Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 906865, 9 pages
doi:10.1155/2009/906865

Research Article

Combining Dissimilarities in a Hyper Reproducing Kernel Hilbert Space for Complex Human Cancer Prediction

Manuel Martín-Merino,[1] Ángela Blanco,[1] and Javier De Las Rivas[2]

[1] *Department of Computer Science, Universidad Pontificia de Salamanca (UPSA), C/Compañía 5, 37002 Salamanca, Spain*
[2] *Cancer Research Center (CIC-IBMCC, CSIC/USAL), Campus Miguel De Unamuno s/n, 37007 Salamanca, Spain*

Correspondence should be addressed to Manuel Martín-Merino, mmartinmac@upsa.es

Received 16 January 2009; Accepted 24 March 2009

Recommended by Dechang Chen

DNA microarrays provide rich profiles that are used in cancer prediction considering the gene expression levels across a collection of related samples. Support Vector Machines (SVM) have been applied to the classification of cancer samples with encouraging results. However, they rely on Euclidean distances that fail to reflect accurately the proximities among sample profiles. Then, non-Euclidean dissimilarities provide additional information that should be considered to reduce the misclassification errors. In this paper, we incorporate in the ν-SVM algorithm a linear combination of non-Euclidean dissimilarities. The weights of the combination are learnt in a (Hyper Reproducing Kernel Hilbert Space) HRKHS using a Semidefinite Programming algorithm. This approach allows us to incorporate a smoothing term that penalizes the complexity of the family of distances and avoids overfitting. The experimental results suggest that the method proposed helps to reduce the misclassification errors in several human cancer problems.

1. Introduction

DNA Microarray technology provides us a way to monitor the expression levels of thousands of genes simultaneously across a collection of related samples. This technology has been applied particularly to the prediction of different types of human cancer with encouraging results [1].

Support Vector Machines (SVM) [2] are powerful machine learning techniques that have been applied to the classification of cancer samples [3]. However, the categorization of different cancer types remains a difficult problem for classical SVM algorithms. In particular, the SVM is based on Euclidean distances that fail to reflect accurately the proximities among the sample profiles [4]. Non-Euclidean dissimilarities misclassify frequently different subsets of patterns because each one reflects complementary features of the data. Therefore, they should be integrated in order to reduce the fraction of patterns misclassified by the base dissimilarities.

In this paper, we introduce a framework to learn a linear combination of non-Euclidean dissimilarities that reflect better the proximities among the sample profiles. Each dissimilarity is embedded in a feature space using the Empirical Kernel Map [5, 6]. After that, learning the dissimilarity is equivalent to optimize the weights of the linear combination of kernels. Several approaches have been proposed to this aim. In [7, 8] the kernel is learnt optimizing an error function that maximizes the alignment between the input kernel and an idealized kernel. However, this error function is not related to the misclassification error and is prone to overfitting. To avoid this problem, [9] learns the kernel by optimizing an error function derived from the Statistical Learning Theory. This approach includes a term to penalize the complexity of the family of kernels considered. This algorithm is not able to incorporate infinite families of kernels and does not overcome the overfitting of the data.

In this paper, the combination of distances is learnt in a (Hyper Reproducing Kernel Hilbert Space) HRKHS following the approach of hyperkernels proposed in [10]. This formalism exhibits a strong theoretical foundation and is less sensitive to overfitting. Moreover, it allow us to work with infinite families of distances. The algorithm has been

applied to the prediction of different kinds of human cancer. The experimental results suggest that the combination of dissimilarities in a Hyper Reproducing Kernel Hilbert Space improves the accuracy of classifiers based on a single distance, particularly for nonlinear problems. Besides, our approach outperforms the Lanckriet formalism specially for multicategory problems and is more robust to overfitting.

This paper is organized as follows. Section 2 introduces the algorithm proposed, the material and the methods employed. Section 3 illustrates the performance of the algorithm in the challenging problem of gene expression data analysis. Finally, Section 4 gets conclusions and outlines future research trends.

2. Material and Methods

2.1. Distances for Gene Expression Data Analysis. An important step in the design of a classifier is the choice of a proper dissimilarity that reflects the proximities among the objects. However, the choice of a good dissimilarity is not an easy task. Each measure reflects different features of the data and the classifiers induced by the dissimilarities misclassify frequently a different set of patterns. In this section, we comment shortly the main differences among several dissimilarities proposed to evaluate the proximity between biological samples considering their gene expression profiles. For a deeper description and definitions see [11].

Let $\mathbf{x} = [x_1, \ldots, x_d]$ be the vectorial representation of a sample where x_i is the expression level of gene i. The *Euclidean distance* evaluates if the gene expression levels differ significantly across different samples:

$$d_{\text{euclid}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}. \tag{1}$$

An interesting alternative is the *cosine dissimilarity*. This measure will become small when the ratio between the gene expression levels is similar for the two samples considered. It differs significantly from the Euclidean distance when the data is not normalized by the L_2 norm:

$$d_{\text{cosine}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}. \tag{2}$$

The *correlation measure* evaluates if the expression level of genes change similarly in both samples. Correlation-based measures tend to group together samples whose expression levels are linearly related. The correlation differs significantly from the cosine if the means of the sample profiles are not zero. This measure is more sensitive to outliers:

$$d_{\text{cor}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{i=1}^{d} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{d} (x_i - \overline{x})^2} \sqrt{\sum_{j=1}^{d} (y_j - \overline{y})^2}}, \tag{3}$$

where $\overline{x}$ and $\overline{y}$ are the means of the gene expression profiles.

The *Spearman rank dissimilarity* is less sensitive to outliers because it computes a correlation between the ranks of the gene expression levels:

$$d_{\text{speam}}(x', y') = 1 - \frac{\sum_{i=1}^{d} (x_i' - \overline{x'})(y_i' - \overline{y'})}{\sqrt{\sum_{i=1}^{d} (x_i' - \overline{x'})^2} \sqrt{\sum_{i=1}^{d} (y_i' - \overline{y'})^2}}, \tag{4}$$

where $\mathbf{x_i}' = \text{rank}(\mathbf{x}_i)$ and $\mathbf{y}_j' = \text{rank}(\mathbf{y}_j)$.

An alternative measure that helps to overcome the problem of outliers is the *Kendall-τ index* which is related to the Mutual Information probabilistic measure [11]:

$$d_{\text{kendall}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{i=1}^{d} \sum_{j=1}^{d} C_{x_{ij}} - C_{y_{ij}}}{d(d-1)}, \tag{5}$$

where $C_{x_{ij}} = \text{sign}(x_i - x_j)$ and $C_{y_{ij}} = \text{sign}(y_i - y_j)$.

Finally, the dissimilarities have been transformed using the inverse multiquadratic kernel because this transformation helps to discover certain properties of the underlying structure of the data [12, 13]. The inverse multiquadratic transformation is based on the inverse multiquadratic kernel defined as follows:

$$k(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{\|\mathbf{x} - \mathbf{y}\|^2 + c^2}}, \tag{6}$$

where c is a smoothing parameter. Considering that $\|\mathbf{x} - \mathbf{y}\|$ is the Euclidean distance, (6) can be rewritten in terms of a dissimilarity as follows:

$$k(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{d_{ij}^2 + c^2}}. \tag{7}$$

The above nonlinear transformation gives more weight to small dissimilarities, particularly when c becomes small.

2.2. ν-Support Vector Machines. Support Vector Machines [2] are powerful classifiers that are able to deal with high dimensional and noisy data keeping a high generalization ability. They have been widely applied in cancer classification using gene expression profiles [1, 14]. In this paper, we will focus on the ν-Support Vector Machines (SVM). The ν-SVM is a reparametrization of the classical C-SVM [2] that allows to interpret the regularization parameter in terms of the number of support vectors and margin errors. This property helps to control the complexity of the approximating functions in an intuitive way. This feature is desirable for the application we are dealing with because the sample size is frequently small and the resulting classifiers are prone to overfitting.

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ be the training set codified in $\mathbb{R}^d$. We assume that each $\mathbf{x}_i$ belongs to one of the two classes labeled by $y_i \in \{-1, 1\}$. The SVM algorithm looks for the linear hyperplane $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x} + b$ that maximizes the margin $\gamma = 2/\|\mathbf{w}\|^2$. γ determines the generalization ability of the SVM. The slack variables ξ_i allow to consider classification errors and are defined as $\xi_i = \max\{0, 1 - y_i f(\mathbf{x}_i)\}$.

For the ν-SVM, the hyperplane that minimizes the prediction error is obtained solving the following optimization problem [2]:

$$\begin{aligned} &\min_{w,\{\xi_i\},\rho} \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{m}\sum_i \xi_i \\ &\text{s.t. } y_i(\langle \mathbf{w},\ \mathbf{x_i}\rangle + b) \geq \rho - \xi_i, \quad i = 1,\ldots,m, \\ &\xi_i \geq 0, \qquad \rho \geq 0 \quad i = 1,\ldots,m, \end{aligned} \tag{8}$$

where ν is an upper bound on the fraction of margin errors and a lower bound on the number of support vectors. Therefore, this parameter controls the complexity of the approximating functions.

The optimization problem can be solved efficiently in the dual space and the discriminant function can be expressed exclusively in terms of scalar products:

$$f(\mathbf{x}) = \sum_{\alpha_i>0} \alpha_i y_i \langle \mathbf{x}, \mathbf{x_i}\rangle + b, \tag{9}$$

where α_i are the Lagrange multipliers in the dual optimization problem. The ν-SVM algorithm can be easily extended to the nonlinear case substituting the scalar products by a Mercer kernel [2]. Besides, non-Euclidean dissimilarities can be incorporated into the ν-SVM via the kernel of dissimilarities [5].

Finally, several approaches have been proposed in the literature to extend the SVM to deal with multiple classes. In this paper, we have followed the one-against-one (OVO) strategy. Let k be the number of classes, in this approach $k(k-1)/2$ binary classifiers are trained and the appropriate class is found by a voting scheme. This strategy compares favorably with more sophisticated methods and it is more efficient computationally than the one-against-rest (OVR) approach [15].

2.3. Empirical Kernel Map. The Empirical Kernel Map allows us to incorporate non-Euclidean dissimilarities into the SVM algorithm using the kernel trick [5, 13].

Let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a dissimilarity and $R = \{p_1,\ldots,p_n\}$ a subset of representatives drawn from the training set. Define the mapping $\phi : \mathcal{F} \to \mathbb{R}^n$ as

$$\phi(z) = D(z,R) = [d(z,p_1), d(z,p_2),\ldots,d(z,p_n)]. \tag{10}$$

This mapping defines a dissimilarity space where feature i is given by $d(\cdot, p_i)$.

The set of representatives R determines the dimensionality of the feature space. The choice of R is equivalent to select a subset of features in the dissimilarity space. Due to the small number of samples in our application, we have considered the whole training set as representatives. Notice that it has been suggested in literature [13] that for small samples reducing the set of representatives does not help to improve the classifier performance.

2.4. Learning a Linear Combination of Dissimilarities in an HRKHS. In order to learn a linear combination of non-Euclidean dissimilarities, we follow the approach of Hyperkernels developed by [10]. To this aim, each distance is embedded in an RKHS via the Empirical Kernel Map presented in Section 2.3. Next, a regularized quality functional is introduced that incorporates an l_2-penalty over the complexity of the family of distances considered. The solution to this regularized quality functional is searched in a Hyper Reproducing Kernel Hilbert Space. This allows to minimize the quality functional using an SDP approach.

Let $X_{\text{train}} = \{x_1, x_2, \ldots, x_m\}$ and $Y_{\text{train}} = \{y_1, y_2, \ldots, y_n\}$ be a finite sample of training patterns where $y_i \in \{-1, +1\}$. Let $\mathcal{K}$ be a family of semidefinite positive kernels. Our goal is to learn a kernel of dissimilarities $k \in \mathcal{K}$ that represents the combination of dissimilarities and minimizes the following empirical quality functional:

$$Q_{\text{emp}}(f, X_{\text{train}}, Y_{\text{train}}) = \frac{1}{m}\sum_{i=1}^{m} l(x_i, y_i, f(x_i)) + \frac{\lambda}{2}||f||^2_{\mathcal{H}}, \tag{11}$$

where l is a loss function, $\|\ \|_{\mathcal{H}}$ is the L_2 norm defined in a reproducing kernel Hilbert space, and λ is a regularization parameter that controls the balance between training error and the generalization ability.

By virtue of the representer theorem [2], we know that (11) can be written as a kernel expansion:

$$Q_{\text{emp}} = \min_{\alpha,k}\left[\frac{1}{m}\sum_{i=1}^{m} l(x_i, y_i, [K\alpha]_i) + \frac{\lambda}{2}\alpha^T K\alpha\right]. \tag{12}$$

However, if the family of kernels $\mathcal{K}$ is complex enough it is possible to find a kernel that achieves zero error overfitting the data. To avoid this problem, we introduce a term that penalizes the kernel complexity in an HRKHS. A rigorous definition of the HRKHS is provided in the appendix:

$$Q_{\text{reg}}(k, X, Y) = Q_{\text{emp}}(k, X, Y) + \frac{\lambda_Q}{2}\|k\|^2_{\underline{H}}, \tag{13}$$

where $\|\ \|_{\underline{\mathcal{H}}}$ is the L_2 norm defined in the Hyper Reproducing Kernel Hilbert space generated by the hyperkernel $\underline{k}$. λ_Q is a regularization parameter that controls the complexity of the resulting kernel.

The following theorem allows us to write the solution to the minimization of this regularized quality functional as a linear combination of hyperkernels in an HRKHS.

Theorem 1 (Representer theorem for Hyper-RKHS [10]). *Let X, Y be the combined training and test set, then each minimizer $k \in \underline{\mathcal{H}}$ of the regularized quality functional $Q_{reg}(k, X, Y)$ admits a representation of the form*

$$k(x, x') = \sum_{i,j=1}^{m} \beta_{ij}\underline{k}\Big(\big(x_i, x_j\big), (x, x')\Big), \tag{14}$$

for all $x,\ x' \in X$, where $\beta_{ij} \in \mathbb{R}$, for each $1 \leq i, j \leq m$.

However, we are only interested in solutions that give rise to positive semidefinite kernels. The following condition over the hyperkernels [10] allows us to guarantee that the solution is a positive semidefinite kernel.

Property 1. *Given a hyperkernel $\underline{k}$ with elements such that for any fixed $\underline{x} \in \underline{X}$, the function $k(x_p, x_q) = \underline{k}(\underline{x}, (x_p, x_q))$, with $x_p, x_q \in \mathcal{X}$, is a positive semidefinite kernel, and $\beta_{ij} \geq 0$ for all $i, j = 1, \ldots, m$, then the kernel*

$$k\left(x_p, x_q\right) = \sum_{i,j=1}^{m} \beta_{ij} \underline{k}\left(x_i, x_j, x_p, x_q\right) \tag{15}$$

is positive semidefinite.

Now, we address the problem of combining a finite set of dissimilarities. As we mentioned in Section 2.3, each dissimilarity can be represented by a kernel using the Empirical Kernel Map. Next, the hyperkernel is defined as

$$\underline{k}(\underline{x}, \underline{x}') = \sum_{i=1}^{n} c_i k_i(\underline{x}) k_i(\underline{x}'), \tag{16}$$

where each k_i is a positive semidefinite kernel of dissimilarities and c_i is a constant ≥ 0.

Now, we show that $\underline{k}$ is a valid hyperkernel. First, $\underline{k}$ is a kernel because it can be written as a dot product $\langle \underline{\Phi}(\underline{x}), \underline{\Phi}(\underline{x}') \rangle$ where

$$\underline{\Phi}(\underline{x}) = (\sqrt{c_1}\ k_1(\underline{x}), \sqrt{c_2}\ k_2(\underline{x}), \ldots, \sqrt{c_n}\ k_n(\underline{x})). \tag{17}$$

Next, the resulting kernel (15) is positive semidefinite because for all $\underline{x}, \underline{k}(\underline{x}, (x_p, x_q))$ is a positive semidefinite kernel and β_{ij} can be constrained to be ≥ 0. Besides, the linear combination of kernels is a kernel and therefore is positive semidefinite. Notice that $\underline{k}(\underline{x}, (x_p, x_q))$ is positive semidefinite if $c_i \geq 0$ and k_i are pointwise positive for training data. Both RBF and multiquadratic kernels verify this condition.

Finally, we show that the resulting kernel is a linear combination of the original k_i. Substituting the expression of the hyperkernel (16) in (15), the kernel is written as

$$k\left(x_p, x_q\right) = \sum_{i,j=1}^{m} \beta_{ij} \sum_{l=1}^{n} c_l k_l\left(x_i, x_j\right) k_l\left(x_p, x_q\right). \tag{18}$$

Now the kernel can be written as a linear combination of base kernels:

$$k\left(x_p, x_q\right) = \sum_{l=1}^{n} \left[c_l \sum_{i,j=1}^{m} \beta_{ij} k_l\left(x_i, x_j\right) \right] k_l\left(x_p, x_q\right). \tag{19}$$

Therefore, the above kernel introduces into the ν-SVM a linear combination of base dissimilarities represented by k_l with coefficients $\gamma_l = c_l \sum_{i,j=1}^{m} \beta_{ij} k_l(x_i, x_j)$.

The previous approach can be extended to an infinite family of distances. In this case, the space that generates the kernel is infinite dimensional. Therefore, in order to work in this space, it is necessary to define a hyperkernel and to optimize it using an HRKHS. Let k be a kernel of dissimilarities. The hyperkernel is defined as follows [10]:

$$\underline{k}(\underline{x}, \underline{x}') = \sum_{i=0}^{\infty} c_i (k(\underline{x}) k(\underline{x}'))^i, \tag{20}$$

where $c_i \geq 0$ and $i = 0, \ldots, \infty$. In this case, the nonlinear transformation to feature space is infinite dimensional. Particularly, we are considering all powers of the original kernels which is equivalent to transform nonlinearly the original dissimilarities:

$$\underline{\Phi}(\underline{x}) = \left(\sqrt{(c_1)} k(\underline{x}), \sqrt{(c_2)} k^2(\underline{x}), \ldots, \sqrt{(c_n)} k^n(\underline{x}) \right), \tag{21}$$

where n is the dimensionality of the space which is infinite in this case. As we mentioned in Section 2.1, nonlinear transformations of a given dissimilarity provide additional information that may help to improve the classifier performance.

As for the finite family, it can be easily shown that $\underline{k}$ is a valid hyperkernel provided that the kernels considered are pointwise positive. The Inverse Multiquadratic kernel satisfies this condition. Next, we derive the hyperkernel expression for the multiquadratic kernel.

Proposition 1 (see [Harmonic Hyperkernel]). *Suppose k is a kernel with range $[0, 1]$ and $c_i = (1 - \lambda_h)\lambda_h^i$, $i \in \mathbb{N}$, $0 < \lambda_h < 1$. Then, computing the infinite sum in (20), one has the following expression for the harmonic hyperkernel:*

$$\underline{k}(\underline{x}, \underline{x}') = (1 - \lambda_h) \sum_{i=0}^{\infty} (\lambda_h k(\underline{x}) k(\underline{x}'))^i = \frac{1 - \lambda_h}{1 - \lambda_h k(\underline{x}) k(\underline{x}')}, \tag{22}$$

λ_h is a regularization term that controls the complexity of the resulting kernel. Particularly, larger values for λ_h give more weight to strongly nonlinear kernels while smaller values give coverage for wider kernels.

In this paper one has considered the inverse multiquadratic kernel defined in (6). Substituting in (22), one gets the inverse multiquadratic hyperkernel:

$$\underline{k}(\underline{x}, \underline{x}\prime) = \frac{1 - \lambda_h}{1 - \lambda_h \left(\left(\|x - x'\|^2 + c^2 \right) \left(\|x'' - x'''\|^2 + c^2 \right) \right)^{-1/2}}, \tag{23}$$

where $\underline{x} = (x, x')$ and $\underline{x}' = (x'', x''')$.

2.5. ν-SVM in an HRKHS. In this section, we detail how to learn the kernel for a ν-Support Vector Machine in an HRKHS. First, we will introduce the optimization problem and next, we will explain shortly how to solve it using an SDP approach.

We start some notation that is used in the ν-SVM algorithm. For $p, q, r \in \mathbb{R}^n$, $n \in \mathbb{N}$ let $r = p \circ q$ be defined as element by element multiplication, $r_i = p_i \times q_i$. The pseudoinverse of a matrix K is denoted by $K^{\dagger}$. Define the hyperkernel Gram matrix $\underline{K}$ by $\underline{K}_{ijpq} = \underline{k}((x_i, x_j), (x_p, x_q))$, the kernel matrix $K = \text{reshape}\,(\underline{K}\beta)$ (reshaping an m^2 by 1 vector, $\underline{K}\beta$, to an $m \times m$ matrix), $Y = \text{diag}(y)$ (a matrix with y on the diagonal and zero otherwise), $G(\beta) = YKY$ (the dependence on β is made explicit), and 1 is a vector of ones.

The ν-SVM considered in this paper uses an l_1 soft margin, where $l(x_i, y_i, f(x_i)) = \max(0, 1 - y_i f(x_i))$. This error

is less sensitive to outliers which are convenient features for microarray datasets. Let ξ_i be the slack variables that allow for errors in the training set. Substituting in (13) Q_{emp} by the one optimized by ν-SVM (8) the regularized quality functional in an HRKHS can be written as

$$\begin{aligned} \min_{\underline{k}\in\underline{H}} \min_{\mathbf{w}\in\mathcal{H}_k} \; & \frac{1}{m}\sum_{i=1}^{m}\xi_i + \frac{1}{2}\|\mathbf{w}\|_{\mathcal{H}}^2 - \nu\rho + \frac{\lambda_Q}{2}\|k\|_{\underline{H}}^2 \\ \text{s.t.} \; & y_i f(x_i) \geq \rho - \xi_i, \quad i = 1,\ldots,m, \\ & \xi_i \geq 0 \quad i = 1,\ldots,m, \end{aligned} \tag{24}$$

where ν is the regularization parameter that achieves a balance between training error and the complexity of the approximating functions and λ_Q is a parameter that penalizes the complexity of the family of kernels considered. The minimization of the previous equation leads to the following SDP optimization problem [10].

$$\min_{\beta,\gamma,\eta,\xi,\chi} \; \frac{1}{2}t_1 - \chi\nu + \frac{1}{m}\xi^T 1 + \frac{\lambda_Q}{2}t_2 \tag{25}$$

$$\text{s.t.} \;\; \chi \geq 0,\; \eta \geq 0,\; \xi \geq 0,\; \beta \geq 0, \tag{26}$$

$$\left\|\underline{K}^{1/2}\beta\right\| \leq t_2,\; 1^T\beta = 1, \tag{27}$$

$$\begin{bmatrix} G(\beta) & z \\ z^T & t_1 \end{bmatrix} \succcurlyeq 0, \tag{28}$$

where $z = \gamma y + \chi 1 + \eta - \xi$

The value of α which optimizes the corresponding Lagrange function is $G(\beta)^{\dagger}z$, and the classification function, $f = \text{sign}(K(\alpha \circ y) - b_{\text{offset}})$, is given by

$$f = \text{sign}\Big(KG(\beta)^{\dagger}(y \circ z) - \gamma\Big), \tag{29}$$

$\underline{K}$ is the hyperkernel defined in Section 2.4 which represents the combination of dissimilarities considered. Finally, the algorithm proposed can be easily extended to deal with multiple classes via a one-against-one approach (OVO). This strategy is simple, more efficient computationally than the OVR, and compares well with more sophisticated multicategory SVM methods [15].

2.6. Implementation. The optimization problem (25) were solved using SeDuMi 1.1R3 [16] and YALMIP [17] SDP optimization packages running under MATLAB.

As in the SDP problem there are m^2 coefficients β_{ij}, the computational complexity is high. However, it can be significantly reduced if the Hyperkernel $\{\underline{k}((x_i, x_j), \cdot) \mid 1 \leq i, j \leq m^2\}$ is approximated by a small fraction of terms, $p \ll m^2$ for a given error. In particular, we have chosen an $m \times p$ truncated lower triangular matrix G which approximate the hyperkernel matrix to an error $\delta = 10^{-6}$ using the incomplete Cholesky factorization method [18].

2.7. Datasets and Preprocessing. The gene expression datasets considered in this paper correspond to several human cancer problems and exhibit different features as shown in Table 1. We have considered both, binary and multicategory problems with a broad range of signal to noise ratio (Var/Samp.), different number of samples, and varying priors for the larger category. All the datasets are available from the Broad Institute of MIT and Harvard http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi/. Next we detail the features and preprocessing applied to each dataset.

Table 1: Features of the different cancer datasets

	Clases	Samples	Genes	Var/Samp.	Priors %
Lymphoma DLBCL	2	77	6817	88	75.3
Lymphoma MLBCL/DLBCL	2	210	44928	213	84
Breast cancer LN	2	49	7129	145	51
Medulloblastoma	2	60	7129	119	65
Breast cancer B	3	49	1213	24.7	52
DLBCL survival C	4	58	3795	65.4	27
DLBCL survival D	4	129	3795	29.4	38

The first dataset was obtained from 77 patients with (diffuse large B-cell lymphoma) DLBCL (58 samples) or FL (follicular lymphoma) (19 samples) and they were subjected to transcriptional profiling using oligonucleotide Affymetrix gene chip *hu*68000 containing probes for 6817 genes [19]. The second dataset consists of frozen tumors specimens from newly diagnosed, previously untreated MLBCL patients (34 samples) and DLBCL patients (176 samples). They were hybridized to Affymetrix *hgu*133*b* gene chip containing probes for 44000 genes [20]. In both cases the raw intensities have been normalized using the rma algorithm [21] available from Bioconductor package [11]. The third problem we address concerns the clinically important issue of metastatic spread of the tumor. The determination of the extent of lymph node involvement in primary breast cancer is the single most important risk factor in disease outcome and here the analysis compares primary cancers that have not spread beyond the breast to ones that have metastasized to axillary lymph nodes at the time of diagnosis. We identified tumors as "reported negative" (24) when no positive lymph nodes were discovered and "reported positive" (25) for tumors with at least three identifiably positive nodes [22]. All assays used the human HuGeneFL Genechip microarray containing probes for 7129 genes. The fourth dataset [23] address the clinical challenge concerning medulloblastoma due to the variable response of patients to therapy. Whereas some patients are cured by chemotherapy and radiation, others have progressive disease. The dataset consists of 60 samples containing 39 medulloblastoma survivors and 21 treatment failures. Samples were hybridized to Affymetrix HuGeneFL arrays containing 5920 known genes and 897 expressed sequence tags.

All the datasets have been standarized subtracting the median and dividing by the Inter-quantile range. The rescaling were performed based only on the training set to avoid bias.

Regarding the identification of multiple classes of cancer we have considered three different datasets. The first one consists of 49 samples of Breast Cancer generated using 1-channel oligonucleotide Affymetrix HuGeneFl [1]. The second and third datasets consist of 58 and a129 samples from Diffuse large B-cell lymphoma with survival data. Fourth different subclasses can be identified. Data preparatory steps have been performed by the authors of the primary study [1]. The 10% oligonucleotides with smaller Interquantile Range were filtered to remove genes with expression level constant across samples.

2.8. Performance Evaluation. In order to assure an honest evaluation of all the classifiers we have performed a double loop of crossvalidation [15]. The outer loop is based on stratified tenfold cross-validation that iteratively splits the data in ten sets, one for testing and the others for training. The inner loop perform stratified ninefold cross-validation over the training set and is used to estimate the optimal parameters avoiding overfitting. The stratified variant of cross-validation keeps the same proportion of patterns for each class in training and test sets. This is necessary in our problem because the class proportions are not equal. Finally, the error measure considered to evaluate the classifiers has been accuracy. This metric computes the proportion of samples misclassified. The accuracy is easy to interpret and allows us to compare with the results obtained by previously published studies.

2.9. Parameters for the Classification Algorithm. The parameters for the ν-SVM and for the classifiers based on a linear combination of dissimilarities have been set up by a nested stratified tenfold crossvalidation procedure [15]. This method avoids the overfitting as is described in Section 2.8 and takes into account the asymmetric distribution of class priors.

For the ν-SVM we have considered both, linear and inverse multiquadratic kernels. The optimal parameters have been obtained by a grid search strategy over the following set of values: $\nu = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\sigma = \{d/2, d, 2d\}$, where d denotes the dimensionality of the input space.

Additionally, for the finite family of distances $c_i = 1/M$ where M is the number of dissimilarities considered, and $\lambda_Q = 1$ because the misclassification errors are hardly sensitive to the regularization parameter that controls the kernel complexity. Finally, for the infinite family of dissimilarities, the regularization parameter λ_h in the Harmonic hyperkernel (22) has been set up to 0.6 which gives an adequate coverage of various kernel widths. Smaller values emphasizes only wide kernels. All the base kernel of dissimilarities have been normalized so that all ones have the same scale.

Regarding the Lanckriet [9] formalism that allows to combine a finite set of dissimilarities, several values for the regularization parameter C have been tried, $C = \{0.1, 1, 10, 100, 1000\}$. A grid search strategy has been applied to determine the best values for both, the kernel parameters and the regularization parameter. The kernel matrices have been normalized by the trace as recommended in the original paper.

Table 2: Accuracy for the ν-SVM using a linear combination of non-Euclidean dissimilarities in an HRKHS. The ν-SVM based on the best distance and coordinates and the Lanckriet formalism have been taken as a reference.

Technique	Limphoma	Limphoma cell B	Breast LN	Brain
ν-SVM (coordinates)	6.66%	7.14%	8.16%	16.6%
ν-SVM (best distance)	6.66%	5.71%	8.16%	13.3%
ν-SVM (nonlinear kernel)	6.25%	5.71%	8.16%	11.6%
Lanckriet (finite family)	5%	7.62%	8.16%	11.67%
Finite family of distances	5%	7.14%	10%	10%
Infinite family of distances	5%	5.71%	8%	8.33%

2.10. Gene Selection. Gene selection can improve significantly the classifier performance [24]. Therefore, we have evaluated the classifiers for the following subsets of genes $\{280, 146, 101, 56, 34\}$. The ν-SVM is robust against noise and is able to deal with high dimensional data. However, the empirical evidence suggests that considering a larger subset of genes or even the whole set of genes increases the misclassification errors.

The genes are ranked according to the ratio of between-group to within-group sums of squares defined in [25]:

$$BW(j) = \frac{\sum_i \sum_k I(y_i = k)\left(\overline{x}_{\cdot j}^{(k)} - \overline{x}_{\cdot j}\right)^2}{\sum_i \sum_k I(y_i = k)\left(x_{ij} - \overline{x}_{\cdot j}^{(k)}\right)^2}, \tag{30}$$

where $\overline{x}_{\cdot j}^{(k)}$ and $\overline{x}_{\cdot j}$ denote "respectively" the average expression level of gene j for class k and the overall average expression level of gene j across all samples, y_i denotes the class of sample i, and $I(\cdot)$ is the indicator function. Next, the top ranked genes are chosen. This feature selection method is simple but compares well with more sophisticated methods [24]. Finally, the ranking of genes has been carried out considering only the training set to avoid bias. Therefore, feature selection is repeated in each iteration of cross-validation.

3. Results and Analysis

The algorithms proposed have been applied to the identification of several cancer human samples using microarray gene expression data.

First, we address several binary categorization problems.

Table 2 reports the accuracy for the two combination approaches proposed in this paper. The first one considers the finite set of dissimilarities introduced in Section 2.1. The second one considers an infinite family of distances obtained by transforming nonlinearly the base dissimilarities

Table 3: Accuracy for the ν-SVM using a linear combination of non-Euclidean dissimilarities in an HRKHS. The ν-SVM based on the best distance, the classical ν-SVM, and the Lanckriet formalism have been taken as a reference.

Technique	Breast B	DLBCL C	DLBCL D
ν-SVM (Coordinates)	10.20%	6.89%	12.96%
ν-SVM (Best Distance)	8.6%	6.89%	14.81%
ν-SVM (Nonlinear kernel)	8.16%	6.89%	12.96%
Lanckriet (finite family)	8%	10.3%	25.2%
Infinite family of distances	6%	5.33%	16%

to feature space. We have compared with the ν-SVM based on the best distance (linear and nonlinear kernel) and the classical ν-SVM. The performance for the Lanckriet formalism [9] that allow us to incorporate a finite linear combination of dissimilarities is also reported.

Before computing the kernel of dissimilarities, all the distances have been transformed using the multiquadratic kernel introduced in Section 2.1. This nonlinear transformation helps to improve the accuracy for all the techniques evaluated. From the analysis of Table 2, the following conclusions can be drawn.

(i) The ν-SVM based on a finite set of distances improves the ν-SVM based on the best dissimilarity for brain prognosis and Lymphoma datasets. The error is not reduced for Lymphoma cell B and Breast LN. This may be explained because the ratio (var/samp.) in Table 1 suggests that both datasets are quite noisy and nonlinear. The combination of a finite set of dissimilarities is not able to improve the separation between classes and increases slightly the overfitting of the data. Similarly, our algorithm helps to improve the SVM based on coordinates, particularly for the previous problems. We also report that working directly from a dissimilarity matrix may help to reduce the misclassification errors.

(ii) The infinite family of distances outperforms the ν-SVM based on the best distance disregarding the kernel considered for all the datasets. The improvement is more relevant in brain cancer prognosis. Brain cancer prognosis is a complex problem according to the original study [23] and the nonlinear transformations of the dissimilarities help to reduce the misclassification errors. Besides, the infinite family improves the accuracy of the finite family of distances particularly for lymphoma cell B and Breast LN. This suggests that both datasets are nonlinear.

(iii) The Lanckriet formalism and the finite family of dissimilarities perform similarly. However, the infinite family of distances outperforms the Lanckriet formalism particularly for brain and Lymphoma cell B which are more complex problems.

(iv) The best distance depends on the dataset considered.

Next we move to the categorization of multiple cancer types.

Table 3 compares the proposed algorithms with ν-SVM based on the best distance (linear and nonlinear kernel) and the classical ν-SVM. The accuracy for the Lanckriet formalism has also been reported. Our approach considers an infinite family of distances obtained by transforming nonlinearly the base dissimilarities to feature space.

Before computing the kernel of dissimilarities, all the distances have been transformed using the multiquadratic kernel introduced in Section 2.1. From the analysis of Table 3, the following conclusions can be drawn.

(i) The combination of non-Euclidean dissimilarities helps to improve the SVM based on the best dissimilarity disregarding the kernel considered for the two first datasets. The error is slightly larger for the third dataset which may suggest that the problem is linear.

(ii) Our algorithm improves the SVM based on coordinates. The experimental results suggest that the nonlinear transformations of the dissimilarities help to increase the separation among classes.

(iii) The Hyperkernel classifier outperforms the Lanckriet formalism for multicategory problems. As the number of classes growths the number of samples per class comes down and the Lanckriet formalism seems to be less robust to overfitting.

Finally, notice that our algorithm allow us to work with applications in with only a dissimilarity is defined. Moreover, we avoid the complex task of choosing a dissimilarity that reflects properly the proximities among the sample profiles.

4. Conclusions

In this paper, we propose two methods to incorporate in the ν-SVM algorithm a linear combination of non-Euclidean dissimilarities. The family of distances is learnt in a (Hyper Reproducing Kernel Hilbert Space) HRKHS using a Semidefinite Programming approach. A penalty term has been added to avoid the overfitting of the data. The algorithm has been applied to the classification of complex cancer human samples. The experimental results suggest that the combination of dissimilarities in a Hyper Reproducing Kernel Hilbert Space improves the accuracy of classifiers based on a single distance particularly for nonlinear problems. Besides, this approach outperforms the Lanckriet formalism specially for multi-category problems and is more robust to overfitting. Future research trends will focus on learning the combination of dissimilarities for other classifiers such as k-NN.

Appendix

In this section we define rigorously the Hyper-Reproducing Kernel Hilbert Spaces. First, we define a Reproducing Kernel Hilbert Space.

Definition 1 (see [Reproducing Kernel Hilbert Space]). Let $\mathcal{X}$ be a nonempty set and $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. Let $\langle \cdot, \cdot \rangle$ be a dot product in $\mathcal{H}$ which induces a norm as $\|f\| = \sqrt{\langle f, f \rangle}$. $\mathcal{H}$ is called an RKHS if there is a function $k : \mathcal{X} \times \mathcal{X}$ with the following properties:

(i) k has the reproducing property $\langle f, k(x, \cdot) \rangle = f(x)$ for all $f \in \mathcal{H}, x \in \mathcal{X}$;

(ii) k spans $\mathcal{H}$, that is, $\mathcal{H} = \overline{\text{span}\{k(x, \cdot) \mid x \in \mathcal{X}\}}$, where $\overline{X}$ is the completion of the set X;

(iii) k is symmetric, that is, $k(x, y) = k(y, x)$.

Next, we introduce the Hyper Reproducing Kernel Hilbert Space.

Definition 2 (see [Hyper-Reproducing Kernel Hilbert Space]). Let $\mathcal{X}$ be a nonempty set and $\underline{\mathcal{X}} = \mathcal{X} \times \mathcal{X}$ be the Cartesian product. Let $\underline{\mathcal{H}}$ be the Hilbert space of functions $k : \underline{\mathcal{X}} \to \mathbb{R}$ with a dot product $\langle \cdot, \cdot \rangle$ and a norm $\|k\| = \sqrt{\langle k, k \rangle}$. $\underline{\mathcal{H}}$ is a Hyper Reproducing Kernel Hilbert Space if there is a hyperkernel $\underline{k} : \underline{X} \times \underline{X} \to \mathbb{R}$ with the following properties:

(i) $\underline{k}$ has the reproducing property $\langle k, \underline{k}(\underline{x}, \cdot) \rangle = k(\underline{x})$ for all $k \in \underline{\mathcal{H}}$;

(ii) $\underline{k}$ spans $\underline{H} = \overline{\text{span}\{\underline{k}(\underline{x}, \cdot) \mid \underline{x} \in \underline{X}\}}$;

(iii) $\underline{k}(x, y, s, t) = \underline{k}(y, x, s, t)$ for all $x, y, s, t \in \mathcal{X}$.

Acknowledgments

The authors would like to thanks two anonymous referees by their useful comments and suggestions. Financial support from Grant S02EIA-07L01 is gratefully appreciated.

References

[1] Y. Hoshida, J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Subclass mapping: identifying common subtypes in independent disease data sets," *PLoS ONE*, vol. 2, no. 11, article e1195, pp. 1–8, 2007.

[2] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[3] T. S. Furey, N. Cristianini, 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*, vol. 16, no. 10, pp. 906–914, 2000.

[4] Á. Blanco, M. Martín-Merino, and J. De Las Rivas, "Combining dissimilarity based classifiers for cancer prediction using gene expression profiles," *BMC Bioinformatics*, vol. 8, supplement 8, article S3, pp. 1–2, 2007.

[5] K. Tsuda, "Support vector classifier with asymmetric kernel function," in *Proceedings of the 7th European Symposium on Artificial Neural Networks (ESANN '99)*, pp. 183–188, Bruges, Belgium, April 1999.

[6] B. Schölkopf, J. Weston, E. Eskin, C. Leslie, and W. Stafford Noble, "A kernel approach for learning from almost orthogonal patterns," in *Proceedings of the 13th European Conference on Machine Learning (ECML '02)*, vol. 2430 of *Lecture Notes in Computer Science*, pp. 511–528, Springer, Helsinki, Finland, August 2002.

[7] N. Cristianini, J. Kandola, J. Elisseeff, and A. Shawe-Taylor, "On the kernel target alignment," *Journal of Machine Learning Research*, vol. 1, pp. 1–31, 2002.

[8] J. Kandola, J. Shawe-Taylor, and N. Cristianini, "Optimizing kernel alignment over combinations of kernels," Tech. Rep. NC-TR-02-121, NeuroCOLT, London, UK, 2002.

[9] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[10] C. S. Ong, A. J. Smola, and R. C. Williamson, "Learning the kernel with hyperkernels," *Journal of Machine Learning Research*, vol. 6, pp. 1043–1071, 2005.

[11] R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, Berlin, Germany, 2006.

[12] G. Wu, E. Y. Chang, and N. Panda, "Formulating distance functions via the kernel trick," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 703–709, Chicago, Ill, USA, August 2005.

[13] E. Pekalska, P. Paclick, and R. Duin, "A generalized kernel approach to dissimilarity-based classification," *Journal of Machine Learning Research*, vol. 2, pp. 175–211, 2001.

[14] 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.

[15] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[16] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, no. 1–4, pp. 625–653, 1999.

[17] J. Löfberg, YALMIP, yet another LMI parser, 2002, http://control.ee.ethz.ch/~joloef/wiki/pmwiki.php.

[18] S. Fine and K. Scheinberg, "Efficient SVM training using low-rank kernel representations," *Journal of Machine Learning Research*, vol. 2, pp. 243–264, 2001.

[19] M. A. Shipp, K. N. Ross, P. Tamayo, et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[20] K. J. Savage, S. Monti, J. L. Kutok, et al., "The molecular signature of mediastinal large B-cell lymphoma differs from that of other diffuse large B-cell lymphomas and shares features with classical Hodgkin lymphoma," *Blood*, vol. 102, no. 12, pp. 3871–3879, 2003.

[21] R. A. Irizarry, B. Hobbs, F. Collin, et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[22] M. West, C. Blanchette, H. Dressman, et al., "Predicting the clinical status of human breast cancer by using gene expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 20, pp. 11462–11467, 2001.

[23] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[24] I. B. Jeffery, D. G. Higgins, and A. C. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," *BMC Bioinformatics*, vol. 7, article 359, pp. 1–16, 2006.

[25] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–86, 2002.