BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 17 2008, pages 1889–1895
doi:10.1093/bioinformatics/btn349

Gene expression

Classification with reject option in gene expression data

Blaise Hanczar[1,2,3] and Edward R. Dougherty[1,4,*]

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA, [2]Laboratoire d'Informatique Medicale et Bioinformatique (Lim&Bio), Universite Paris 13, 93017 Bobigny, France, [3]INSERM, U872, Equipe7 nutriomique, centre de recherche des Cordelier, Université Pierre et Marie Curie, 75004 Paris, France and [4]Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA

Received on January 22, 2008; revised on July 2, 2008; accepted on July 8, 2008

Advance Access publication July 10, 2008

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: The classification methods typically used in bioinformatics classify all examples, even if the classification is ambiguous, for instance, when the example is close to the separating hyperplane in linear classification. For medical applications, it may be better to classify an example only when there is a sufficiently high degree of accuracy, rather than classify all examples with decent accuracy. Moreover, when all examples are classified, the classification rule has no control over the accuracy of the classifier; the algorithm just aims to produce a classifier with the smallest error rate possible. In our approach, we fix the accuracy of the classifier and thereby choose a desired risk of error.

Results: Our method consists of defining a rejection region in the feature space. This region contains the examples for which classification is ambiguous. These are rejected by the classifier. The accuracy of the classifier becomes a user-defined parameter of the classification rule. The task of the classification rule is to minimize the rejection region with the constraint that the error rate of the classifier be bounded by the chosen target error. This approach is also used in the feature-selection step. The results computed on both synthetic and real data show that classifier accuracy is significantly improved.

Availability: Companion Website. http://gsp.tamu.edu/Publications/rejectoption/

Contact: edward@ece.tamu.edu, hanczar_blaise@yahoo.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Microarrays provide simultaneous expression measurements for thousands of genes and are now used in many fields of medical research. One of the most promising applications is the prediction of a biological parameter based on the gene-expression profile. For example, expression profiles can be used to differentiate different types of tumors with different outcomes and thereby assist in the selection of a therapeutic treatment. This task consists of using a training microarray dataset to build a classifier with which to make a prediction for an unknown patient. Diverse methods from pattern recognition have been used: linear discriminant analysis (Dudoit *et al*., 2002), support vector machines (Furey *et al*., 2000), neural networks (Khan *et al*., 2001), etc. Even if these methods produce classifiers with a good accuracy, very often they are still insufficiently accurate to be used in medical applications. A diagnostic or a choice of therapeutic strategy must be based on a very high confidence classifier.

As typically applied in the context of gene-expression classification (for instance, in the previously cited works), classifiers classify all examples even if the classification is unsafe, for example when the example is close to the separating hyperplane. On the other hand, a physician confronted with ambiguous symptoms may refer the patient to another specialist instead of giving an unsafe diagnosis. If this concept is implemented in the classification model, then it may be more useful in practical medial application. For instance, in cancer treatment, knowing the type of cancer is a crucial factor to defining a efficient therapeutic strategy. A classifier with a 20% error rate in predicting the cancer type of an arbitrary patient may be useless. It can preferable to have a classifier that predicts the cancer type of only a part of the patients with a very high accuracy, with the other patients being handled by other techniques.

In this article, we recall the concept of classification with reject option based on Chow's theory. A rejection option is added to classical classification methods and determines whether a given example will be classified or rejected (not classified). Then we present our method of classification based on Chow's works (Chow, 1970) in the context of gene-expression data. The error rate of the classifier becomes a parameter of the classification rule that is chosen by the user. The learning task is to minimize the rate of rejection with respect to the given error rate. We show how to implement this kind of classifier in the context of wrapper feature selection. We test and show the usefulness of the proposed method on both artificial and real data.

2 THEORY OF CLASSIFICATION WITH REJECT OPTION

Consider a classification problem with two classes, $C=\{C_1,C_2\}$, where an example is characterized by a feature vector $x\in R^p$ and a label $y\in C$. The posterior probability is defined by the Bayes's formula:

$$p(C_i|x)=\frac{p(x|C_i)p(C_i)}{p(x)}=\frac{p(x|C_i)p(C_i)}{\sum_{i=1}^{2}p(x|C_i)p(C_i)}$$

*To whom correspondence should be addressed.

where $p(C_i)$ is the prior probability of class C_i, $p(x|C_i)$ is the conditional probability of x given C_i and $p(x)$ is the probability of x. A classifier is a function $f: R^p \rightarrow C$ which divides the feature space into two regions, R_1, R_2, one for each predicted class, such that $x \in R_i$ means that $f(x) = C_i$. The performance of a classifier is measure by its error rate,

$$\varepsilon[f] = p(f(x) \neq y) = \sum_{i=1}^{2} \int_{R_i} \sum_{j=1; j \neq i}^{2} p(x|C_j) p(C_j)\, dx$$

which is the probability of making an incorrect classification. The accuracy of a classifier is defined as the probability of making a correct decision.

$$a[f] = 1 - \varepsilon[f]$$

The classifier minimizing the error is called the Bayes classifier. It predicts the class having the highest posterior probability:

$$f_{Bayes}(x) = \arg\max_{C_i}(p(C_i|x))$$

It is not possible to obtain a better accuracy than with the Bayes classifier.

If the accuracy of the Bayes classifier is not sufficient for the task at hand, then one can take the approach not to classify all examples, but only the those for which the posterior probability is sufficiently high. Based on this principle, Chow (1970) presented an optimal classifier with reject option. A rejection region R_{reject} is defined in the feature space and all examples belonging to this region are rejected by the classifier. An example x is accepted only if the probability that x belongs to C_i is higher than or equal to a given probability threshold t:

$$f(x) = \begin{cases} \arg\max_{C_i}(p(C_i|x)) \text{ if } \max_{C_i}(p(C_i|x)) \geq t \\ \text{reject if } p(C_i|x) < t \ \forall i \end{cases}$$

The classifier rejects an example if the prediction is not sufficiently reliable. The rejection rate is the probability that the classifier rejects the example,

$$p(reject) = \int_{R_{reject}} p(x)\, dx = p(\max(p(C_i|x)) \leq t)$$

The acceptance rate is the probability that the classifier accepts an example,

$$p(accept) = 1 - p(reject)$$

In classification with reject option, we can define two types of error. The error, $\varepsilon[f]$, is the probability of making an incorrect classification. The conditional error,

$$\varepsilon^{cond}[f] = p(f(x) \neq y | accept)$$

is the probability of making an incorrect classification, given the classifier has accepted the example. We have the following basic properties:

$$\begin{aligned} p(accept) + p(reject) &= 1 \\ p(f(x) = y) + p(f(x) \neq y) + p(reject) &= 1 \\ p(f(x) = y | accept) + p(f(x) \neq y | accept) &= 1 \end{aligned}$$

There is a general relation between the error and rejection rate: the error rate decreases monotonically while the rejection rate increases (Chow, 1970). Based on this relation, Chow proposes an optimal error versus reject tradeoff.

In Chow's theory, an optimal classifier can be found only if the true posterior probabilities are known. This is rarely the case in practice. Fumera *et al*. (2000) show that Chow's rule does not perform well if a significant error in probability estimation is present. In this case, they claim that defining different thresholds for each class gives better results. The classification rule becomes:

$$f(x) = \begin{cases} \arg\max_{C_i}(p(C_i|x)) \text{if } \max_{C_i}(p(C_i|x)) \geq t_i \\ \text{reject if } p(C_i|x) < t_i \ \forall i \end{cases}$$

Although this kind of classifier is popular in the machine learning community, it is rarely used in microarray-based classification. Note that this method is close to the notion of soft classification. The main difference is that in soft classification, the posterior probabilities are the output of the classifier. In classification with reject option, a decision is made based on these posterior probabilities. The output of the classifier is a class or a rejection.

In classifier with rejection option, the key parameters are the thresholds t_i that define the reject areas. Several strategies have been proposed to find an optimal reject rule. Landgrebe *et al*. (2006) define 3D ROC curves for a classifier, where the axes represent the true positive rate, the false positive rate rejected by the classifier and the false positive rate accepted by the classifier. The optimal thresholds are chosen by maximizing the volume under the 2D surface. Dubuisson and Masson (1993) propose a rejection rule for problems where the classes are not well known. They include two rejection options: an ambiguity reject when an example is situated in the area between several classes and a distance reject for examples far from the samples of known classes. Li and Sehi (2006) propose to control the error instead of finding a trade-off between rejection and error rates. They reformulate the problem as: given an error rate for each class, design a classifier with the smallest rejection rate. Our approach is similar in that we propose to control the conditional error rate of the classifier, not the error.

3 IMPLEMENTATION FOR BIOINFORMATICS

In this section, we present our method of classification with reject option in the context of gene-expression-based classification. A classifier with reject option is composed of two elements: a classifier model and a set of thresholds. We explain this in the following sections and show how include this concept in the feature selection. In this article, we restrain our work to 2-class classification problems; multi-class problem will be studied in future works.

3.1 Classifier model

For binary classification, a classifier is a mapping $f: R^p \rightarrow \{0, 1\}$; however, a classifier can be defined via a discriminant function $d: R^p \rightarrow R$, where the sign of the function is used to predict the label of a given example: $d(x) \leq 0$ implies $f(x) = C_1$ and $d(x) > 0$ implies $f(x) = C_2$. By treating classification in this context, the distance, $|d(x)|$, of the output from the origin can be used to represent the confidence of the classification. Of interest in the present circumstance is that, whereas Chow's theory is defined using the posterior probabilities, it is not necessary to compute them to apply a rejection rule. The rejection region can be defined directly via $d(x)$.

Figure 1 illustrates two class-conditional densities for the discriminant, the density corresponding to C_i determining probabilities corresponding to $d(x)$ given C_i. The two vertical lines

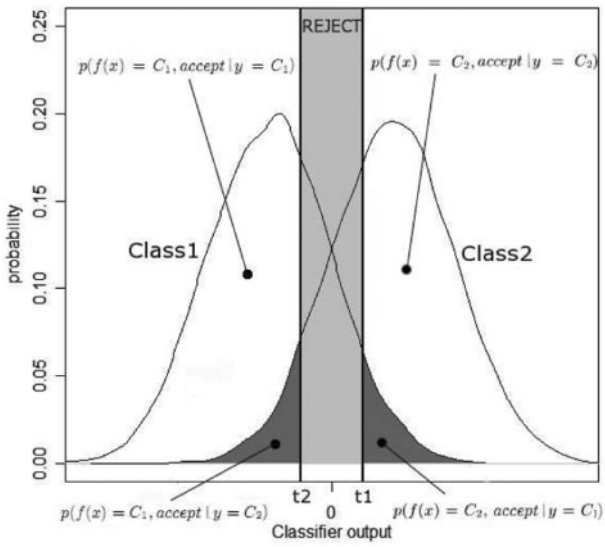

Fig. 1. Probability distribution of the two classes on the classifier output.

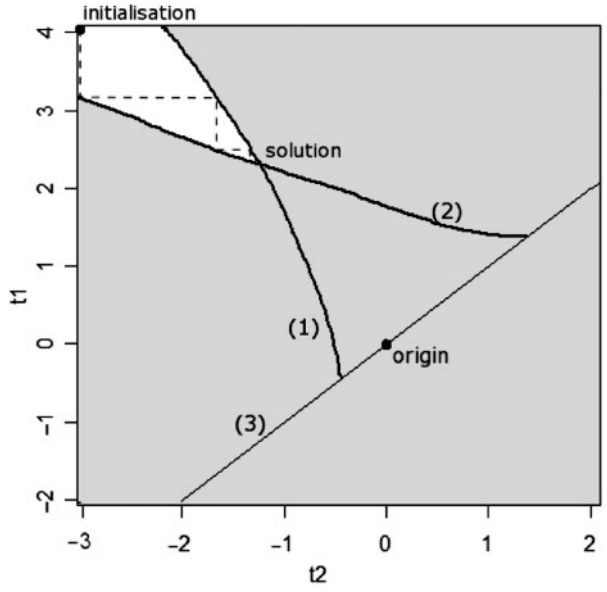

Fig. 2. Representation of the optimization problem. The two axis correspond the values of the two thresholds t_1 and t_2. The three lines (1), (2) and (3) represent the three constraints of the optimization problem. The white region represents the domain of validity. The dotted lines represent the heuristic search to find the optimal solution.

represent two thresholds, t_1 and t_2, the light gray area between t_1 and t_2 being the rejection region. The area to the left of t_2 is the region where examples are classified into the class C_1. In this region, the dark gray area represents the probability $p(f(x) = C_1, accept | y = C_2)$. We define the conditional error of class C_2 by

$$\varepsilon_2^{cond} = p(f(x) = C_1, y = C_2 | accept)$$

Equivalently,

$$\varepsilon_2^{cond} = \frac{p(f(x) = C_1, accept | y = C_2) p(Y = C_2)}{P(accept)}$$

which shows how the dark gray region gives the conditional error of class C_2. The conditional error ε_1^{cond} is defined analogously. Note that the conditional errors depend on both thresholds.

3.2 Threshold selection

The task is to select thresholds to define regions for the two classes and the rejection region. This choice determines the error reject trade-off. As seen in Section 2, several optimization strategies have been proposed. Our method is to fix a target condition error, ε_i^*, for each class. These conditional errors become parameters of the algorithm and the learning objective is not to minimize the error but to minimize the rejection rate under the constraints $\varepsilon_i^{cond} \leq \varepsilon_i^*$. If t_1 and t_2 are two thresholds, $t_2 < t_1$, then the problem can be formalized as an optimization problem with three constraints:

$$minimize\ (t_1 - t_2)$$
$$\begin{cases} (1)\ \varepsilon_1^{cond} \leq \varepsilon_1^* \\ (2)\ \varepsilon_2^{cond} \leq \varepsilon_2^* \\ (3)\ t_2 \leq t_1 \end{cases}$$

This minimization problem is represented by Figure 2. The two axes correspond to the values of the thresholds t_1 and t_2, and the three constraints are represented by the three lines, (1), (2) and (3). The domain of validity is represented by the white region. Minimizing $t_1 - t_2$ is equivalent to minimizing t_1 and maximizing t_2.

The solution is represented on the figure by the junction point of the lines (1) and (2). Note the bound of the constraint (3) corresponds to classifiers where $t_1 = t_2$, i.e. classifiers with no reject option. On this line, the origin corresponds to the regular classifier where there is a single threshold at 0.

We propose an iterative procedure to find the solution of this optimization problem. For a given value of ε_1^*, let the function $g_{\varepsilon_1^*}(t_1) = t_2$ [resp. $g_{\varepsilon_2^*}(t_2) = t_1$] gives the value of t_2 (resp. t_1) for any value of t_1(resp. t_2). t_1 and t_2 are initialized with their maximum and minimum values, respectively, in the search space, represented by the point on the upper left corner in Figure 2. We alternately minimize t_1 with respect to the constraint $\epsilon_1^{cond} \leq \epsilon_1^*$ and maximize t_2 with respect to the constraint $\epsilon_2^{cond} \leq \epsilon_2^*$. At the i-th iteration, the threshold pair is (t_1^i, t_2^i), and at the next iteration, $t_1^{i+1} = g_{\varepsilon_1^*}(t_2^i)$ and $t_2^{i+1} = g_{\varepsilon_2^*}(t_1^{i+1})$. This procedure is iterated until t_1 cannot be decreased and t_2 cannot be increased. The search is represented in Figure 2 by the dotted line.

Since the functions $g_{\varepsilon_1^*}(t_1) = t_2$ and $g_{\varepsilon_2^*}(t_2) = t_1$ are monotonely decreasing, the search converges to a unique solution, except in two special cases. First, when the domain of validity is empty there is no solution. It does not exist a classifier satisfying the constraints for the target errors. Second, there are several solutions, all solutions being of the type $t_1 = t_2$, meaning these solutions are correspond to classifiers with no reject option. In this case it not necessary to use a reject option; the regular classifier is sufficiently accurate to respect the target errors.

Resolving this minimization problem requires estimating the density probabilities of the two classes on the classifier output. This estimation is done by Gaussian kernel density estimation method (Silverman, 1986), the principle being to applied a Gaussian distribution on all points and sum all these distributions.

Since the conditional errors depend on the classifier, it is important to use different subsets to learn the classifier and to compute the thresholds; otherwise the probability estimates used for finding the thresholds will tend to be low-biased. This means that the training dataset, S_{train}, should be split into S_{model} and S_{thres}, with the classifier learned on S_{model} and then the thresholds constructed using S_{thres} and the learned classifier.

3.3 Feature selection

For feature selection we adapt sequential forward search (SFS) to classification with reject option. In the usual application of SFS, the features providing the lowest error rate are selected; however, in the reject scenario, the selection criterion is no longer the error rate but is instead the size of the rejection region. As the search proceeds, we select the feature providing the lowest rejection rate under the conditional error constraints $\varepsilon_1^{cond} \leq \varepsilon_1^*$ and $\varepsilon_2^{cond} \leq \varepsilon_2^*$. As we have previously noted, the threshold computation can fail when there is no solution to the optimization problem. If the selection of a feature leads to this case, then there is no classifier and this feature is directly removed from the potential selectable features for this iteration. For the next iteration, this feature will be tested again. In the case where all features lead to failed classifiers, the selection is done by selecting the feature that minimizes the error rate of the classifier with no reject option. This case may occur in the first iterations of the feature selection, when the information contained in the selected features does not allow the construction a classifier respecting the target error constraints.

4 RESULTS AND DISCUSSION

We present results showing the advantage of using a rejection option in classification and the limitations of this method. The experiments use both synthetic and real data. The experiments on synthetic data permit very accurate estimations of the error and rejection rates. The experiments on real data require the use of sampling methods to estimate the error rate and it has been shown that these methods are inaccurate for small-sample problems (Hanczar *et al*., 2007); nonetheless, we present them under this codicil to illustrate the method on real data, keeping in mind that, as always with small samples, the experiments using synthetic data are more definitive owing to better error estimation. In all experiments, we are interested only in the conditional error and to simplify the notation we will call this term the error. We assume that the target errors are equal: $\varepsilon^* = \varepsilon_1^* = \varepsilon_2^*$, and we compute only the total error rate ε. We compare our method with classifiers with no reject option and with classifiers using posterior probabilities. The classifier with posterior probabilities, described in Section 2, has a fixed pre-defined threshold. If the posterior probabilities are lower than this threshold, then the example is rejected. In the following sections, we present some representative results. Supplementary results and details on experimental design can be found in the companion website.

4.1 Synthetic data

The synthetic data are generated from real microarray dataset. We use three microarray datasets: colon, breast and lung cancer datasets, which are detailed in the next sections. A dataset is reduced to its 30 best genes, based on their *t*-test scores. Then Gaussian mixture models are fit for each of the two classes. $N/2$ and 5000 examples are, respectively, generated for each class to form the training and test set. Finally, 1970 noise features are added to the training and test sets. A noise feature is generated for the two classes from the same Gaussian distribution whose mean and standard error are of same order as the other features. Altogether, the synthetic data has two equally likely classes, a training set of N examples, a test set of 10 000 examples, 30 relevant features and 1970 irrelevant features. More details are presented on the companion website.

Figure 3 shows the results on synthetic data generated from a colon cancer dataset, $N = 200$, the classification rule is the a SVM with linear kernel. The dotted line corresponds to the error rate of the classifier with no reject option. In panel A, the full line represents the error rate of classifier with reject option whose the target error rate is 0.1. In panel B, the full line represents the error rate of classifier using the posterior probabilities whose threshold is to 0.1. The gray histogram represents the rejection rate whose scale is on the left axis. Up to 20 features are selected by the SFS procedure. In panel A, we see that with no reject option the error rate decreases during the first four iterations and then stays around 0.15. If we apply our algorithm with target error rate 0.1, the error is always around 0.1. The reject rate is around 0.6 with a minimum at 0.4 for three selected features. Note that in classification with no reject option the error rate begins to decrease strongly then increase slowly with the number of selected features, thereby exhibiting the peaking phenomenon (Hua *et al*., 2005). For the classifier with reject option, the error rate is stable around the target error for any number of selected features. It is interesting to note that the peaking phenomenon can be observed with the rejection rate, the optimal solution in the reject setting corresponding to the classifier that accepts the maximum of examples. In panel B, we see the error rate of classifier using posterior probabilities is between 0.07 and 0.08 and the rejection rate is higher than 0.75. Compared to our method, the classifier using posterior probabilities is more accurate but rejects more examples. The tradeoff error/rejection is better in our method because in both methods we respect the constraints ($error \leq 0.1$) but our method rejects less examples.

Another experiment using the colon cancer dataset has been done in which we vary the target error rate. We use the same parameters as in the previous experiment except that the number of selected features is fixed to 10. The classifier with no reject option still produces an error rate of 0.15. We construct classifiers with reject option with different target error rates. The results are presented in the Figure 4. We see that the errors of classifiers are very close to the target error, meaning that the constraint on target error is respected. The rejection rate decreases as the target error increases, going from 0.91 for $\varepsilon^* = 0.05$ to 0.12 for $\varepsilon^* = 0.15$. Increasing the target error makes the problem easier, the threshold region decreases, and more examples are accepted.

At the last point of the figure ($\varepsilon^* = 0.15$), the target error is the same as the error of the classifier with no reject option, which has been found by directly applying the classification rule. One might expect that in this situation there would be no rejection area and all the examples would be accepted. This is not the case: 12% of the examples are rejected, even through the classifier with no reject option shows that it would be possible to classify all examples at the target error. This apparent anomaly occurs because for the classifier with reject option the training data have been evenly split into two sets, one for model learning and the other for threshold computation. That means the classification rule is applied on only the half of

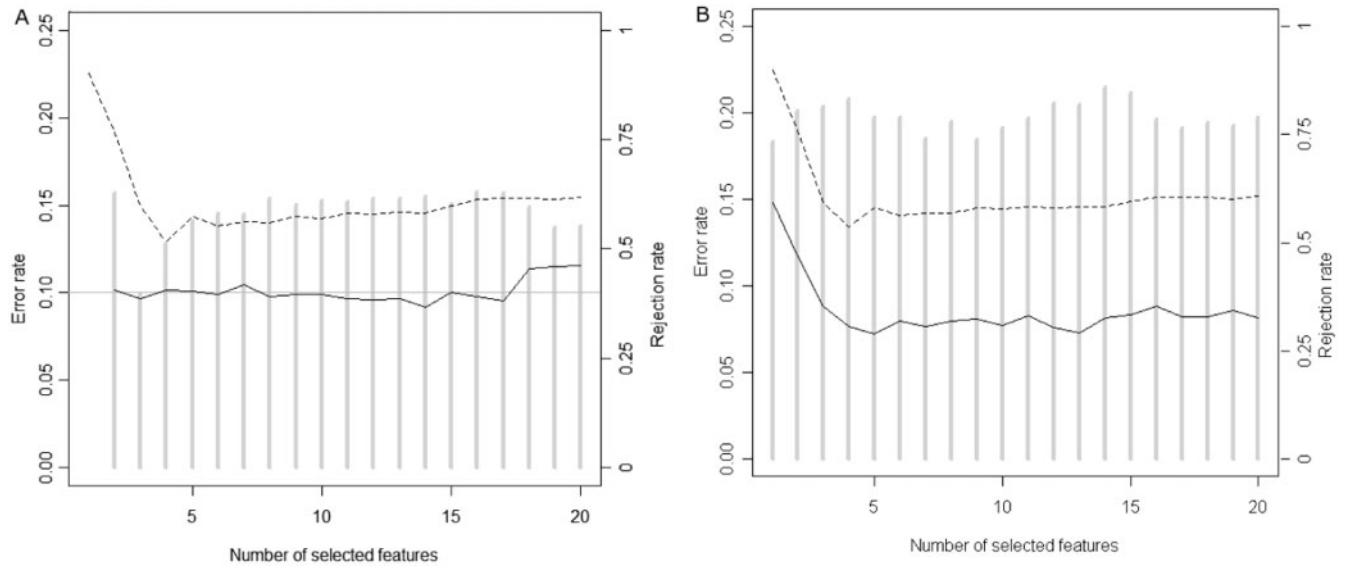

Fig. 3. Result of classification on artificial data based on colon dataset. $N = 200$ and the classification rule is a linear SVM. The dotted line represents the error rate of classifier with no rejection. In panel A, the full line represents the error rate of classifier with reject option whose target error rate is 0.1. In panel B, the full line represents the error rate of classifier using the posterior probabilities whose threshold is to 0.1. The gray histogram represents the rejection rate whose scale is on the left axis.

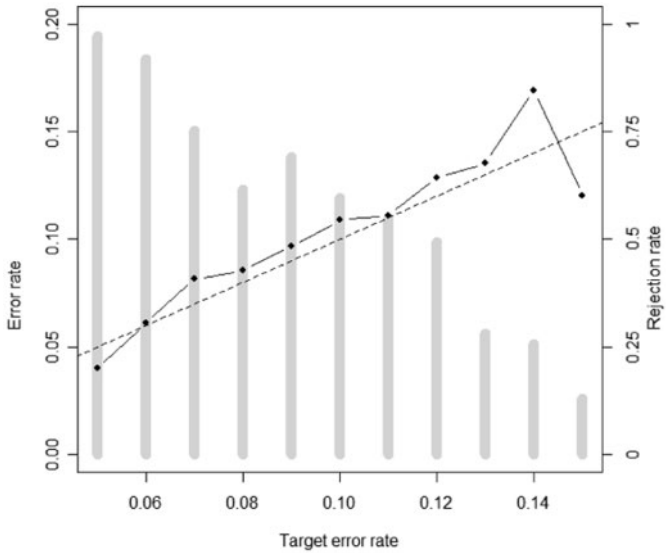

Fig. 4. Results of classification with reject option on artificial data. The full line represents the error rate of the classifier and the dotted line represents the situation where error equal target error. The gray histogram represents the rejection rate whose scale is on the left axis.

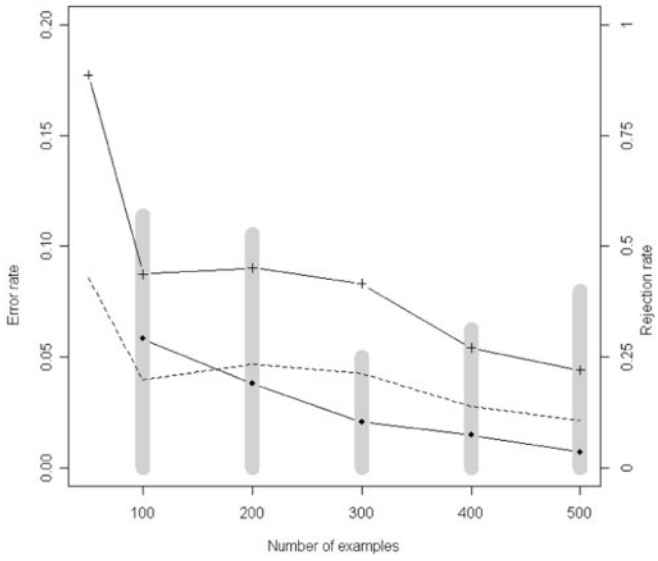

Fig. 5. Result of classification on artificial data based on lung dataset. $N = 200$ and the classification rule is a linear SVM. The cross line represents the error rate of classifier with no rejection option. The circle line represents the error rate of classifier with reject option whose target error rate is represented by the dotted line . The gray histogram represents the rejection rate whose scale is on the left axis.

the training set in the case of classification with reject option and, therefore, the classifier designed with reject option is less powerful than the classifier designed with no reject option. This is the first limitation: if the target error is close to the error obtained by the classifier with no reject option, then there is no benefit to using the classifier with reject option.

Algorithm performance is influenced by the training set size. Figure 5 shows the results on the lung cancer dataset with a SVM classifier. The error rates of the classifier without and with reject option are represented with the cross and circle lines, respectively. The difficulty of classifier design depends on the size of the training set, the larger the training set, the easier the design. Therefore, it is not appropriate to fix the same target error for all training set sizes. We have chosen to set the target error to the half of the error obtained

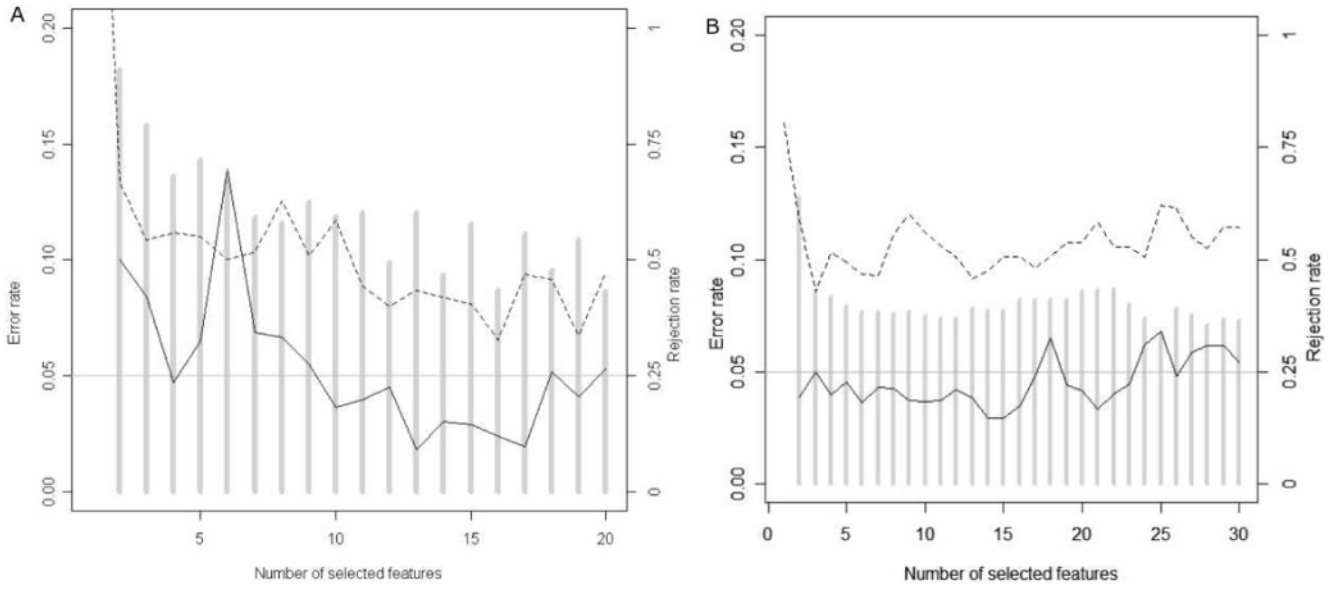

Fig. 6. Result of classification on lung cancer dataset. In panel A, the classification rule is the Fisher discriminant. Panel B is a SVM with linear kernel. The dotted line represents the error rate of classifier with no rejection option and the full line represents the error rate of classifier with reject option whose target error rate is 0.05. The gray histogram represents the rejection rate whose scale is on the left axis.

by the classifier with no reject option. This target error is represented by the dotted line. The rejection rate is represented by the gray histogram. We see that for a training set size of $N = 200$ or more the target error is respected with a rejection rate between 0.25 to 0.5. For $N = 100$ the error rate of the classifier with no reject option is 0.09 and the target error is 0.045. The classifier with reject option does not respect the target error constraint, its error being 0.06. With $N = 50$ the error rate of the classifier with no reject option is 0.18 and the target error is 0.09. There are no results for the classifier with reject option because classifier construction fails, there were no solution to the optimization problem during the threshold computation step. These problems are related to the density estimations of the classes on the classifier output. This estimation is done with $N/4$ examples for each of the two classes, which means 12 and 25 examples for the $N = 50$ and $N = 100$ problems, respectively. When the number of examples is too low, the density estimations are very inaccurate and lead to bad reject options. This is the second limitation of the method. If the number of examples is too low to estimate accurately the class densities, then the classifier construction may fail or the classifier not respect the target error constraints.

4.2 Real data

We have applied our approach on three real microarray datasets. We have used the lung cancer dataset (Bhattacharjee, 2001) whose the task is to discriminate the adenocarcinmas from the other type of cancers. The data contains 139 adenocarcinomas and 64 cancers of another type. The colon cancer dataset (Alon *et al*., 1999) contains the genetic profile of 39 patients affected by a colon cancer and 23 non-affected patients. The breast cancer dataset has (van de Vijver, 2002) 295 patients affected by a breast cancer, 115 belonging to the good-prognosis class and 180 to the poor-prognosis class. We have reduced the three datasets to a selection of the 2000 genes with highest variance.

Unlike the synthetic data, there is no test set to estimate classifier performances. We use k-fold cross-validation, which is an iterative procedure where the data are randomly divided into k subsets. During the i-th iteration, the feature selection and model learning are done on the $k-1$ subsets not containing the i-th subset and the designed classifier is evaluated on the i-th subset. The final estimate is the mean of the results of the k iterations. We use 10-fold cross-validation in our experiments. As noted previously, cross-validation is not very reliable in small sample settings (Hanczar *et al*., 2007), and therefore these results should be viewed with caution.

Figure 6 shows the results of classification on the lung cancer dataset as a function of the number of selected features. The target error of each class is fixed to 0.05. In panel A, the classification rule is the Fisher discriminant. Owing to the high variance of cross-validation, the error curves are unstable; nonetheless, we can put forth some putative statements. The error rate for the classifier with no reject option is decreasing until 15 features and then stays around 0.08. For the classifier with reject option, the error rate is higher than the target error rate and the rejection rate is high with <10 features, perhaps meaning that there are insufficient features, which would be consistent with the classifier with no reject option. From 10 features onward, the error rate respects the target error constraints and the rejection rate is stabilized at around 0.5. In panel B the classification rules is SVM with linear kernel. The error of the classifier with no reject option is around 0.1. For any number of selected features, the classifier with reject option reaches the target error rate. From 3 features onward, the rejection rate is stabilized at around 0.37. These results indicate that on real data, using a reject option with the two classifiers improves their accuracy.

As previously remarked, there are limitations to these methods. For instance, a low number of examples has a bad impact on the results. For the colon cancer dataset, the classifier with no reject

option has an error rate from 0.15 to 0.17. With the target error set to 0.1, the error rate of classifier with reject option is highly variable, from 0.14 to 0.53, and is much higher than the target error. The rejection rate is very high, always >0.9. These poor results are not unexpected because the colon cancer dataset contains 39 and 23 patients for the two classes. That means during the cross-validation procedure, the probability densities of the two classes are estimated with only 17 and 10 examples, respectively. With so small a number of examples, density estimation is very inaccurate and leads to wrong thresholds.

The breast cancer dataset illustrates another limitation of trying to improve classification accuracy with a reject option. The target error is set to 0.2 and the error of the classifier with reject option varies between 0.28 to 0.53. The rejection rate is very high, >0.85. Moreover, classifier construction fails 75% of the time. In this case, the problem does not come from the threshold computation but from the feature-label distribution and the class split in the sample data. The error rate of the classifier with no reject option is between 0.3 and 0.35 but the good-prognosis class represents only 34% of all examples. This means that the classifier has the same accuracy as the majority classifier that predicts all examples to be in the poor-diagnosis class. In effect, the classifier does not discriminate between the two class densities. Computation of the threshold cannot improve the accuracy of the classifier. This result demonstrates the last limitation of our method: if the regular classifier has no discriminatory power, then the incorporation of a reject option will not improve its accuracy.

5 CONCLUSION

We have presented a new approach of the classification of gene-expression data. The principle is to add an reject option to the regular classifier. Only the examples for which the classification is sufficiently reliable are classified. The rejection region is defined by two thresholds. If an example belongs to the reject region, then the example is rejected; otherwise, it is accepted. Unlike regular classifier, the proposed method allows the user to control the error rate of the classifier. The error rate become a parameter of the classifier design and performance now depends of the rejection rate. The classifier respecting the target error constraint with minimal rejection rate is the best. We have also shown how to include this approach in feature selection. A reject option can be added to many classification rules. We have tested it on the Fisher discriminant and SVM.

We have shown on both synthetic and real data that this method can significantly improve classifier accuracy; however, we have shown three conditions for which the method cannot be used: (1) if the target error is close to the error obtained by the classifier with no reject option, then there is no benefit to use the classifier with reject option; (2) if the sample size is too low to obtain a decent estimates of the class densities, then the classifier design may fail or the classifier not respect the target error constraint and (3) if the regular classifier lacks discriminatory power beyond that of the majority classifier on the sample data, then adding of a rejection option will not improve its accuracy. Using a classifier with constraint option can facilitate the construction of more reliable classifiers for medical application where confidence of the diagnosis must be very high.

ACKNOWLEDGEMENTS

We would like to acknowledge the Translational Genomics Research Institute, the French Ministry of Foreign Affairs, and the National Science Foundation (CCF-0514644) for providing support for this research.

Conflict of Interest. none declared.

REFERENCES

Alon,U. *et al*. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, **96**, 6745–6750.

Bhattacharjee,A. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, **98**, 13790–13795.

Chow,C. (1970) On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory*, **16**, 41–46.

Dubuisson,B. and Masson,M. (1993) A statistical decision rule with incomplete knowledge about classes. *Pattern Recognit.*, **26**, 155–165.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Fumera,G. *et al*. (2000) Multiple reject thresholds for improving classification reliability. In Heidelberg,S.B. (ed), *Advances in Pattern Recognition: Joint IAPR International Workshops, SSPR 2000 and SPR 2000*. Alicante, Spain, pp. 863.

Furey,T. *et al*. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Hanczar,B. *et al*. (2007) Decorrelation of the true and estimated classifier errors in high-dimensional settings. *EURASIP J. Bioinform. Syst. Biol.*, Article ID 38473.

Hua,J. *et al*. (2005) Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*, **21**, 1509–1515.

Khan,J. *et al*. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Landgrebe,T. *et al*. (2006) The interaction between classification and reject performance for distance-based reject-option classifiers. *Pattern Recognit. Lett.*, **27**, 908–917.

Li,M. and Sethi,I. (2006) Confidence-based classifier design. *Pattern Recogn.*, **39**, 1230–1240.

Silverman,B.W. (1986) *Density Estimation.* Chapman and Hall.

van de Vijver,M. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.