Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 608701, 10 pages
doi:10.1155/2009/608701

Research Article

Gene-Based Multiclass Cancer Diagnosis with Class-Selective Rejections

Nisrine Jrad, Edith Grall-Maës, and Pierre Beauseroy

Institut Charles Delaunay (ICD, FRE CNRS 2848), Université de Technologie de Troyes, LM2S 12 rue Marie Curie, BP 2060, 10010 Troyes cedex, France

Correspondence should be addressed to Nisrine Jrad, nisrine.jrad@utt.fr

Received 15 January 2009; Accepted 13 March 2009

Recommended by Dechang Chen

Supervised learning of microarray data is receiving much attention in recent years. Multiclass cancer diagnosis, based on selected gene profiles, are used as adjunct of clinical diagnosis. However, supervised diagnosis may hinder patient care, add expense or confound a result. To avoid this misleading, a multiclass cancer diagnosis with class-selective rejection is proposed. It rejects some patients from one, some, or all classes in order to ensure a higher reliability while reducing time and expense costs. Moreover, this classifier takes into account asymmetric penalties dependant on each class and on each wrong or partially correct decision. It is based on ν-1-SVM coupled with its regularization path and minimizes a general loss function defined in the class-selective rejection scheme. The state of art multiclass algorithms can be considered as a particular case of the proposed algorithm where the number of decisions is given by the classes and the loss function is defined by the Bayesian risk. Two experiments are carried out in the Bayesian and the class selective rejection frameworks. Five genes selected datasets are used to assess the performance of the proposed method. Results are discussed and accuracies are compared with those computed by the Naive Bayes, Nearest Neighbor, Linear Perceptron, Multilayer Perceptron, and Support Vector Machines classifiers.

1. Introduction

Cancer diagnosis, based on gene expression profiling, have improved over the past 40 years. Many microarray technologies studies were developed to analyze the gene expression. These genes are later used to categorize cancer classes. Two different classification approaches can be used: class discovery and class prediction. The first is an unsupervised learning approach that allows to separate samples into clusters based on similarities in gene expression, without prior knowledge of sample identity. The second is a supervised approach which predicts the category of an already defined sample using its gene expression profiles. Since these classification problems are described by a large number of genes and a small number of samples, it is crucial to perform genes selection before the classification step. One way to identify informative genes pointed in [1] is the test statistics.

Researches show that the performance of supervised decisions based on selected gene expression can be comparable to the clinical decisions. However, no classification strategy is absolutely accurate. First, many factors may effectively decrease the predictive power of a multiclass problem. For example, findings of [2] imply that information useful for multiclass tumor classification is encoded in a complex gene expression and cannot be given by a simple one. Second, it is not possible to find an optimal classification method for all kinds of multiclass problems. Thus, supervised diagnosis are always considered as an important adjunct of traditional diagnostics and never like its substitute.

Unfortunately, supervised diagnosis can be misleading. They may hinder patient care (wrong decision on a sick patient), add expense (wrong decision on a healthy patient) or confound the results of cancer categories. To overcome

these limitations, a multi-SVM [3] classifier with class-selective rejection [4–7] is proposed. Class-selective rejection consists of rejecting some patients from one, some, or all classes in order to ensure a higher reliability while reducing time and expense costs. Moreover, any of the existing multiclass [8–10] algorithms have taken into consideration asymmetric penalties on wrong decisions. For example, in a binary cancer problem, a wrong decision on a sick patient must cost more than a wrong decision on a healthy patient. The proposed classifier handles this kind of problems. It minimizes a general loss function that takes into account asymmetric penalties dependant on each class and on each wrong or partially correct decision.

The proposed method divides the multiple class problem into several unary classification problems and train one ν-1-SVM [11–13] coupled with its regularization path [14, 15] for each class. The winning class or subset of classes is determined using a prediction function that takes into consideration the costs asymmetry. The parameters of all the ν-1-SVMs are optimized jointly in order to minimize a loss function. Taking advantage of the regularization path method, the entire parameters searching space is considered. Since the searching space is widely extended, the selected decision rule is more likely to be the optimal one. The state-of-art multiclass algorithms [8–10] can be considered as a particular case of the proposed algorithm where the number of decisions is given by the existing classes and the loss function is defined by the Bayesian risk.

Two experiments are reported in order to assess the performance of the proposed approach. The first one considers the proposed algorithm in the Bayesian framework and uses the selected microarray genes to make results comparable with existing ones. Performances are compared with those assessed using Naive Bayes, Nearest Neighbor, Linear Perceptron, Multilayer Perceptron, and Support Vector Machines classifiers, invoked in [1]. The second one shows the ability of the proposed algorithm solving multiclass cancer diagnosis in the class-selective rejection scheme. It minimizes an asymmetric loss function. Experimental results show that, a cascade of class-selective classifiers with class-selective rejections can be considered as an improved supervised diagnosis rule.

This paper is outlined as follows. Section 2 presents a description of the model as a gene selection task. It introduces the multiclass cancer diagnosis problem in the class-selective rejection scheme. It also proposes a supervised training algorithm based on ν-1-SVM coupled with its regularization path. The two experiments are carried out in Section 3, results are reported, compared and discussed. Finally, a conclusion is presented in Section 4.

2. Models and Methods

This section describes the multiclass cancer diagnosis based on microarray data. Feature selection is evoked as a first process in a gene-based cancer diagnosis. Test statistics are used as a possible way for informative genes identification [1]. Once genes selection is processed, a classification problem should be solved. The multiclass cancer diagnosis problem, formulated in the general framework of class-selective rejection, is introduced. A solution based on ν-1-SVM [11–13] is proposed. First a brief description of ν-1-SVM and the derivation of its regularization path [14, 15] is presented. Second, the proposed algorithm [3] is explained. It allows to determine a multiclass cancer diagnosis that minimizes an asymmetric loss function in the class-selective rejection scheme.

2.1. Genes Selection Using Test Statistics. Gene profiles are successfully applied to supervised cancer diagnosis. Since cancer diagnosis problems are usually described by a small set of samples with a large number of genes, feature or gene selection is an important issue in analyzing multiclass microarray data. Given a microarray data with N tumor classes, n tumor samples and g genes per sample, one should identify a small subset of informative genes that contribute most to the prediction task. Various feature selection methods exist in literature. One way pointed in [1] is to use test statistics for the equality of the class means. Authors of [1] formulate first the expression levels of a given gene by a one-way analysis of variance model. Second, the power of genes in discriminating between tumor types is determined by a test statistic. The discrimination power is the value of the test evaluated at the expression level of the gene. The higher the discrimination power is, the more powerful the gene is in discriminating between tumor types. Thus, genes with higher power of discrimination are considered as informative genes.

Let Y_{jp} be the expression level from the pth sample of the jth class, the following general model is considered:

$$Y_{jp} = \mu_j + \epsilon_{jp} \quad \text{for } j = 1,\ldots,N;\ p = 1,\ldots,n_j \text{ with} \sum_{j=1}^{N} n_j = n. \tag{1}$$

In the model μ_j represents the mean expression level of the gene in class w_j, ϵ_{jp} are independent random variables and $E(\epsilon_{jp}) = 0,\ V(\epsilon_{jp}) = \sigma_j^2 < \infty$ for $j = 1,\ldots,N$; $p = 1,\ldots,n_j$.

For the case of homogeneity of variances, the ANOVA F or F test [16] is the optimal one testing the means equality hypothesis. With heterogeneity of variances, the task is challenging. However, it is known that, with a large number of genes present, usually in thousands, no practical test is available to locate the best set of genes. Thus, the authors of [1] studied six different statistics.

(i) ANOVA F test statistic, the definition of this test is

$$F = \frac{(n-N)\sum n_j\left(\overline{Y_{j\cdot}} - \overline{Y_{\cdot\cdot}}\right)^2}{(N-1)\sum\left(n_j - 1\right)s_j^2}, \tag{2}$$

where $\overline{Y_{j\cdot}} = \sum_{p=1}^{n_j} Y_{jp}/n_j$ and $\overline{Y_{\cdot\cdot}} = \sum_{j=1}^{N} n_j\overline{Y_{j\cdot}}/n$, $s_j^2 = \sum_{p=1}^{n_j}(Y_{jp} - \overline{Y_{j\cdot}})^2/(n_j - 1)$. For simplicity, $\sum$ is used to indicate the sum taken over the index j. Under means equality hypothesis and assuming variance homogeneity, this test has a distribution of $F_{N-1,n-N}$ [16].

(ii) Brown-Forsythe test statistic [17], given by

$$B = \frac{\sum n_j \left(\overline{Y_{j\cdot}} - \overline{Y_{\cdot\cdot}}\right)^2}{\sum \left(1 - n_j/n\right) s_j^2}. \quad (3)$$

Under means equality hypothesis, B is distributed approximately as $F_{N-1,\tau}$ where

$$\tau = \frac{\left[\sum (1 - n_j/n) s_j^2\right]^2}{\sum \left(1 - n_j/n\right)^2 s_j^4 / \left(n_j - 1\right)}. \quad (4)$$

(iii) Welch test statistic [18], defined as

$$W = \frac{\sum \omega_j \left(\overline{Y_{j\cdot}} - \sum h_j \overline{Y_{j\cdot}}\right)^2}{(N-1) + 2(N-2)(N+1)^{-1} \sum \left(n_j - 1\right)^{-1} \left(1 - h_j\right)^2}, \quad (5)$$

with $\omega_j = n_j / s_j^2$ and $h_j = \omega_j / \sum \omega_j$. Under means equality hypothesis, W has an approximate distribution of F_{N-1,τ_ω} where

$$\tau_\omega = \frac{N^2 - 1}{3 \sum \left(n_j - 1\right)^{-1} \left(1 - h_j\right)^2}. \quad (6)$$

(iv) Adjusted Welch test statistic [19]. It is similar to Welch statistic and defined to be

$$W^* = \frac{\sum \omega_j^* \left(\overline{Y_{j\cdot}} - \sum h_j^* \overline{Y_{j\cdot}}\right)^2}{(N-1) + 2(N-2)(N+1)^{-1} \sum \left(n_j - 1\right)^{-1} \left(1 - h_j^*\right)^2}, \quad (7)$$

where $\omega_j^* = n_j / (\Phi_j s_j^2)$ with Φ_j chosen such that $1 \leq \Phi_j \leq (n_j - 1)/(n_j - 3)$ and $h_j^* = \omega_j^* / \sum \omega_j^*$. Under means equality hypothesis, W^* has an approximate distribution of F_{N-1,τ_ω^*} where

$$\tau_\omega^* = \frac{N^2 - 1}{3 \sum \left(n_j - 1\right)^{-1} \left(1 - h_j^*\right)^2}. \quad (8)$$

(v) Cochran test statistic [20]. This test statistic is simply the quantity appearing in the numerator of the Welch test statistic W, that is,

$$C = \sum \omega_j \left(\overline{Y_{j\cdot}} - \sum h_j \overline{Y_{j\cdot}}\right)^2. \quad (9)$$

Under means equality hypothesis, C has an approximate distribution of χ^2_{N-1}.

(vi) Kruskal-Wallis test statistic. This is the well-known nonparametric test given by

$$H = \frac{12}{n(n+1)} \sum \frac{R_j^2}{n_j} - 3(n+1), \quad (10)$$

where R_j is the rank sum for the jth class. The ranks assigned to Y_{jp} are those obtained from ranking the entire set of Y_{jp}. Assuming each $n_j \geq 5$, then under means equality hypothesis, H has an approximate distribution of χ^2_{N-1} [21].

These tests performances are evaluated and compared over different supervised learning methods applied to publicly available microarray datasets. Experimental results show that the model for gene expression values without assuming equal variances is more appropriate than that assuming equal variances. Besides, under heterogeneity of variances, Brown-Forsythe test statistic, Welch test statistic, adjusted Welch test statistic, and Cochran test statistic, perform much better than ANOVA F test statistic and Kruskal-Wallis test statistic.

2.2. Multitumor Classes with Selective Rejection. Once gene selection is processed, the classification problem should be solved. Let us define this diagnosis problem in the class-selective rejection scheme. Assuming that the multiclass cancer problem deals with N tumor classes noted $w_1 \ldots w_N$ and that any patient or sample x belongs to one tumor class and has d informative genes, a decision rule consists in a partition Z of $\Re^d$ in I sets Z_i corresponding to the different decision options. In the simple classification scheme, the options are defined by the N tumor classes. In the class-selective rejection scheme, the options are defined by the N tumor classes and the subsets of tumor classes (i.e. assigning patient x to the subset of tumor classes $\{w_1, w_3\}$ means that x is assigned to cancer categories w_1 and w_3 with ambiguity).

The problem consists in finding the decision rule Z^* that minimizes a given loss function $c(Z)$ defined by

$$c(Z) = \sum_{i=1}^{I} \sum_{j=1}^{N} c_{ij} P_j P\left(D_i / w_j\right), \quad (11)$$

where c_{ij} is the cost of assigning a patient x to the ith decision option when it belongs to the tumor class w_j. The values of c_{ij} being relative since the aim is to minimize $c(Z)$, the values can be defined in the interval $[0; 1]$ without loss of generality. P_j is the a priori probability of tumor class w_j and $P(D_i/w_j)$ is the probability that patients of the tumor class w_j are assigned to the ith option.

2.3. μ-1-SVM. To solve the multiclass diagnosis problem, an approach based on ν-1-SVM is proposed. Considering a set of m samples of a given tumor classes $X = \{x_1, x_2, \ldots, x_m\}$ drawn from an input space $\mathcal{X}$, ν-1-SVM computes a decision function $f_X^\lambda(\cdot)$ and a real number b^λ in order to determine the region $\mathcal{R}^\lambda$ in $\mathcal{X}$ such that $f_X^\lambda(x) - b^\lambda \geq 0$ if the sample $x \in \mathcal{R}^\lambda$ and $f_X^\lambda(x) - b^\lambda < 0$ otherwise. The decision function $f_X^\lambda(\cdot)$ is parameterized by $\lambda = \nu m$ (with $0 \leq \nu < 1$) to control the number of outliers. It is designed by minimizing the volume of $\mathcal{R}^\lambda$ under the constraint that all the samples of X, except the fraction ν of outliers, must lie in $\mathcal{R}^\lambda$. In order to determine $\mathcal{R}^\lambda$, the space of possible functions $f_X^\lambda(\cdot)$ is reduced to a Reproducing Kernel Hilbert Space (RKHS) with kernel function $K(\cdot, \cdot)$. Let $\Phi : \mathcal{X} \to \mathcal{H}$ be the mapping defined over the input space $\mathcal{X}$. Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ be a dot product defined in $\mathcal{H}$. The kernel $K(\cdot, \cdot)$ over $\mathcal{X} \times \mathcal{X}$ is defined by:

$$\forall \left(x_p, x_q\right) \in \mathcal{X} \times \mathcal{X} \quad K\left(x_p, x_q\right) = \left\langle \Phi\left(x_p\right), \Phi\left(x_q\right) \right\rangle_{\mathcal{H}}. \quad (12)$$

Without loss of generality, $K(\cdot,\cdot)$ is supposed normalized such that for any $x \in \mathcal{X}$, $K(x,x) = 1$. Thus, all the mapped vectors $\Phi(x_p)$, $p = 1,\ldots,m$ are in a subset of a hypersphere with radius one and center O. Provided $K(\cdot,\cdot)$ is always positive and $\Phi(X)$ is a subset of the positive orthant of the hypersphere. A common choice of $K(\cdot,\cdot)$ is the Gaussian RBF kernel $K(x_p,x_q) = \exp[-1/2\sigma^2\|x_p - x_q\|^2_{\mathcal{X}}]$ with σ the parameter of the Gaussian RBF kernel. ν-1-SVM consists of separating the mapped samples in $\mathcal{H}$ from the center O with a hyperplane $\mathcal{W}^\lambda$. Finding the hyperplane $\mathcal{W}^\lambda$ is equivalent to find the decision function $f^\lambda_X(\cdot)$ such that $f^\lambda_X(x) - b^\lambda = \langle w^\lambda, \Phi(x)\rangle_{\mathcal{H}} - b^\lambda \geq 0$ for the $(1-\nu)m$ mapped samples while $\mathcal{W}^\lambda$ is the hyperplane with maximum margin $b^\lambda/\|w^\lambda\|_{\mathcal{H}}$ with w^λ the normal vector of $\mathcal{W}^\lambda$.

This yields $f^\lambda_X(\cdot)$ as the solution of the following convex quadratic optimization problem:

$$\min_{w^\lambda,b^\lambda,\xi_p} \sum_{p=1}^{m} \xi_p - \lambda b^\lambda + \frac{\lambda}{2}\left\|w^\lambda\right\|^2_{\mathcal{H}}$$
$$\text{subject to } \left\langle w^\lambda, \Phi\left(x_p\right)\right\rangle_{\mathcal{H}} \geq b^\lambda - \xi_p, \quad \xi_p \geq \quad \forall p = 1,\ldots,m \tag{13}$$

where ξ_p are the slack variables. This optimization problem is solved by introducing lagrange multipliers α_p. As a consequence to Kuhn-Tücker conditions, w^λ is given by

$$w^\lambda = \frac{1}{\lambda}\sum_{p=1}^{m} \alpha_p \Phi\left(x_p\right), \tag{14}$$

which results in

$$f^\lambda_X(\cdot) - b^\lambda = \frac{1}{\lambda}\sum_{p=1}^{m} \alpha_p K\left(x_p,\cdot\right) - b^\lambda. \tag{15}$$

The dual formulation of (13) is obtained by introducing Lagrange multipliers as

$$\min_{\alpha_1,\ldots,\alpha_m} \frac{1}{2\lambda}\sum_{p=1}^{m}\sum_{q=1}^{m} \alpha^\lambda_p \alpha^\lambda_q K\left(x_p,x_q\right)$$
$$\text{with} \sum_{p=1}^{m} \alpha^\lambda_p = \lambda, \quad 0 \leq \alpha^\lambda_p \leq 1 \quad \forall p = 1,\ldots,m. \tag{16}$$

A geometrical interpretation of the solution in the RKHS is given by Figure 1. $f^\lambda_X(\cdot)$ and b^λ define a hyperplane $\mathcal{W}^\lambda$ orthogonal to $f^\lambda_X(\cdot)$. The hyperplane $\mathcal{W}^\lambda$ separates the $\Phi(x_p)$s from the sphere center, while having $b^\lambda/\|w^\lambda\|_{\mathcal{H}}$ maximum which is equivalent to minimize the portion $\mathcal{S}^\lambda$ of the hypersphere bounded by $\mathcal{W}^\lambda$ that contains the set $\{\Phi(x) \text{ s.t. } x \in \mathcal{R}^\lambda\}$.

Tuning ν or equivalently λ is a crucial point since it enables to control the margin error. It is obvious that changing λ leads to solve the optimization problem formulated in (16) in order to find the new region $\mathcal{R}^\lambda$. To obtain great computational savings and extend the search space of λ, we proposed to use ν-1-SVM regularization path [14, 15]. Regularization path was first introduced by Hastie et al. [14] for a binary SVM. Later, Rakotomamojy and Davy [15] developed the entire regularization path for a ν-1-SVM. The basic idea of the ν-1-SVM regularization path is that the parameter vector of a ν-1-SVM is a piecewise linear function of λ. Thus the principle of the method is to start with large λ, (i.e., $\lambda = m - \epsilon$) and decrease it towards zero, keeping track of breaks that occur as λ varies.

As λ decreases, $\|w^\lambda\|_{\mathcal{H}}$ increases and hence the distance between the sphere center and $\mathcal{W}^\lambda$ decreases. Samples move from being outside (non-margin SVs with $\alpha^\lambda_p = 1$ in Figure 1) to inside the portion $\mathcal{S}^\lambda$ (non-SVs with $\alpha^\lambda_p = 0$). By continuity, patients must linger on the hyperplane $\mathcal{W}^\lambda$ (margin SVs with $0 < \alpha^\lambda_p < 1$) while their α^λ_ps decrease from 1 to 0. α^λ_ps are piecewise-linear in λ. Break points occur when a point moves from a position to another one. Since α^λ_p is piecewise-linear in λ, $f^\lambda(\cdot)$ and b^λ are also piecewise-linear in λ. Thus, after initializing the regularization path (computing α^λ_p by solving (16) for $\lambda = m - \epsilon$), almost all the α^λ_ps are computed by solving linear systems. Only for some few integer values of λ smaller than m, α^λ_ps are computed by solving (16) according to [15].

Using simple linear interpolation, this algorithm enables to determine very rapidly the ν-1-SVM corresponding to any value of λ.

2.4. Multiclass SVM Based on μ-1-SVM. Given N classes and N trained ν-1-SVMs, one should design a supervised decision rule Z, moving from unary to multiclass classifier by assigning samples to a decision option. To determine the decision rule, first a prediction function should decide the winning option. A distance measure between x and the training class set w_j, using the ν-1-SVM parameterized by λ_j, is defined as follows:

$$d^{\lambda_j}(x) = \frac{\cos\left(\widehat{w^{\lambda_j}, \Phi(x)}\right)}{\cos\left(\theta^{\lambda_j}\right)} = \frac{\left\|w^{\lambda_j}\right\|_{\mathcal{H}}}{b^{\lambda_j}} \cos\left(\widehat{w^{\lambda_j}, \Phi(x)}\right), \tag{17}$$

where θ^{λ_j} is the angle delimited by w^{λ_j} and the support vector as shown in Figure 1. $\cos(\theta^{\lambda_j})$ is a normalizing factor which is used to make all the $d^\lambda_j(x)$ comparable.

Using $\|\Phi(x)\| = 1$ in (17) leads to the following:

$$d^{\lambda_j}(x) = \frac{\left\langle w^{\lambda_j}, \Phi(x)\right\rangle_{\mathcal{H}}}{b^{\lambda_j}} = \frac{1/\lambda_j \sum_{p=1}^{n_j} \alpha_p^{\lambda_j} K\left(x_p,x\right)}{b^{\lambda_j}}. \tag{18}$$

Since the $\alpha_p^{\lambda_j}$ are obtained by the regularization path for any value of λ_j, computing d^{λ_j} is considered as an easy-fast task. The distance measure $d^{\lambda_j}(x)$ is inspired from [22]. When data are distributed in a unimodal form, the $d^{\lambda_j}(x)$ is a decreasing function with respect to the distance between a sample x and the data mean. The probability density function is also a decreasing function with respect to the distance from the mean. Thus, $d^{\lambda_j}(x)$ preserves distribution order relations. In such case, and under optimality of the ν-1-SVM classifier, the use of $d^{\lambda_j}(x)$ should reach the same performances as the one obtained using the distribution.

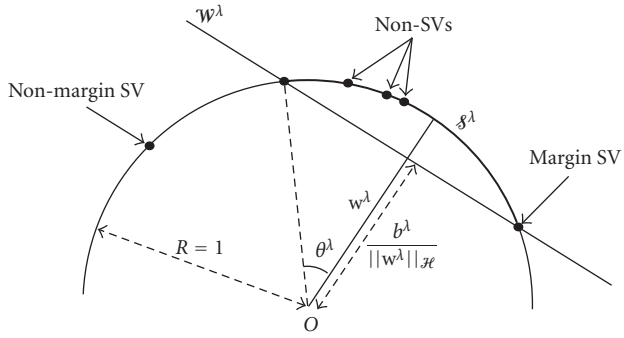

FIGURE 1: Training data mapped into the feature space on a portion $\mathcal{S}^\lambda$ of a hypersphere.

In the simplest case of multiclass problems where the loss function is defined as the error probability, a patient x is assigned to the tumor class maximizing $d^{\lambda_j}(x)$.

To extend the multiclass prediction process to the class-selective scheme, a weighted form of the distance measure is proposed. The weight β_j is associated to d^{λ_j}. β_j reflects an adjusted value of the distance d^{λ_j} according to the penalty associated with the tumor class w_j. Thus, introducing weights leads to treat differently each tumor class and helps solving problems with different costs c_{ij} on the classification decisions.

Finally, in the general case where the loss function is considered in the class-selective rejection scheme, the prediction process can be defined as follows: a blinded sample x is assigned to the ith option if and only if

$$\sum_{j=1}^{N} c_{ij} P_j \beta_j d^{\lambda_j}(x) \leq \sum_{j=1}^{N} c_{lj} P_j \beta_j d^{\lambda_j}(x), \quad \forall l = 1 \ldots I,\ l \neq i. \tag{19}$$

Thus, in contrast to previous multiclass SVMs, which construct the maximum margin between classes and locate the decision hyperplane in the middle of the margin, the proposed approach resembles more to the robust Bayesian classifier. The distribution of each tumor class is considered and the optimal decision is slightly deviated toward the class with the smaller variance.

The proposed decision rule depends on $\boldsymbol{\sigma}, \boldsymbol{\nu}$ and $\boldsymbol{\beta}$ vectors of σ_j, ν_j and β_j for $j = 1, \ldots, N$. Tuning ν_j is the most time expensive task since changing ν_j leads to solve the optimization problem formulated in (16). Moreover, tuning ν_j is a crucial point, it enables to control the margin error. In fact, it was shown in [11] that this regularization parameter is an upper bound on the fraction of outliers and a lower bound on the fraction of the SVs. In [9, 23] a smooth grid search was supplied in order to choose the optimal values of $\boldsymbol{\nu}$. The N values ν_js were chosen equal to reduce the computational costs. However, this assumption reduces the search space of parameters too. To avoid this restriction, the proposed approach optimizes all the ν_j with $j = 1, \ldots, N$ corresponding to the $N\nu$-1-SVMs using regularization path and consequently explores the entire parameters space. Thus the tuned ν_j are most likely to be the optimal ones. The parameter $\boldsymbol{\sigma}$ are set equals $\sigma_1 = \sigma_2 = \cdots = \sigma_N$.

The optimal vector of σ_j, λ_j and β_j, $j = 1, \ldots, N$, is the one which minimizes an estimator of $c(Z)$ using a validation set. Since the problem is described by a sample set, an estimator $\hat{c}(Z)$ of $c(Z)$ given by (11) is used:

$$\hat{c}(Z) = \sum_{i=1}^{I} \sum_{j=1}^{N} c_{ij} \hat{P}_j \hat{P}\left(\frac{D_i}{w_j}\right), \tag{20}$$

where $\hat{P}_j$ and $\hat{P}(D_i/w_j)$ are the empirical estimators of P_j and $P(D_i/w_j)$, respectively.

The optimal rule is obtained by tuning λ_j, β_j and σ_j so that the estimated loss $\hat{c}(Z)$ computed on a validation set is minimum. This is accomplished by employing a global search for λ_j and β_j and an iterative search over the kernel parameter. For each given value σ of the parameter kernels, ν-1-SVMs are trained using the regularization path method on a training set. Then the minimization of $\hat{c}(Z)$ over a validation set is sought by solving an alternate optimization problem over λ_j and β_j which is easy since all ν-1-SVM solutions are easily interpolated from the regularization path. σ is chosen from a previously defined set of real numbers $[\sigma_0, \ldots, \sigma_s]$ with $s \in \aleph$. Algorithm 1 elucidates the proposed approach.

3. Experimental Results

In this section, two experiments are reported in order to assess the performance of the proposed approach. First, the cancer diagnosis problem is considered in the traditional Bayesian framework. Five gene expression datasets and five supervised algorithms are considered. Each gene dataset was selected using the six test statistics of [1]. The decisions are given by the possible set of tumor classes and the loss function is defined as the probability of error to make results comparable with those of [1]. Second, in order to show the advantages of considering the multiclass cancer diagnosis in class-selective rejection scheme, one gene dataset is considered and studied with an asymmetric loss function. A cascade of classifiers with rejection options is used to ensure a reliable diagnosis. For both experiments, the loss function was minimized by determining the optimal parameters β_j and λ_j for $j = 1, \ldots, N$ for a given kernel parameter $\boldsymbol{\sigma}$ and by testing different values of $\boldsymbol{\sigma}$ in the set $[2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2]$. Finally, the decision rule which minimizes the loss function is selected.

3.1. Bayesian Framework. Five multiclass gene expression datasets leukemia72 [24], ovarian [25], NCI [26, 27], lung cancer [28] and lymphoma [29] were considered. Table 1 describes the five genes datasets. For each dataset, the six test statistics F, B, W, W^*, C, and H were used to select informative genes.

The cancer diagnosis problem was considered in the traditional Bayesian framework. Decisions were given by the set of possible classes and loss function was defined by the error risk. This means that in (20) c_{ij} are defined according

$\boldsymbol{\theta} := \emptyset$
$C := \emptyset$
for $\sigma \leftarrow \sigma_0$ to σ_s do
  /*Using the Training Set*/
  for $j \leftarrow 1$ to N do
    Train ν-1-SVM on w_j, namely solving the QP (16)
    Derive the regularization path for w_j, namely compute the $\boldsymbol{\alpha}^{\lambda_j}$s
  end
  /*Using the Validation Set*/
  $\boldsymbol{\lambda} := \boldsymbol{\lambda}_0$
  $\boldsymbol{\beta} := \boldsymbol{\beta}_0$
  repeat
    $d^{\lambda_j}(x) := \frac{1}{\lambda_j} \sum_{p=1}^{n_j} \alpha_p^{\lambda_j} K(x_p, x)/b^{\lambda_j}$
    $\widehat{P}_j := |w_j| / \sum_{j=1}^{N} |w_j|$ /*| | = cardinality*/
    Assign x to a decision ψ_i according to (19)
    $\widehat{P}(D_i/w_j) := |\{x \text{ of } w_j \text{ assigned to } \psi_i\}|/|\{x/x \in w_j\}|$
    $\widehat{c}(Z) := \sum_{i=1}^{I} \sum_{j=1}^{N} c_{ij} \widehat{P}_j \widehat{P}(D_i/w_j)$
    $\boldsymbol{\lambda} := \boldsymbol{\lambda}_{\text{new}}$/* construct the new vector according to the direction of greatest decrease */
    $\boldsymbol{\beta} := \boldsymbol{\beta}_{\text{new}}$
  until $\widehat{c}(Z)$*is minimum*
  $\boldsymbol{\theta} := \boldsymbol{\theta} \cup \{\sigma, \boldsymbol{\lambda}, \boldsymbol{\beta}\}$
  $C := C \cup \{\widehat{c}(Z)\}$
end
index := min{C}
$\boldsymbol{\theta}_{\text{optimal}} := \boldsymbol{\theta}_{\text{index}}$

Algorithm 1: Multiclass SVM minimizing an asymmetric loss function.

Table 1: Multiclass gene expression datasets.

Dataset	Leukemia72	Ovarian	NCI	Lung cancer	Lymphoma
No. of gene	6817	7129	9703	918	4026
No. of sample	72	39	60	73	96
No. of class	3	3	9	7	9

Table 2: Loss function cost matrix in the Bayesian framework.

		Patient class				
		1	2	.	.	N
Prediction	1	0	1	·	·	1
	2	1	0	1		·
	·	·	·	·	·	·
	·	·		·	·	1
	N	1	·	·	1	0

to the Table 2. The performance of the proposed method was measured by evaluating its accuracy rate and it was compared to results obtained by the five predictors evoked in [1]: Naive Bayes, Nearest Neighbor, Linear Perceptron, Multilayer Perceptron Neural Network with five nodes in the middle layer, and Support Vector Machines with second-order polynomial kernel.

To compute the generalization accuracy of the proposed classifier, Leave One Out (LOO) resampling method is used to divide a gene dataset of n patients into two sets, a set of $n - 1$ patients and a test set of 1 blinded patient. This method involves n separate runs. For each run, the first set of $n - 1$ samples is divided using 5 Cross-validation (5-CV) into a training set and a validation set. $N\nu$-1-SVMs are trained using the training set for all values of ν_j. The decision is obtained by tuning the parameters β_j, λ_j and σ_j for $j = 1, \ldots, N$ so that the loss function computed on the validation set is minimum. Optimal parameters are then used to build the decision rule using the whole $n - 1$ samples. The blinded test set is classified according to this rule. The overall prediction error is the sum of the patients misclassified on all n runs.

Table 3 reports errors of the proposed algorithm, the average value and the median value of the 5 classifiers prediction errors reported in [1] when 50 informative genes are used. Table 4 reports values when 100 informative genes are used. F, B, W, W^*, C, and H represent the six test statistics.

Experimental results show that, for ovarian, NCI, lung cancer and lymphoma multiclass genes problems, the proposed approach achieves competitive performances compared to the 5 classifiers reported in [1]. For these datasets, prediction errors of the proposed approach are less than the mean and median values of the 5 classifiers prediction errors reported in [1]. However, for leukemia72, the proposed

Table 3: Prediction errors of the proposed classifier, mean and median values of the **5** classifiers prediction errors according to [1] with 50 informative selected genes.

		F	*B*	*W*	*W**	*C*	*H*
Leukemia	Proposed algorithm	4	3	5	5	3	2
	Mean	3.4	2.4	2.8	2.8	3.2	3.0
	Median	3	2	3	3	3	3
Ovarian	Proposed algorithm	0	0	0	0	0	0
	Mean	0.2	0.0	0.0	0.0	0.0	0.0
	Median	0	0	0	0	0	0
NCI	Proposed algorithm	31	26	27	27	27	33
	Mean	36.0	32.0	27.4	26.0	27.0	35.4
	Median	35	29	27	27	27	35
Lung cancer	Proposed algorithm	14	16	16	16	16	15
	Mean	17.6	17.0	17.6	17.6	18.0	18.0
	Median	17	17	18	18	18	18
Lymphoma	Proposed algorithm	18	16	9	10	9	15
	Mean	23.8	19.8	14.0	14.0	12.8	22.0
	Median	23	19	12	12	13	20

Table 4: Prediction errors of the proposed classifier, mean and median values of the **5** classifiers prediction errors according to [1] with 100 informative selected genes.

		F	*B*	*W*	*W**	*C*	*H*
Leukemia	Proposed algorithm	5	2	3	3	4	6
	Mean	3.4	3.0	3.0	3.0	3.2	3.0
	Median	3	3	4	3	3	3
Ovarian	Proposed algorithm	0	0	0	0	0	0
	Mean	0.2	0.0	0.0	0.0	0.0	0.0
	Median	0	0	0	0	0	0
NCI	Proposed algorithm	33	21	26	25	26	36
	Mean	33.0	22.6	23.8	25.2	25.2	31.6
	Median	33	22	25	26	26	31
Lung cancer	Proposed algorithm	11	10	11	11	11	13
	Mean	12.2	12.2	11.4	12.2	12.2	15.8
	Median	12	12	11	11	11	14
Lymphoma	Proposed algorithm	16	16	11	10	11	17
	Mean	21.8	19.2	13.0	13.8	14.4	18.2
	Median	17	16	12	12	12	18

Table 5: Confusion matrix of 50 *W** lung cancer dataset. Total of misclassified is equal to 16.

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	6	0	0	0	0	0	0
	SCLC	0	4	0	0	0	1	0
	LCLC	0	0	3	0	0	4	1
	SCC	0	0	0	16	0	3	0
	AC2	0	0	0	0	4	0	0
	AC3	0	1	1	0	1	4	0
	AC1	0	0	1	0	2	1	20

Table 6: Confusion Matrix of 50 *H* lung cancer dataset. Total of misclassified is equal to 15.

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	5	0	0	0	0	0	0
	SCLC	0	4	0	0	0	0	0
	LCLC	0	0	1	1	0	2	2
	SCC	0	0	2	14	0	1	0
	AC2	0	0	0	0	7	0	0
	AC3	0	0	2	1	0	8	0
	AC1	1	1	0	0	0	2	19

Table 7: Asymmetric cost matrix of the loss function.

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	0	1	1	1	1	1	1
	SCLC	1	0	1	1	1	1	1
	LCLC	1	1	0	0.9	0.9	1	1
	SCC	1	1	0.9	0	0.9	1	0.9
	AC2	1	1	0.9	0.9	0	0.9	0.9
	AC3	1	1	0.9	0.9	0.9	0	0.9
	AC1	1	1	0.9	0.9	0.9	0.9	0
	{*LCLC, SCC, AC3*}	1	1	0.6	0.6	0.9	0.2	0.9
	All tumors	1	0.2	0.6	0.6	0.2	0.2	0.5
	All classes	0.6	0.2	0.6	0.6	0.2	0.6	0.6

Table 8: Confusion matrix of the 50 W^* lung cancer problem with class-selective rejection using cost matrix defined in Table 7. Total of misclassified is equal to 10, total of partially and totally rejected samples is equal to 8.

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	6	0	0	0	0	0	0
	SCLC	0	3	0	0	0	0	0
	LCLC	0	0	3	0	0	4	0
	SCC	0	0	0	16	0	2	0
	AC2	0	0	0	0	4	0	0
	AC3	0	0	0	0	1	3	0
	AC1	0	0	1	0	1	1	20
	{*LCLC, SCC, AC3*}	0	0	1	0	0	2	0
	All tumors	0	2	0	0	1	1	1
	All classes	0	0	0	0	0	0	0

algorithm performances are almost in the same range of those provided by the 5 classifiers reported in [1]. The proposed approach prediction error is equal, or in the worst case, slightly higher than the mean and median errors.

Moreover, we can note that focussing on the test statistics comparison, experimental results confirm those of [1]. B, W and W^* can be the most performing tests under variances heterogeneity assumptions.

3.2. Class-Selective Rejection Framework. In the following, we present the study of lung cancer problem in the class-selective rejection scheme. Lung cancer diagnosis problem is determined by the gene expression profiles of 67 lung tumors and 6 normal lung specimens from patients whose clinical course was followed for up to 5 years. The tumors comprised 41 Adenocarcinomas (ACs), 16 squamous cell carcinomas (SCCs); 5 cell lung cancers (LCLCs) and 5 small cell lung cancers (SCLCs). ACs are subdivided into three subgroups 21 AC of group 1 tumors, 7 AC of group 2 tumors and 13 AC of group 3 tumors. Thus, the multiclass diagnosis cancer consists of 7 classes.

Authors in [28] observed that AC of group 3 tumors shared strong expression of genes with LCLC and SCC tumors. Thus, poorly differentiated AC is difficult to distinguish from LCLC or SCC. Confusion matrices (Tables 5 and 6) computed in the Bayesian framework, with 50 W^* and 50 H prove well these claims. It can be noticed that 8 of the 16 misclassified 50 W^* patients and 8 of the 15 misclassified 50 H patients correspond to confusion between these three subcategories. Therefore, one may define a new decision option as a subset of these three classes to reduce error.

Moreover, same researches affirm that distinction between patients with nonsmall cell lung tumors (SCC, AC and LCLC) and those with small cell tumors or SCLC is extremely important, since they are treated very differently. Thus, a confusion or wrong decision among patients of nonsmall cell lung tumors should cost less than a confusion between nonsmall and small lung cells tumors. Besides, one may provide an extra decision option that includes all the subcategories of tumors to avoid this kind of confusion. Finally, another natural decision option can be the set of all

Table 9: Confusion matrix of the cascade classifier (50 W^* with rejection and 50 H classifier). Total of misclassified is equal to 13.

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	6	0	0	0	0	0	0
	SCLC	0	4	0	0	0	0	0
	LCLC	0	0	3	0	0	4	1
	SCC	0	0	0	16	0	2	0
	AC2	0	0	0	0	5	0	0
	AC3	0	1	1	0	1	6	0
	AC1	0	0	1	0	1	1	20

classes, which means that the classifier has totally withhold taking a solution.

Given all these information, the loss function can be empirically defined according to the asymmetric cost matrix given in Table 7. Solving 50 W^* lung cancer problem in this scheme leads to the confusion matrix presented in Table 8. As a comparison with Table 5, one may mainly note that the number of misclassified patients decreases from 16 to 10 and 8 withhold decisions or rejected patients. This partial rejection contributes to avoid confusion between nonsmall and small lung cells tumors and reduces errors due to indistinctness among LCLC, SCC and AC3. Besides, according to the example under study, no patient is totally rejected. It is an expected result since initially (Table 5) there was no confusion between normal and tumor samples.

To take a decision concerning the rejected patients, we may refer to clinical analysis. It is worth to note that for partially rejected patients, clinical analysis will be less expensive in terms of time and money than those on completely blinded patients. Moreover, a supervised solution can be also proposed. It aims to use genes selected from another test statistic in order to assign rejected patients to one of the possible classes. According to Tables 3 and 4, prediction errors computed on same patients using genes selected by different test statistics may decrease since errors of two different test statistics do not occur on the same patients. Thus, we chose 50 H lung cancer dataset to reclassify the 8 rejected patients of Table 8. Five of them were correctly classified while three remained misclassified. Results are reported in Table 9. The number of misclassified patients decreases to 13 which is less than all the prediction errors obtained with 50 informative genes (lung cancer problem prediction errors of Table 3). In fact, many factors play an important role in the cascade classifiers system such as the asymmetric costs matrix which has been chosen empirically, the choice of test statistics, the number of classifiers in a cascade system,.... Such concerns are under study.

4. Conclusion

Cancer diagnosis using genes involve a gene selection task and a supervised classification procedure. This paper tackles the classification step. It considers the problem of gene-based multiclass cancer diagnosis in the general framework of class-selective rejection. It gives a general formulation of the problem and proposes a possible solution based on ν-1-SVM coupled with its regularization path. The proposed classifier minimizes any asymmetric loss function. Experimental results show that, in the particular case where decisions are given by the possible classes and the loss function is set equal to the error rate, the proposed algorithm, compared with the state of art multiclass algorithms, can be considered as a competitive one. In the class-selective rejection, the proposed classifier ensures higher reliability and reduces time and expense costs by introducing partial and total rejection. Furthermore, results prove that a cascade of classifiers with class-selective rejections can be considered as a good way to get improved supervised diagnosis. To get the most reliable diagnosis, the confusion matrix defining the loss function should be carefully chosen. Finding the optimal loss function according to performance constraints is an promising approach [30] which is actually under investigation.

Acknowledgments

The authors thank Dr. Dechang Chen of the Uniformed Services University of the Health Sciences for providing the gene selected data of leukemia72, ovarian, NCI, lung cancer and lymphoma using six test statistics ANOVA F, Brown-Forsythe test, Welch test, Adjusted Welch test, Cochran test and Kruskal-Wallis test. This work was supported by the "Conseil Régional Champagne Ardenne" and the "Fonds Social Européen".

References

[1] D. Chen, Z. Liu, X. Ma, and D. Hua, "Selecting genes by test statistics," *Journal of Biomedicine and Biotechnology*, vol. 2005, no. 2, pp. 132–138, 2005.

[2] S. Ramaswamy, P. Tamayo, R. Rifkin, et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 15149–15154, 2001.

[3] N. Jrad, E. Grail-Maës, and P. Beauseroy, "A supervised decision rule for multiclass problems minimizing a loss function," in *Proceedings of the 7th International Conference on Machine Learning and Applications (ICMLA '08)*, pp. 48–53, San Diego, Calif, USA, December 2008.

[4] T. M. Ha, "The optimum class-selective rejection rule," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 6, pp. 608–615, 1997.

[5] T. Horiuchi, "Class-selective rejection rule to minimize the maximum distance between selected classes," *Pattern Recognition*, vol. 31, no. 10, pp. 1579–1588, 1998.

[6] E. Grall-Maës, P. Beauseroy, and A. Bounsiar, "Multilabel classification rule with performance constraints," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, vol. 3, pp. 784–787, Toulouse, France, May 2006.

[7] N. Jrad, E. Grall-Maës, and P. Beauseroy, "Gaussian mixture models for multiclass problems with performance constraints," in *Proceedings of the 17th European Symposium on Artificial Neural Networks (ESANN '09)*, Bruges, Belgium, April 2009.

[8] L. Bottou, C. Cortes, J. Denker, et al., "Comparison of classifier methods: a case study in handwritten digit recognition," in *Proceedings of the 12th IAPR International Conference on Pattern Recognition (ICPR '94)*, vol. 2, pp. 77–82, Jerusalem, Israel, October 1994.

[9] X. Yang, J. Liu, M. Zhang, and K. Niu, "A new multiclass SVM algorithm based on one-class SVM," in *Proceedings of International Conference on Computational Science (ICCS '07)*, pp. 677–684, Beijing, China, May 2007.

[10] P.-Y. Hao and Y.-H. Lin, "A new multi-class support vector machine with multi-sphere in the feature space," in *Proceedings of the 20th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE '07)*, vol. 4570 of *Lecture Notes in Computer Science*, pp. 756–765, Kyoto, Japan, June 2007.

[11] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[12] D. Tax, *One-class classification: concept learning in the absence of counter-examples*, Ph.D. thesis, Technische Universiteit Delft, Delft, The Netherlands, 2001.

[13] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, Mass, USA, 2001.

[14] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *The Journal of Machine Learning Research*, vol. 5, pp. 1391–1415, 2004.

[15] A. Rakotomamojy and M. Davy, "One-class SVM regularization path and comparison with alpha seeding," in *Proceedings of the 15th European Symposium on Artificial Neural Networks (ESANN '07)*, pp. 271–276, Bruges, Belgium, April 2007.

[16] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, McGraw-Hill, New York, NY, USA, 5th edition, 2005.

[17] M. B. Brown and A. B. Forsythe, "The small sample behavior of some statistics which test the equality of several means," *Technometrics*, vol. 16, no. 1, pp. 129–132, 1974.

[18] B. L. Welch, "On the comparison of several mean values: an alternative approach," *Biometrika*, vol. 38, no. 3-4, pp. 330–336, 1951.

[19] J. Hartung, D. Argaç, and K. H. Makambi, "Small sample properties of tests on homogeneity in one-way Anova and meta-analysis," *Statistical Papers*, vol. 43, no. 2, pp. 197–235, 2002.

[20] W. G. Cochran, "Problems arising in the analysis of a series of similar experiments," *Journal of the Royal Statistical Society*, vol. 4, pp. 102–118, 1937.

[21] W. Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences*, John Wiley & Sons, New York, NY, USA, 1999.

[22] M. Davy, F. Desobry, A. Gretton, and C. Doncarli, "An online support vector machine for abnormal events detection," *Signal Processing*, vol. 86, no. 8, pp. 2009–2025, 2006.

[23] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[24] T. R. Golub, D. K. Slonim, P. Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[25] J. B. Welsh, P. P. Zarrinkar, L. M. Sapinoso, et al., "Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 3, pp. 1176–1181, 2001.

[26] D. T. Ross, U. Scherf, M. B. Eisen, et al., "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics*, vol. 24, no. 3, pp. 227–235, 2000.

[27] U. Scherf, D. T. Ross, M. Waltham, et al., "A gene expression database for the molecular pharmacology of cancer," *Nature Genetics*, vol. 24, no. 3, pp. 236–244, 2000.

[28] M. E. Garber, O. G. Troyanskaya, K. Schluens, et al., "Diversity of gene expression in adenocarcinoma of the lung," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13784–13789, 2001.

[29] A. A. Alizadeh, M. B. Elsen, R. E. Davis, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[30] E. Grall-Maës and P. Beauseroy, "Optimal decision rule with class-selective rejection and performance constraints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. 1, 2009.