0	0	0	0
Ripper	0	0	0	0	1
LP	0	0	0	0	0
NB	0	0	1	0	0

7.3. SUMMARY

The test procedure can be complicated when different datasets, different threshold values and different performance metrics are used. Therefore, we give a pseudocode for our proposed procedures in Figure 7.1. Comparison type is taken as a parameter, then after making k-fold cross validation on the dataset and trainining and testing our classifiers, comparison is performed according to the type of the comparison. The type of comparison is composed of three parts. If different performance metrics are used, the comparison is a multivariate comparison. Otherwise, the comparison is a univariate comparison. Two classifiers or multiple classifiers can be compared. The classifiers can be compared over one dataset or multiple datasets.

```
1  Results StatisticalComparison(ComparisonType)
2      Make k-fold cross validation on the dataset.
3      switch ComparisonType
4      case univariate, two classifiers, one dataset
5          Paired t test to the values of error, AUC and AUC-PR.
6      case univariate, multiple classifiers, one dataset
7          ANOVA with blocking factor to the values of error, AUC and AUC-PR.
8          Tukey test for pairwise comparisons.
9      case multivariate, multiple classifiers, one dataset
10         MANOVA with blocking to the TP,FP,TN,FN and Precision, Recall.
11         Two-sample Hotelling's T^2 test with Bonferroni correction.
12     case univariate, multiple classifiers, multiple datasets
13         ANOVA with blocking factor to the values of error, AUC and AUC-PR.
14         Tukey test for pairwise comparisons.
15         Calculate total risk values, compare and count wins, losses and ties.
16         Binomial Sign test with Bonferroni correction to wins, losses and ties.
17     case  multivariate, multiple classifiers, multiple datasets
18         MANOVA with blocking to the TP,FP,TN,FN and Precision, Recall
19         Two-sample Hotelling's T^2 test with Bonferroni correction.
20         Calculate total risk values, compare and count wins, losses and ties.
21         Binomial Sign test with Bonferroni correction to wins, losses and ties.
22     case multivariate, multiple classifiers, multiple datasets, different thresholds
23         MANOVA with blocking to the TP,FP,TN,FN and Precision, Recall.
24         Two-sample Hotelling's T^2 test with Bonferroni correction.
25         For each dataset, right tailed Binomial Sign test over all thresholds.
26         Sign test with Bonferroni correction to total grid over all datasets.
```

Figure 7.1. Pseudocode

8. CONCLUSION

It has been known that the ROC curve or the AUC value gives more information than the misclassification error [4], but still, most of the tests in literature use misclassification error.

In this work, we propose a novel statistical comparison procedure based on AUC of the ROC curves. To check for significant difference (unaffected by randomness), for each classifier, we use k-fold cross validation to construct multiple ROC curves and calculate an AUC value for each. We then use the paired t test to test hypotheses on such AUC distributions.

To validate our test, we compare it with the paired t test on misclassification errors. We see that our AUC test and the one using error give consistent decisions on a high proportion of cases. When they disagree, we believe that the one using AUC values are more to be trusted because they compare under a set of possible losses and not just a single one of equal loss for false positives and false negatives.

Both the error test and our test use the central limit theorem which states that the sum of a large number of iid random variables (the Bernoulli random variables corresponding to 0/1 decisions on test instances) is approximately normal. We see in practice that the distributions for error or AUC are sometimes not normal, probably due to dependence between folds which share data and the fact that 30 is a relatively small number for central limit theorem to hold.

For obtaining a comparison for different threshold points, we also use Precision-Recall curves which are the alternatives to ROC curves. After applying k-fold paired t test to AUC-PR, we observe that error makes a local decision and AUC-PR makes a global decision like AUC. However, we see that a classifier can be better according to AUC and can be worse according to AUC-PR. We argue that although there is a correspondence between ROC and PR curves, that is, one curve dominates the other

in ROC space, if and only if it dominates other one in also PR space [16], we can not guarantee the magnitude of this domination. The reason behind this situation is that different metrics constitude the ROC and PR curves. It has been argued that one can not assume the interval between points in PR curve as linear since Precision does not change linearly as Recall changes and one should take into account the skewness in interpolation [16]. Including skewness of PR curves in AUC-PR calculation can be a future work.

Instead of making pairwise comparisons, multiple comparisons of classifiers at the same time is also another problem. We use ANOVA for multiple comparisons. We use MANOVA for comparing classifiers by using different performance metrics at the same time. We observe that the multivariate test results in more reliable results since it takes into account the correlation of different metrics.

We use AUC and error as univariate performance metrics and compare multiple datasets over multiple classifiers using ANOVA and Binomial Sign test and compare it with the results obtained by applying Friedman test and post-hoc test of Nemenyi. We observe that Nemenyi test and Sign test have low powers and also AUC and error have different test results. The reason behind this difference is the difference we indicated in AUC-error metric comparisons. Since AUC makes a global decision and error makes a local decision they can have different decisions. We also use AUC-PR as an alternative to AUC, we obtain different results than the results of AUC and error.

We also try to simulate the behavior of the ROC space by applying MANOVA at different threshold points and compare it with the test results of MANOVA with the threshold point of 0.5 and obtain differences in their decisions because of risk conditions. Therefore we use both different metrics and different threshold points which is a desirable scenario. As we expected, results of MANOVA with the threshold point of 0.5 and MANOVA with different threshold points are different. We also apply MANOVA with threshold point of 0.5 and MANOVA at different threshold points to Precision and Recall metrics to observe the PR curve behaviour.

We also propose novel methods for comparing multiple classifiers and comparing multiple classifiers over multiple datasets. Taking into account the correlation of metrics and also different threshold points makes the statistical comparison more robust. We use Sign test for comparing classifiers, however Bonferroni correction and low power of this statistical test prevents obtaining the statistical differences. It can be replaced with a more powerful test as a future work.

Since Sign Rank test takes into account the magnitudes of the differences and Binomial test does not [27], we have also tried Sign Rank test, however it does not detect the differences again. A special test can be designed for this case.

MANOVA and ANOVA are valid if their assumptions are met. We use nonparametric version of ANOVA which is Friedman test. However, since its post-hoc test is a weak test like Sign test, it is not preferable. It is known that nonparametric versions of parametric tests should not be used if the assumptions are met, because parametric tests performs better than nonparametric ones when assumptions are met [2]. The assumptions of ANOVA and MANOVA should be checked and powerful nonparametric versions of ANOVA and MANOVA can be applied as a future work.

APPENDIX A: CLASSIFICATION ALGORITHMS

LP is the linear perceptron includes one perceptron with inputs of features of the instance and an output which is weighted sum of the features. Then, softmax enables us using the posterior probabilities that is used in classification.

C4.5 is the decision tree algorithm. At each node, the attribute that provides the maximum information gain is selected for splitting and childrens are crated. The decision tree is constructed recursively with the same criteria.

k-NN is k nearest neighbor is the classifier that assigns class of the instance to the class that is most frequently occured in its k nearest neighbors that have minimum Euclidean distance.

NB is the naive Bayes that is a case of the discriminant function with normally distributed class-conditional densities which assumes that variables are independent.

Ripper is an acronym for Repeated Incremental Pruning to Produce Error Reduction and it is the decision tree algorithm that optimizes an initial rule set that is constructed using IREP. For each rule in the initial rule set, the algorithm constructs two rules that are alternatives to that rule. For each rule, it performs a *replacement* after growing an prunning, prunning is done for minimizing error of overall rule set. *Revision* is done in greedy manner. MDL heuristic is used to select from three alternatives of replacement, revision or original. After this optimization step, IREP is used again for addition of rules.

APPENDIX B: DATASETS

Table B.1. Properties of datasets used

Datasets	Number of Features	Datasize
aibocolor	3	71962
chess	6	6451
connect-4	42	61108
mushroom	22	5234
nursery	8	8310
pageblock	10	5242
report	21	10878
shuttle	9	34240
spambase	57	4601
thyroid	27	2785
wave	21	3353
ada	48	4147
caravan	85	9822
gina	970	3153
sylva	216	13086

REFERENCES

1. Dietterich, T. G., "Approximate Statistical Tests for Comparing Supervised Classification Learning Classifiers", *Neural Computation*, Vol. 10, pp. 1895–1923, 1998.

2. Demsar, J., "Statistical Comparisons of Classifiers over Multiple Data Sets", *Journal of Machine Learning Research*, Vol. 7, pp. 1–30, 2006.

3. Alpaydın, E., "Combined 5×2 cv F Test for Comparing Supervised Classification Learning Classifiers", *Neural Computation*, Vol. 11, pp. 1975–1982, 1999.

4. Fawcett, T., "An introduction to ROC analysis", *Pattern Recognition Letters*, Vol. 27, pp. 861–874, 2006.

5. Provost, F. and T. Fawcett, "Robust Classification for Imprecise Environments", *Machine Learning*, Vol. 42, pp. 203–231, 2001.

6. Hanley, J. A. and B. J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve", *Radiology*, Vol. 143, pp. 29–36, 1982.

7. Ling, C. X., J. Huang and H. Zhang, "AUC: a Better Measure than Accuracy in Comparing Learning Algorithms", *Proceedings of International Joint Conferences on Artificial Intelligence (2003)*, Springer, pp. 329–341, 2003.

8. Huang, J., J. Lu and C.X. Ling, "Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy", *Proceedings of the Third IEEE International Conference on Data Mining*, pp. 553-556, 2003.

9. Bradley, A. P., "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition*, Vol. 30, pp. 1145–1159, 1997.

10. Bravo, H. C., G. Wahba, K. E. Lee, B. E. K. Klein, R. Klein and S. K. Iyengar, "Examining the relative influence of familial, genetic, and environmental covariate

information in flexible risk models", *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, Vol. 106, pp. 8128-8133, 2009.

11. Cortes, C. and M. Mohri, "Confidence Intervals for the Area under the ROC Curve", *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, Vol. 17, pp. 305–312, Vancouver, Canada, 2004.

12. Agarwal, S., T. Graepel, R. Herbrich and D. Roth, "Generalization Bounds for the Area Under the ROC Curve", *Journal of Machine Learning Research*, Vol. 6, pp. 393–425, 2005.

13. Weiss, G. M. and F. Provost, "Learning when Training Data are Costly: The Effect of Class Distribution on Tree Induction", *Journal of Artificial Intelligence Research*, Vol. 19, pp. 315–354, 2003.

14. Folleco, A., T. M. Khoshgoftaar and A. Napolitano, "Comparison of Four Performance Metrics for Evaluating Sampling Techniques for Low Quality Class-Imbalanced Data", *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications*, Vol. 00, pp. 153-158, 2008.

15. Bloedorn, E., I. Mani and T. R. Macmillan, "Machine Learning of User Profiles: Representational Issues", *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, AAAI/MIT Press, pp. 433–438, 1996.

16. Davis, J. and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves", *Proceedings of the 23rd international conference on Machine learning*, Vol. 148, pp. 233 - 240, 2006.

17. Landgrebe, T. C. W, P. Paclik and R. P. W. Duin, "Precision-recall operating characteristic (P-ROC) curves in imprecise environments", *Proceedings of the 18th International Conference on Pattern Recognition*, Vol. 04, pp. 123 - 127, 2006.

18. Clémençon, S. and N. Vayatis, "Nonparametric Estimation of the Precision-Recall

Curve", *Proceedings of the 26th Annual International Conference on Machine Learning*, Vol. 382, pp. 185-192, 2009.

19. Raghavan, V. V., G. S. Jung and P. Bollmann, "A critical investigation of recall and precision as measures of retrieval system performance", *ACM Transactions on Information Systems*, Vol. 7, pp. 205–229, 1989.

20. Montgomery, D. C., *Design and Analysis of Experiments*, John Wiley and Sons, 2008.

21. Mardia, K. V., J. T. Kent and J. M. Bibby, *Multivariate Analysis*, Academic Press, 1979.

22. Guyon, I., A. R. S. Azar, G. Dror and G. Cawley, *Agnostic Learning vs. Prior Knowledge Challenge & Data Representation Discovery Workshop, International Joint Conference on Neural Networks 2007*, http://www.agnostic.inf.ethz.ch/datasets.php, Florida, 2007.

23. Yıldız, O. T., *ISELL Machine Learning Open Source Software*, Işık University.

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

25. Cohen, W. W., "Fast Effective Rule Induction", *The Twelfth International Conference on Machine Learning*, pp. 115–123, 1995.

26. Hanley, J. A. and B. J. McNeil, "A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases", *Radiology*, Vol. 148, pp. 839–843, 1983.

27. Sheskin, D. J., *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall, 2000.