JMLR: Workshop and Conference Proceedings 8: 65-81 Machine Learning in Systems Biology

Accuracy-Rejection Curves (ARCs) for Comparing Classification Methods with a Reject Option

Malik Sajjad Ahmed Nadeem MSAJJADNADEEM@GMAIL.COM
INSERM U872 Equipe 7CRC
Universite Pierre et Marie Curie(UPMC)
15 rue des Ecole de Medecine, 75005 Paris, France

LIM&BIO, UFR-SMBH
Universite Paris 13
74 rue Marcel Cachin, 93017 Bobigny, France

Jean-Daniel Zucker JEAN-DANIEL.ZUCKER@IRD.FR
IRD, UMI 209, UMMISCO
IRD France Nord
F-93143, Bondy, France

INSERM U872 Equipe 7CRC
Universite Pierre et Marie Curie(UPMC)
15 rue des Ecole de Medecine, 75005 Paris, France

LIM&BIO, UFR-SMBH
Universite Paris 13
74 rue Marcel Cachin, 93017 Bobigny, France

Blaise Hanczar HANCZAR_BLAISE@YAHOO.FR
LIPADE
Universite Paris Descartes
45 rue des Saint-Peres, 75006, Paris, France

Editor: Sašo Džeroski, Pierre Geurts, and Juho Rousu

Abstract

Data extracted from microarrays are now considered an important source of knowledge about various diseases. Several studies based on microarray data and the use of receiver operating characteristics (ROC) graphs have compared supervised machine learning approaches. These comparisons are based on classification schemes in which all samples are classified, regardless of the degree of confidence associated with the classification of a particular sample on the basis of a given classifier. In the domain of healthcare, it is safer to refrain from classifying a sample if the confidence assigned to the classification is not high enough, rather than classifying all samples even if confidence is low. We describe an approach in which the performance of different classifiers is compared, with the possibility of rejection, based on several reject areas. Using a tradeoff between accuracy and rejection, we propose the use of accuracy-rejection curves (ARCs) and three types of relationship between ARCs for comparisons of the ARCs of two classifiers. Empirical results based on purely synthetic data, semi-synthetic data (generated from real data obtained from patients) and public microarray data for binary classification problems demonstrate the efficacy of this method.

Keywords: Classifier Comparison, Reject Option, Microarray

1. Introduction

Microarray techniques are becoming increasing popular, as they provide researchers and doctors with information about the level of gene expression in biological tissue samples. Microarray systems simultaneously measure the levels of mRNA or thousands of genes in a cell mixture, at a given time and in given environmental conditions. Microarrays are used in many fields of medical research. One of the most widespread applications of this technology is for the prediction of a biological parameter from a gene-expression profile. For example, by comparing the expression profiles of different tissue types, it is possible to predict various biological parameters, such as the occurrence of different types of tumors with different outcomes (Alon et al., 1999; Golub et al., 1999; Shipp et al., 2002), survival times after treatment for cancer patients (Kaderali et al., 2006; Temanni et al., 2007), weight loss after a restrictive diet or bariatric surgery, facilitating selection of the most appropriate treatment, as shown by Dudoit et al. (2002); Braga-Neto and Dougherty (2004); Wang et al. (2007). One of the key characteristics of this kind of data is the huge disparity between the number of examples (generally 10 to 100 samples per microarray) and the number of features studied (several thousands of genes). This constitutes a major challenge for classical machine learning algorithms.

A large number of machine learning methods have been successfully applied to microarray classification: diagonal linear discriminant analysis (DLDA), k-nearest neighbors (Dudoit et al., 2002), support vector machines (Furey et al., 2000) and random forests (Breiman, 2001), for example. The performances of these classifiers are measured by the accuracy with which they predict the class of the examples. The true accuracy of the classifiers cannot be calculated, because the feature-label distribution is unknown. We therefore have to rely on estimates of this accuracy. Given the small number of available examples, this estimation is based on re-sampling procedures, such as cross-validation or bootstrapping. The aim is to identify the best classifier for microarray-based classification. Several comparative studies dealing with methods of microarray classification have been published. Man et al. (2004) claimed that the support vector machine (SVM) and partial least squares discriminant analysis (PLS-DA) methods were the most accurate. Dudoit et al. (2002) showed that simple methods, such as DLDA and k-nearest neighbors gave good results, whereas Lee et al. (2005) conclude that the SVM method was superior. The conclusion of Huang et al. (2005) that no one classifier is systematically better than all the others, is probably the most reliable. All these studies were based on comparisons of the accuracy or error rates of classifiers.

In all these previous publications, the classifier gave a prediction for all the samples. However, for practical applications in medicine, it is better for predictions to be made only when they are sufficiently reliable, with no prediction made by the classifier in other cases. In practice, this results in the addition of a "reject" option to the classifiers. If the category to which an example should be classified cannot be predicted reliably, the classifier rejects the observation by not assigning it any of the class labels. The reject option introduced

by Chow (1970) results in decisions not being taken for samples for which confidence is lowest, to reduce the likelihood of error. Friedel et al. (2006) incorporated a reject option into their methods for improving the prediction accuracy of classifiers. They demonstrated that the inclusion of this option considerably increased the prediction accuracy of classifiers. Hanczar and Dougherty (2008) studied classifiers with reject options for use in microarray classification. The performance of classifiers with reject options was found to depend on accuracy and rejection rate. No study comparing classifiers with reject options has yet been published. It is generally assumed that comparisons of classifiers with and without the reject option are equivalent. We tested this assumption and found it to be wrong. We developed a method for comparing the performances of classifiers in terms of their rejection rates, based on accuracy-rejection curves (ARCs). We assume that, for a given item of data, rejection has different impacts on the accuracy of different classifiers, the best classifier also depending on the rejection rate. Our ARC method compares the performances of classifiers by considering different rejection regions, ranging from 0% to 100%, to provide a complete picture of the effects of different rejection rates on the accuracies of the classifiers considered. Our experimental results, obtained with various sets of purely synthetic data, semi-synthetic data and public microarray data, show that the proposed comparison of different classifiers (with a reject option) makes it easier to select the best available classifier for a given set of data under given selection criteria (i.e. desired accuracy and/or acceptable rejection rate). We have identified three types of result in our simulations, two of which disprove the general assumption that the comparison of classifiers with and without reject options is equivalent.

2. Classification with Reject Option

Let us consider a binary classification problem in which each example belongs to one of two categories. The performance of a classifier is measured by its error rate. The classifier minimizing the error is called the Bayes classifier.

If the Bayes classifier is not sufficiently accurate for the task at hand, then a reject option approach can be used. The reject option introduced by Chow (1957) suggests that samples for which *a posteriori* probabilities are insufficiently high should not be classified, to reduce the likelihood of error. A rejection region is defined in the feature space and all examples within this region are rejected by the classifier. The classifier rejects an example if the prediction is not sufficiently reliable and falls into the rejection region. There is a general relationship between error and rejection rate: according to Chow (1970), the error rate decreases monotonically with increasing rejection rate. Based on this relationship, Chow proposed an optimal error versus rejection tradeoff only if the *a posteriori* probabilities are exactly known. Unfortunately, in real applications, these probabilities are affected by significant estimate error. In classifiers with a reject option, the key parameters are the thresholds defining the rejection regions. Landgrebe et al. (2006), Dubuisson and Masson. (1993), Hanczar and Dougherty (2008) and others have proposed a number of strategies for defining an optimal rejection rule. In this study, we did not deal with the problem of the optimal tradeoff between error and rejection. We used several different rejection regions

and calculated the resulting accuracies. We varied the size of the rejection region from 0% to 100%, by increments of 0.2%, resulting in the definition of 500 rejection regions. We then plotted the rejection regions against the accuracies obtained.

3. Comparing Classifiers with Reject Options

The performances of classifiers are assessed in terms of the accuracy with which they predict the true class. Several studies (some cited in the introduction) claim that certain classifiers are better than others. All previous comparative studies have been based on the error rates obtained for classifiers, but error rate is not the only measurement that can be used to judge a classifier's performance. Indeed, classifier performance also depends heavily on the data, as some classifiers (e.g. LDA, SVM-Linear etc.) perform better with linearly separable data than with non-linearly separable data, whereas the reverse is true for others (SVM-Radial, RF etc). Thus, different classifiers have different accuracies for classification of the very same data. With microarray data, it is often difficult to determine the nature of the data (linear, non-linear etc.), and there may even be a mixture of two or more types of data. So, for each classification task, a comparison study should be carried out to identify the best classifier. In cases of classification with a reject option, accuracy also depends on the rejection rate. It seems likely that different rates of rejection have different effects on the performance of different classifiers.

We describe here a method for comparing classifiers with reject options, by presenting the performances of classifiers on two-dimensional accuracy-rejection curves (ARCs).

Definition 1 (accuracy rejection curve (ARC)) *An accuracy rejection curve (ARC) is a function representing the accuracy of a classifier as a function of its rejection rate.*

An ARC is therefore produced by plotting the accuracy of a classifier against its rejection rate, varying from 0 to 1 (i.e. 100%). All ARCs have an accuracy of 100% for a rejection rate of 100%, and therefore converge on the point (1, 1). They start from a point (0, a), where a% is the percentage accuracy of the classifier when it does not reject any of the observations.

ARCs are useful in that they make it possible to compare graphically the accuracy of several classifiers as a function of their rejection rates. Let us assume that not all the classifiers respond similarly in different rejection regions for a given task of classification. Some have higher accuracies and lower rates of rejection than others. Based on this assumption, we identified three different types of relationship between the ARCs of two classifiers compared with each other: Type 1 ($T1$), Type 2 ($T2$) and Type 3 ($T3$), as illustrated in Figure 1.

1. $T1$ (Crossing-over) ARCs: If the ARC of one of the classifiers (say Cls_1) cuts the ARC of the other classifier (say Cls_2), as shown in Figure 1:A, this situation can be described as a "crossing-over" of ARCs. In Figure 1, we have plotted rejection rates on the x-axis and accuracies on the y-axis. Here, due to the crossing-over of the curves, Cls_1 outperforms Cls_2.

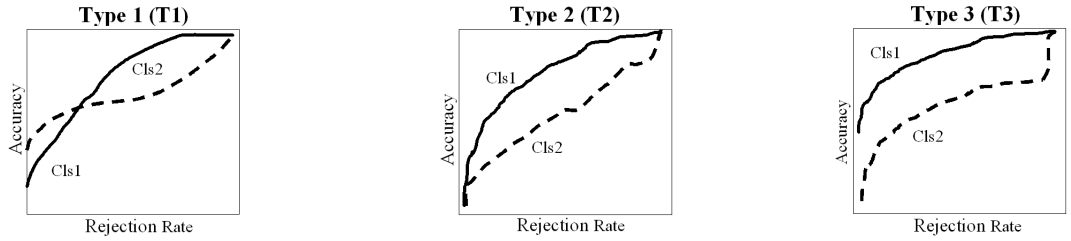

Figure 1: Illustration of the 3 possible relationships between accuracy-rejection curves (ARCs).

2. $T2$ (Diverging) ARCs: In diverging ARCs, the ARCs of two classifiers almost overlap at rejection rates at or close to zero. The ARCs separate as the rejection rates increase. For example, in Figure 1:B, at a rejection rate of zero, the ARCs of he classifiers Cls_1 and Cls_2 overlap. The separation of the ARCs increases with increasing rejection rate, making Cls_1 a better classifier than Cls_2 in this example.

3. $T3$ (Evenly spaced) ARCs: In this case, the distance between the two ARCs is very similar at different rejection rates, as in Figure 1:C.

For the sake of simplicity, we refer to these relationships between ARCs hereafter as $T1$, $T2$, and $T3$.

For selection of the best available classifier by the ARC method, it is necessary to know the desired accuracy, the acceptable rejection rate or both. If the desired accuracy is known, we move horizontally across the ARC plot and select the classifier with the lowest rejection rate. Conversely, if the acceptable rejection rate is known, we select the classifier with the highest prediction accuracy for that rejection rate.

4. Experimentation

We used published real data and two types of data that we generated ourselves: purely synthetic data and semi-synthetic data, which are generated using parameters based on published microarray data for real patients (i.e. (Alon et al., 1999): colon cancer data, (Shipp et al., 2002): lymphoid cancer data, (Golub et al., 1999): acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) data). We will then discuss the experimental design.

4.1 Data

Our experiments are based on three kinds of data:

- Purely synthetic data generated with user-defined parameters and Gaussian models.

Table 1: Parameters for experiments based on purely synthetic data.

Classification rule	R	LDA, QDA, SVM-Radial, SVM-Linear, RF
Training	n_{tr}	50, 100, 200
No. of Gaussians per class	G	1, 2
No. of clusters of features	B_{size}	1, 2, 4, 5, 10
Rejection region	r_{win}	0.002, 0.004, 0.006,..., 100.000

Table 2: Parameters for experiments based on semi-synthetic data from public microarray data sets (Alon et al., 1999; Shipp et al., 2002; Golub et al., 1999).

Classification rule	R	LDA, QDA, SVM-Radial, SVM-Linear, RF
Training	n_{tr}	50, 100, 200
No. of Gaussians per class	G	1, 2, 5
Rejection region	r_{win}	0.002, 0.004, 0.006,..., 100.000

- Semi-synthetic data generated with parameters estimated from real microarray data, using the expectation-maximization (EM) algorithm for microarray studies (Alon et al., 1999; Shipp et al., 2002; Golub et al., 1999).
- Real microarray data: (Alon et al., 1999; Shipp et al., 2002; van de Vijver, 2002).

4.1.1 Purely Synthetic and Semi-Synthetic Data

Experiments on real microarray data require the use of sampling methods to estimate the error rate and these methods are sometimes inaccurate for small samples, with synthetic data giving more accurate estimates of error (Hanczar et al., 2007). So, before using real microarray data, we used synthetic data sets in our experiments.

We considered two-class classification problems in which each class follows a Gaussian distribution. The classes are equally likely and the class-conditional densities are defined as: $N(\mu_1; \sigma_1 \sum)$, and $N(\mu_2; \sigma_2 \sum)$, where $\mu_1 = (-1, -1, -1, ...)$, and $\mu_2 = (1, 1, 1, ...)$. The covariance matrix $K$of each class is defined by $\sigma_i \sum$ where $\sum$ has a block structure.

$$K = \begin{bmatrix} \sum_{B_{size},\rho} & 0 & \cdots & 0 \\ 0 & \sum_{B_{size},\rho} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sum_{B_{size},\rho} \end{bmatrix} \quad where \quad \sum_{B_{size},\rho} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}.$$

Thus each feature is associated with a unique block. The correlation between two features in the same block is ρ, the correlation between two features from different blocks is 0. By varying the parameters of our model, we can construct different kinds of classification

problems (linear or non linear, with or without correlated features). For purely synthetic data, we chose the parameters of the model. Table 1 summarizes the parameters used to generate purely synthetic data. For semi-synthetic data, the parameters of the model were estimated from real microarray data, with the expectation maximization (EM) algorithm. Table 2 summarizes the parameters used to generate semi-syntetic data.

For each classification problem, we generated data with 20 features of interest, or "noise-free" features (d_{nf}). In real microarrays most of the genes are irrelevant for the classification task in hand, as shown by Golub et al. (1999); Furey et al. (2000); Zhou and Mao (2005) and Li et al. (2001). So, to make the analysis more realistic, we added 380 irrelevant or noise features ($d_{irrF} = 380$) to synthetic data sets ($D = d_{nf} + d_{irrF}$). A noise feature follows the same Gaussian distribution for the two classes $N(\mu; \sigma)$. The generated data G_{DATA} has NxD dimensions where N is the number of examples and D is the number of features, as defined above.

4.1.2 Real Microarray Data

We also applied this approach to three real microarray datasets. We used the lymphoid cancer data set (Shipp et al., 2002), in which the task was distinguishing patients with lymphoid cancers from non-patients. The dataset contains information for 58 patients and 19 non-patients. We also used the colon cancer dataset (Alon et al., 1999), which contains the genetic profiles of 39 patients with a colon cancer and 23 non-affected patients, and the breast cancer dataset (van de Vijver, 2002) containing 295 patients with breast cancer, 115 belonging of whom have a good prognosis and 180 of whom have a poor prognosis.

By contrast to the situation for purely synthetic and semi-synthetic data, no test set is available for the estimation of classifier performances with real datasets. We therefore carried out 10-fold cross-validation, an iterative procedure in which the data are randomly assigned to k subsets. During the i-th iteration, feature selection and model learning are carried out on the k-1 subsets not containing the i-th subset and the classifier is evaluated on the i-th subset. This procedure is then repeated 10 times. The final estimate is the mean of the results of the 100 iterations.

4.2 Experimental Design

For a classification task, we need training and test data for learning and testing purposes. We therefore generated class-labeled training data (ntr) and test data (nts).

The ntr contains 50, 100 or 200 samples with $D = d_{nf} + d_{irrF}$ features and nts is a Dx$10,000$ matrix where 10,000 is the sample size for the test data. Most of the features of a microarray are generally irrelevant. The irrelevant features are eliminated by feature selection (Guyon and Elisseeff, 2003; Li et al., 2006). As explained above, we rendered the purely synthetic and semi-synthetic data more like real microarray data (in which we encounter noise), by introducing noise (d_{irrF}) during the data generation process. Our purely

synthetic and semi-synthetic data were therefore very much like real microarray data. We then carried out feature selection, based on t-test statistics, and reduced the number of dimensions of the *nts* and *ntr* by selecting 20 relevant features. The *ntr* was then used to train each of the classifiers, SVM-Radial, SVM-Linear, LDA, QDA and RF. Once the training process had been completed, we assessed the performances of the trained models with the *nts*. The *a posteriori* probabilities were obtained and used as the outcome. These *a posteriori* probabilities were then used to calculate accuracy as a function of rejection rate. The range of *a posteriori* probabilities (from 0 to 1) was subdivided into 500 rejection regions. This process was repeated 100 times to obtain the final mean ARCs.

For real microarray data, we used 10-fold cross-validation, repeated 10 times to obtain a mean ARCs. During the feature selection process, we selected 10%, 50, 100 and 200 relevant genes for the experiments.

5. Results for the Purely Synthetic and Semi-Synthetic Datasets

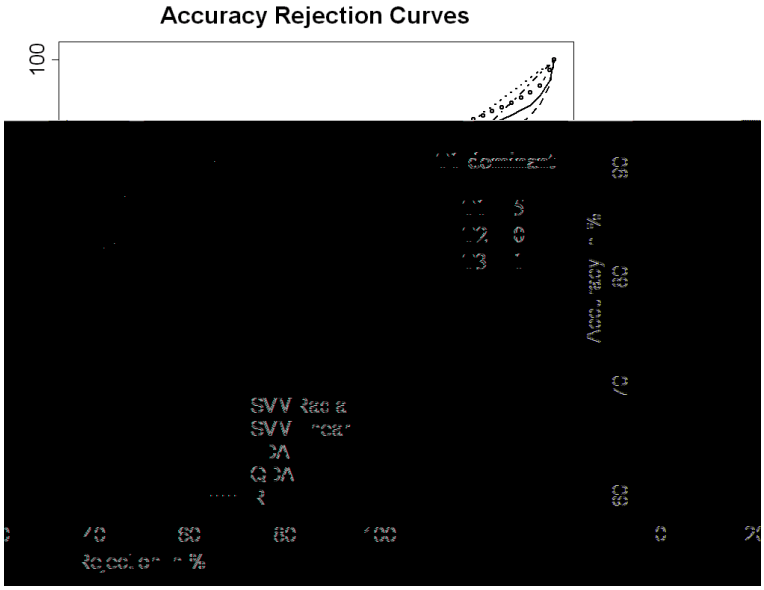

Figure 2: Accuracy-rejection curves with linear, non-correlated data, with 1 Gaussian distribution per class, a training set of 50 examples and a test set of 10,000 examples.

This section focuses on the results of experiments in which we used purely synthetic and semi-synthetic data. The experiments on synthetic data gave very accurate estimates of error and rejection rates, due to large numbers of training and test samples(Hanczar et al., 2007).

In each of the figures below, we have plotted mean rejection rate against mean accuracy for all the classification rules (as listed in Table:1 & 2) and for one of the data sets. We present some typical results. In the plots, solid lines show the ARC of SVM with a radial

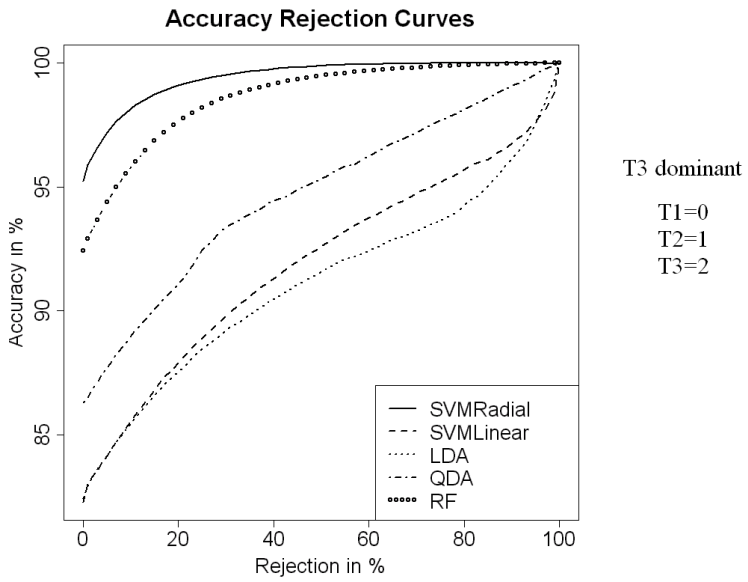

Figure 3: Accuracy-rejection curves for non-linear, correlated data with 1 Gaussian distribution per class, a training data set of 100 examples and a test data set of 10,000 examples. .

kernel, dashed lines show that of SVM with a linear kernel, dotted lines correspond to the ARC of LDA, dashed-dotted lines indicate the ARC for QDA, and lines of circles indicate the ARC for RF.

In Figure:2, SVM-Radial without rejection (0% rejection) has an accuracy of about 87% and RF without rejection (0% rejection) has an accuracy of 85%. If a 50% rejection rate is selected, RF becomes a better classifier than SVM-Radial. Interestingly, in Figure:2, with a rejection rate of 45%, LDA and SVM-Linear have similar accuracies. However, at rejection rates greater than 45%, LDA outperforms SVM-Linear. At a rejection rate of zero, LDA and SVM-Linear have almost identical accuracies. At rejection rates between about 3% and 45%, SVM-Linear performs better than LDA.

Figure:3 shows the similar accuracy of LDA and SVM-Linear from 0% to 18% rejection. However, from a rejection rate of 19% onwards, SVM-Linear outperforms LDA.

A comparison of LDA and SVM-Radial (Figure:4) showed that the ARCs crossed over, indicating that LDA outperformed SVM-Radial. In addition, an evaluation of the performances of QDA and SVM-Radial showed that, QDA outperformed SVM-Radial if there was a reject option.

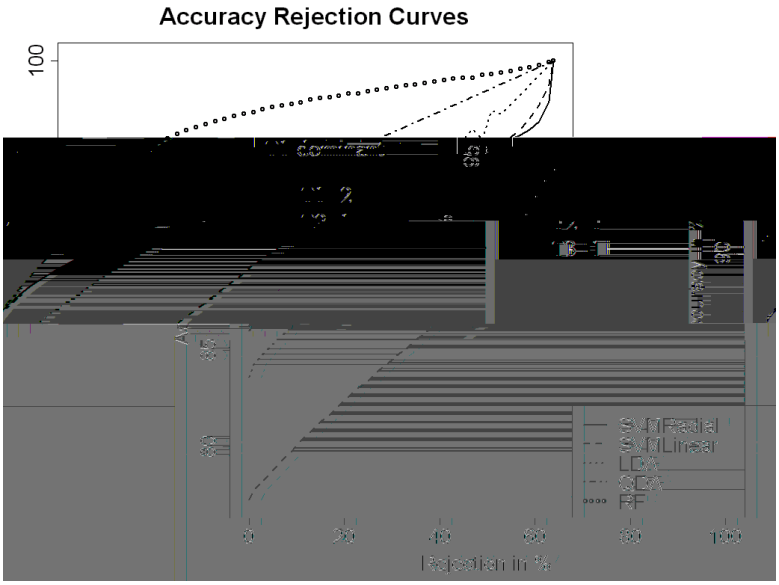

Figure 4: Accuracy-rejection curves based on synthetic data generated from the colon cancer dataset with 5 Gaussian distributions per class, a training dataset of 200 examples and a test dataset of 10,000 examples.

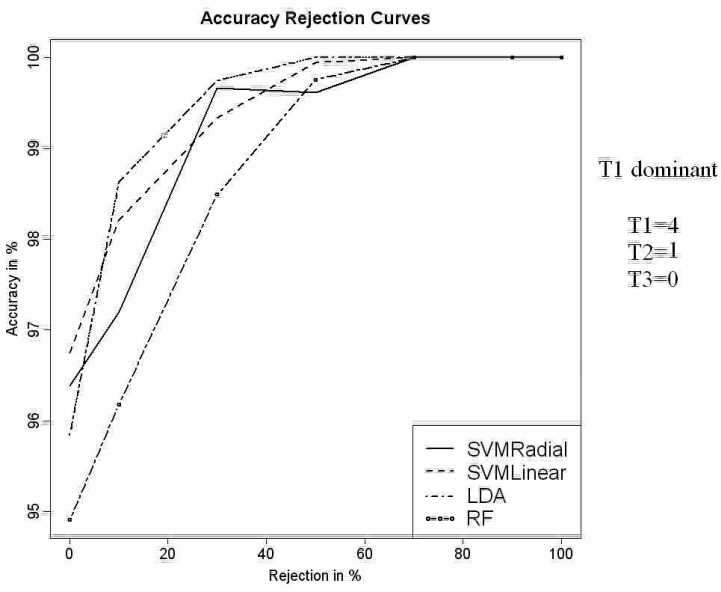

Figure 5: Accuracy-rejection curves based on lymphoid cancer microarray data (Shipp et al., 2002).

6. Results for Public Microarray Data-sets

We used breast cancer data (van de Vijver, 2002), lymphoid cancer data (Shipp et al., 2002), colon cancer data (Alon et al., 1999) as real microarray data-sets. We used 10-fold cross-

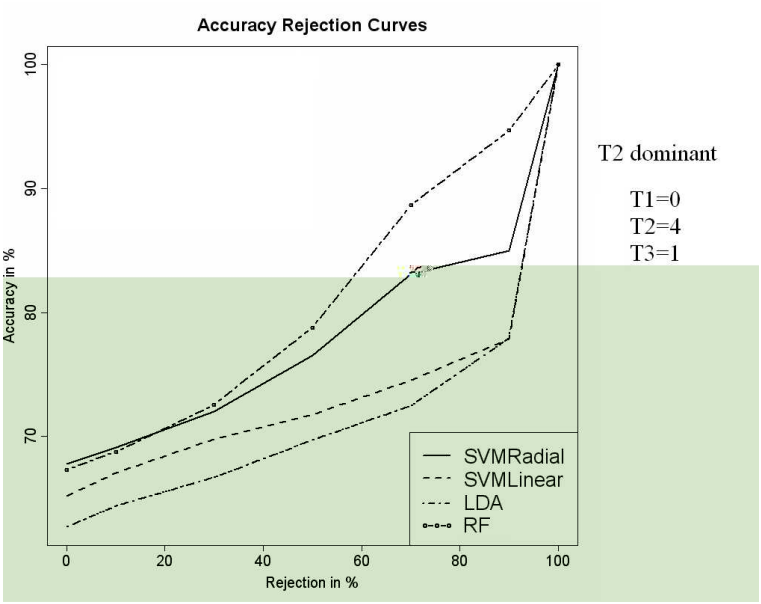

Figure 6: Accuracy-rejection curves based on breast cancer microarray data (van de Vijver, 2002).

validation, repeated 10 times, for these experiments. We used 10x10-fold cross-validation because only a few samples are available in real datasets. We also had to subdivide these samples into training and test data sets, further reducing the size of the dataset used for training. This problem is usually overcome by 10x10-fold cross-validation (i.e., 10-fold cross-validation carried out 10 times, with different random data partitions). Hence, each curve presented here is the mean of the 100 curves produced during this 10x10-fold cross-validation process.

Figure:5 shows the ARCs generated from lymphoid cancer data (Shipp et al., 2002). Four $T1$ type relationships were identified between the ARCs of different classifiers. For example, crossing-over of the ARCs of LDA & SVM-Radial resulted in LDA outperforming SVM-Radial. LDA also outperformed SVM-Linear due to crossing over of their ARCs. There is also a $T2$ type relationship between SVM-Radial & RF.

Figure:6 shows the ARCs for the breast cancer data-set (van de Vijver, 2002). In this figure, RF and SVM-Radial generate almost overlapping ARCs. At rejection rates above 20%, these two ARCs separate to form a $T2$ type ARC relationship. Moreover, comparison of the ARCs for SVM-Radial and SVM-Linear shows a similar $T2$ type relationship, but with no overlap. Instead, the two ARCs are separated from the outset, diverging further at rejection rates above 30%. Similar relationships are also observed between RF and SVM-Linear, RF and LDA and SVM-Radial and LDA. By contrast, there is a $T3$ type relationship between LDA and SVM-Linear.

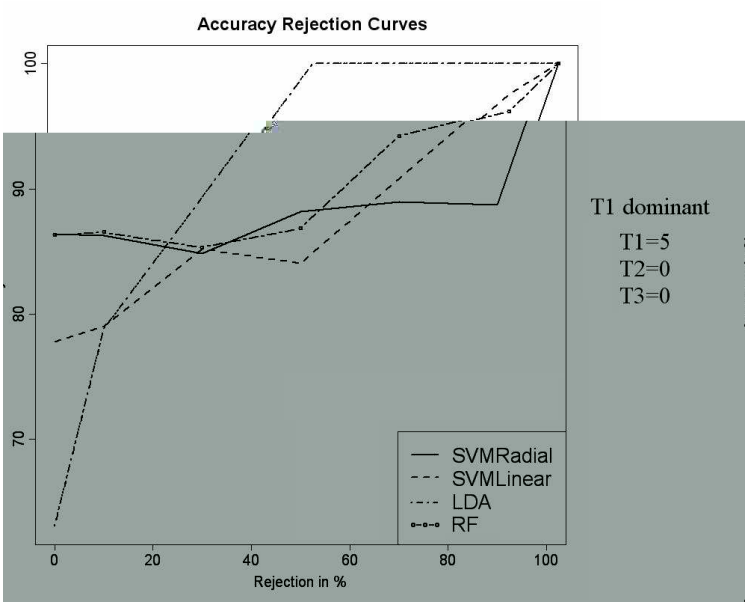

Figure 7: Accuracy-rejection curves based on colon cancer microarray data (Alon et al., 1999).

Figure:7 shows the ARCs obtained for the colon cancer data (Alon et al., 1999). Crossovers ($T1$) are observed between the ARCs of LDA and SVM-Radial, LDA and SVM-Linear, LDA and RF, SVM-Linear and SVM-Radial and between RF and SVM-Linear.

7. Discussion

All the results obtained indicate that accuracy increases with increasing rejection rate. However, the classification rules tested here did not respond identically to the incorporation of a reject option. The results illustrated in the figures above are interesting because they show that different classification rules respond differently at different rejection rates. In our study, some classification rules displayed a more rapid response, resulting in more accurate classification. The empirical results show that in most cases, one or more classification rules outperform the others.

All three types of relationship between ARCs proposed in the section Comparing Classifiers with Reject Options were observed. The identification of these different types of relationship is useful in several ways. First, when trying to select the most suitable classifier for a classification problem, if $T1$ curves are available, then the classifier outperforming the others should be preferred. Second, if T2 curves are obtained, the desired rejection limit should be chosen and the classifier performing best at the desired rejection rate should then be used. Third, if $T3$ curves are obtained, then, at a given rejection rate, the classifier with the best performance for the specific dataset on which the comparison was made should be used. In Tables 3 and 4, we summarize all 90 experiments based on these curve relationship

Table 3: Summary of curves obtained in experiments based on purely synthetic data, where $T1$, $T2$, and $T3$ are cases illustrated in Figure:1

Block size	Training Samples	No. of Gaussian distributions			
		1		2	
		$\sigma_1 = \sigma_2$	$\sigma_2 = \sigma_1/2$	$\sigma_1 = \sigma_2$	$\sigma_2 = \sigma_1/2$
1	50	$T1$ $T2$	$T3$	$T2$	$T2$
	100	$T1$	$T2$	$T1$ $T2$	$T1$
	200	$T1$ $T2$	$T2$	$T1$ $T2$	$T1$ $T2$
2	50	$T1$ $T2$	$T2$ $T3$	$T1$ $T2$	$T2$
	100	$T1$ $T2$	$T3$	$T1$ $T2$	$T1$
	200	$T1$	$T2$	$T1$ $T2$	$T2$
4	50	$T1$ $T2$	$T1$ $T2$	$T1$ $T3$	$T2$
	100	$T1$ $T2$	$T2$	$T1$ $T2$	$T1$
	200	$T2$	$T2$ $T3$	$T1$ $T2$	$T2$ $T3$
5	50	$T1$ $T2$	$T2$	$T1$ $T2$	$T1$
	100	$T1$ $T2$	$T2$	$T1$ $T2$	$T2$
	200	$T1$ $T2$	$T2$ $T3$	$T1$ $T2$	$T2$
10	50	$T2$	$T2$	$T1$ $T3$	$T1$ $T2$
	100	$T1$ $T2$	$T2$	$T1$ $T2$	$T2$
	200	$T1$ $T2$	$T1$ $T3$	$T1$ $T2$	$T2$
No Block (Non-Correl)	50	$T1$ $T2$	$T2$ $T3$	$T1$ $T2$	$T2$
	100	$T3$	$T2$	$T1$ $T2$	$T1$
	200	$T2$ $T3$	$T2$	$T2$	$T1$ $T2$

Table 4: Curves obtained in experiments based on data generated from real microarray data.

Data	Training Samples	No. of Gaussians		
		1	2	5
Golub	100	$T3$	$T3$	$T3$
	200	$T2$	$T3$	$T3$
Alon	100	$T2$	$T2$	$T1$
	200	$T1$ $T2$	$T3$	$T1$ $T2$
Shipp	100	$T3$	$T3$	*xxx
	200	$T3$	$T3$	*xxx

*xxx = No results available.

categories.

In experiments with only purely synthetic data, we observed, 72 experiments, that one or more of the classifiers outperformed the others in 40 instances (crossing over of the curves, category $T1$). There were also 59 situations in which, with or without rejection, two or more classifiers performed almost identically. However, as the rate of rejection increased, one of these classifiers improved in prediction accuracy more rapidly than the other (category $T2$). $T3$ curves were observed in only 12 cases.

In total of 90 experiments, one or more classifiers outperformed the others in 43 instances ($T1$ curves). We also obtained 64 $T2$ curves. The presence of larger numbers of $T2$ and $T1$ curves indicates that reject options are very advantageous in comparisons of classifiers, with the reject option generally leading to more optimal classifier selection. The 22 $T3$ curves indicate that rejection may sometimes have almost identical effects on the performances of classifiers, with no significant difference in performance found between the two classifiers concerned.

Empirical results based on purely synthetic, semi-synthetic and real microarray data demonstrate that the use of ARCs to compare classifiers with reject options is effective and simple. This approach could clearly facilitate the selection of more accurate classifiers.

8. Conclusion and Future Work

In this study, we have introduced the concept of accuracy-rejection curves (ARCs) for accurate representation of the performance of classifiers with reject options. We have identified three different types of relationship that may occur when ARCs are used to compare two classifiers. All three types ($T1$, $T2$ and $T3$) were observed in our empirical results.

We compared classifiers in a large number of experiments based on synthetic data, with 500 different rejection regions, ranging from 0.2% to 100% rejection. We used different parameter settings for purely synthetic data, to construct different kinds of classification problems (linear and non -linear, correlated and non -correlated features, with training sets). For synthetic data generated from data for real patients, the parameters of the model were determined from the real data. The large number of T1 and T2 relationships observed demonstrate the utility of using ARCs to compare the performances of different classifiers. The small number of type relationships observed shows that, although the inclusion of a reject option may result in no significant change in the performance of a classifier, this situation corresponded to a minority of the cases studied.

Many further studies on ARCs are possible, concerning, for example, a cost-benefit analysis of ARCs and studies of their usefulness in classifier comparison. This is an important area of investigation, because the costs of incorrect classification and rejection may be high. Moreover, the concept of the area under the curve (AUC) has proved useful in analyses of ROC curves. The area under the accuracy-rejection curve the AU-ARC, may prove simi-

larly useful.

In addition, for microarrays with very small sample sizes, the behavior of ARCs with bagging and boosting ought to be studied. In this study, we considered three possible types of behavior of pairs of ARCs. Certain refinements of this typology may be potentially useful. Furthermore, when comparing the performances of classifiers, the maximum rejection rate that may be considered acceptable remains to be determined. Moreover, it remains unclear how to obtain and optimal rejection area. Further investigations are required to address these issues.

Acknowledgments

We would like to thank the Government of France, High Education Commission (HEC) Government of Pakistan, *Societe Francaise d'Exportation des Ressources Educatives (SFERE), France*, and University of Azad Jammu & Kashmir, Muzaffarabad AJ&K, Pakistan for providing support for this research.

References

U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad Patterns of Gene Expression revealed by Clustering Analysis of Tumor and normal Colon Tissues probed by Oligonucleotide Arrays. *Proceedings of the National Academy of Sciences (PNAS) USA*, 96(12):6745–6750, 1999.

U. M. Braga-Neto and E. R. Dougherty. Is Cross-Validation valid for Small-Sample Microarray Classification? *Bioinformatics*, 20(3):374–380, 2004.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

C. K. Chow. An Optimum Character Recognition System using Decision Functions. *IRE Trans. on Electronic Computers*, EC-6:247–254, 1957.

C. K. Chow. On Optimum Error and Reject trade-off. *IEEE Trans. on Information Theory*, IT-16(1):41–46, 1970.

B. Dubuisson and M. Masson. A Statistical Decision Rule with Incomplete Knowledge about Classes. *Pattern Recognition*, 26(1):155–165, 1993.

S. Dudoit, J. Fridlyand, and T. Speed. Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data. *Journal of the American Statistical Association*, 97(457):77–87, 2002.

C. C. Friedel, U. Ruckert, and S. Kramer. Cost Curves for Abstaining Classifiers. In *Proceedings of the ICML 2006 workshop on ROC Analysis in Machine Learning*, pages 33–40, Pittsburgh, PA, 2006.

T. S. Furey, N. Christianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support Vector Machine Classification and validation of Cancer Tissue Samples using Microarray Expression Data. *Bioinformatics*, 16(10):906–914, 2000.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, 1999.

I. Guyon and A. Elisseeff. Gene Feature Extraction Using T-Test Statistics and Kernel Partial Least Squares. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

B. Hanczar and E. R. Dougherty. Classification with Reject Option in Gene Expression Data. *Bioinformatics*, 24(17):1889–1895, 2008.

B. Hanczar, J. Hua, and E. R. Dougherty. Decorrelation of the True and Estimated Classifier Errors in High-Dimensional Settings. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007:1–12, 2007.

X. Huang, W. Pan, S. Grindle, X. Han, Y. Chen, S. J. Park, L. W. Miller, and J. Hall. A Comparative Study of Discriminating Human Heart Failure Etiology using Gene Expression profiles. *BMC Bioinformatics*, 6:205–210, 2005.

L. Kaderali, T. Zander, U. Faigle, J. Wolf, J. L. Schultze, and R. Schrader. CASPAR: a Hierarchical Bayesian Approach to Predict Survival Times in Cancer from Gene Expression Data. *Bioinformatics*, 22(12):1495–1502, 2006.

T. C. W. Landgrebe, D. M. J. Tax, P. Paclk, and R. P. W. Duin. The Interaction Between Classification and Reject Performance for Distance-based Reject-Option Classifiers. *Pattern Recognition Letters*, 27(8):908–917, 2006.

J. W. Lee, J. B. Lee, M. Park, and S. H. Songa. An Extensive Comparison of Recent Classification Tools applied to Microarray Data. *Computational Statistics & Data Analysis*, 48(4):869–885, 2005.

L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen. Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method. *Bioinformatics*, 17(12):1131–1142, 2001.

S. Li, C. Liao, and J. T. Kwok. Gene Feature Extraction Using T-test Statistics and Kernel Partial Least Squares. *Lecture notes in computer science*, PT 3(4234):11–20, 2006.

M. Z. Man, G. Dyson, K. Johnson, and B. Liao. Evaluating Methods for Classifying Expression Data. *Journal of Biopharmaceutical Statistics*, 14(4):1065–1084, 2004.

M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mensirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub. Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning. *Nature Medicine*, 8(1):68–74, 2002.

M-R. Temanni, S. A. Nadeem, D. P. Berrar, and J-D. Zucker. Aggregating Abstaining and Delegating Classifiers For Improving Classification performance: An Application to Lung Cancer Survival Prediction. In *CAMDA07*, Valencia, Spain, 2007.

M. van de Vijver. Agene-Expression Signature as a Predictor of Survival in Breast Cancer. *The New England Journal of medicine*, 347(25):1999–2009, 2002.

L. Wang, F. Chu, and W. Xie. Accurate Cancer Classification Using Expression of Very Few Genes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4 (1):40–53, 2007.

X. Zhou and K. Z. Mao. LS Bound Based Gene Selection for DNA Microarray Data. *Bioinformatics*, 21(8):1559–1564, 2005.

Nadeem M.S.A., Zucker Jean-Daniel, Hanczar B. (2010)

Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option

Journal of Machine Learning Research : Workshop and Conference Proceedings, 8, 65-81

International Workshop on Machine Learning in Systems Biology, 3., Ljubljana (SLV), 2009/09/05-06. ISSN 1938-7228