Classification of Multiple Cancer Types in a Hyper Reproducing Kernel Hilbert Space

Ángela Blanco and Manuel Martín-Merino and Javier De Las Rivas

***Abstract*— The classification of multiple cancer types based on the gene expression profiles is a challenging task. Support Vector Machines (SVM) have been applied to this aim but they rely on Euclidean distances that fail to reflect accurately the proximities among sample profiles.**

In this paper, we incorporate in the classical ν-SVM algorithm a linear combination of non-Euclidean dissimilarities. The weights of the combination are learnt in a HRKHS (Hyper Reproducing Kernel Hilbert Space) using an efficient Semidefinite Programming algorithm. This approach allow 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 significantly the misclassification errors in several human cancer problems.

I. INTRODUCTION

DNA Microarray technology provide 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 [20]. Support Vector Machines (SVM) [17] are powerful machine learning techniques that have been applied to the classification of cancer samples [6]. 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 [3], [11]. Besides, non-Euclidean dissimilarities provide complementary information that should be considered in order to reduce the misclassification errors [1].

In this paper, we introduce a method to learn a linear combination of non Euclidean dissimilarities that reflect better the proximities among the sample profiles. The dissimilarity is embedded in a feature space using the Empirical Kernel Map [16]. Next, we apply a family of non-linear transformations to this kernel of dissimilarities using the hyperkernel formalism. Each non-linear transformation gives rise to a non-Euclidean dissimilarity. 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 [2], [10] 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, [13] 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. Therefore, in this paper the combination of distances is learnt in a HRKHS (Hyper Reproducing Kernel Hilbert Space) following the approach of hyperkernels proposed in [18]. 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 with remarkable results. This paper is organized as follows: *Section II* introduces the distances used to measure the similarities in gene expression data . *Section III* presents the method proposed to combine the distances. *Section IV* illustrates the performance of the algorithm in the challenging problem of gene expression data analysis. Finally, *Section V* gets conclusions.

Ángela Blanco and Manuel Martín-Merino are with Universidad Pontificia de Salamanca (UPSA) C/Compañía 5, 37002, Salamanca, Spain ablancogo@upsa.es, mmartinmac@upsa.es

Javier De Las Rivas is with Cancer Research Center (CIC-IBMCC, CSIC/USAL) Salamanca, Spain jrivas@usal.es

II. 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. Therefore, different dissimilarities provide complementary information. In this paper, we have considered an infinite family of dissimilarities.

Finally, the dissimilarities have been transformed using the inverse multiquadratic kernel [15] because this transformation helps to discover certain properties of the underlying structure of the data [7], [14].

A. Empirical Kernel Map

Now we introduce the Empirical Kernel Map that allow us to incorporate non-Euclidean dissimilarities into the SVM algorithm using the kernel trick [15], [14].

Let d: $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a dissimilarity [14] and $R = \{p_1, \ldots, p_n\}$ a subset of representatives drawn from the training set. Define the mapping $\phi : \mathcal{F} \rightarrow \mathbb{R}^n$ as:

$$\phi(z) = D(z, R) = [d(z, p_1), d(z, p_2), \ldots, d(z, p_n)] \quad (1)$$

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

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 training samples in our application, we have considered the whole sample as a representative set [14].

III. LEARNING A LINEAR COMBINATION OF DISSIMILARITIES

In order to incorporate a potentially infinite family of non Euclidean dissimilarities into the SVM, we follow the approach of Hyperkernels developed by [18]. To this aim, a given distance such as χ^2 is embedded in a RKHS via the Empirical Kernel Map [14]. Next, this dissimilarity is non-linearly transformed to a feature space and a regularized quality functional is introduced that incorporates a 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 a SDP approach.

A. Multi-class ν-Support Vector Machines

Support Vector Machines [17] 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 [20], [19]. In this paper, we will focus on the ν-Support Vector Machines (SVM). The ν-SVM is a reparametrization of the classical C-SVM [15] 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 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be the training set codified in $\mathbb{R}^d$. We assume that each $\boldsymbol{x}_i$ belongs to one of the two classes labeled by $y_i \in \{-1, 1\}$. The SVM algorithm looks for the linear hyperplane $f(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{x} + b$ that maximizes the margin $\gamma = 2/\|\boldsymbol{w}\|^2$. γ determines the generalization ability of the SVM. The slack variables ξ_i allow to consider classification errors and are defined as: $\xi_i = \max\{0, 1 - y_i f(\boldsymbol{x}_i)\}$.
For the ν-SVM, the hyperplane that minimizes the prediction error is obtained solving the following optimization problem [17]:

$$\begin{aligned} \min_{w,\{\xi_i\},\rho} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 - \nu\rho + \frac{1}{m}\sum_i \xi_i \\ \text{s. t.} \quad & y_i(<\boldsymbol{w}, \boldsymbol{x}_i> + b) \geq \rho - \xi_i \quad i = 1, \ldots, m \\ & \xi_i \geq 0, \ \rho \geq 0 \quad i = 1, \ldots, m \end{aligned} \tag{2}$$

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,

$$f(\boldsymbol{x}) = \sum_{\alpha_i > 0} \alpha_i y_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle + b \tag{3}$$

where α_i are the Lagrange multipliers in the dual optimization problem. The ν-SVM algorithm can be easily extended to the non-linear case substituting the scalar products by a Mercer kernel [17]. Besides, non-Euclidean dissimilarities can be incorporated into the ν-SVM via the kernel of dissimilarities.

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 appropiate 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 [21].

B. Learning the Kernel in a HRKHS

First, we define a 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 a RKHS if there is a function $k : \mathcal{X} \times \mathcal{X}$ with the following properties:

- k has the reproducing property $\langle f, k(x, \cdot) \rangle = f(x)$ for all $f \in \mathcal{H}$, $x \in \mathcal{X}$
- k spans $\mathcal{H}$, i.e. $\mathcal{H} = \overline{span\{k(x, \cdot) | x \in \mathcal{X}\}}$, where $\overline{X}$ is the completion of the set X.

Next, we introduce the 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:

- $\underline{k}$ has the reproducing property $\langle k, \underline{k}(\underline{x}, \cdot) \rangle = k(\underline{x})$ for all $k \in \underline{\mathcal{H}}$
- $\underline{k}$ spans $\underline{H} = \overline{span\{\underline{k}(\underline{x}, \cdot) | \underline{x} \in \underline{X}\}}$
- $\underline{k}(x, y, s, t) = \underline{k}(y, x, s, t)$ for all $x, y, s, t \in \mathcal{X}$.

Let $X_{train} = \{x_1, x_2, \ldots, x_m\}$ and $Y_{train} = \{y_1, y_2, \ldots, y_n\}$ 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 $k \in \mathcal{K}$ that minimizes the following empirical quality functional :

$$Q_{emp}(f, X_{train}, Y_{train}) = \frac{1}{m}\sum_{i=1}^m l(x_i, y_i, f(x_i)) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \tag{4}$$

By virtue of the representer theorem [17], we know that equation (4) can be written as a kernel expansion:

$$Q_{emp} = \min_{\alpha, k} \left[\frac{1}{m}\sum_{i=1}^m l(x_i, y_i, [K\alpha]_i) + \frac{\lambda}{2}\alpha^T K \alpha \right] \tag{5}$$

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 penalty term in a HRKHS:

$$Q_{reg}(k, X, Y) = Q_{emp}(k, X, Y) + \frac{\lambda_Q}{2}\|k\|^2_{\underline{H}} \quad (6)$$

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

Theorem 1 (Representer theorem for Hyper-RKHS): 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:

$$k(x, x') = \sum_{i,j=1}^{m} \beta_{ij}\underline{k}((x_i, x_j), (x, x')) \quad (7)$$

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 [18] allow 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

$$k(x_p, x_q) = \sum_{i,j=1}^{m} \beta_{ij}\underline{k}(x_i, x_j, x_p, x_q) \quad (8)$$

is positive semidefinite.

Now, we detail how to incorporate a potentially infinite family of non-Euclidean dissimilarities via a hyperkernel. Let k be a kernel of dissimilarities. The hyperkernel is defined as follows [18]:

$$\underline{k}(\underline{x}, \underline{x}') = \sum_{i=0}^{\infty} c_i(k(\underline{x})k(\underline{x}'))^i \quad (9)$$

where $c_i \geq 0$ and $i = 0, \ldots, \infty$. In this case, the non-linear transformation to feature space is infinite dimensional. Particularly, we are considering all powers of the original kernels which is equivalent to transform non-linearly the original dissimilarities. As we have mentioned in section II, non linear transformations of a given dissimilarity provide complementary information of the data.

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

Proposition 1 (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, we have the following expression for the harmonic hyperkernel:

$$\underline{k}(\underline{x}, \underline{x}') = (1-\lambda_h)\sum_{i=0}^{\infty}(\lambda_h k(\underline{x})k(\underline{x}'))^i = \frac{1-\lambda_h}{1-\lambda_h k(\underline{x})k(\underline{x}')}, \quad (10)$$

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 non-linear kernels.

Considering the inverse multiquadratic kernel $\left(k(x, x') = 1/\sqrt{\|\boldsymbol{x} - \boldsymbol{x'}\|^2 + c^2}\right)$ in equation (10), we get the inverse multiquadratic hyperkernel:

$$\underline{k}(\underline{x}, \underline{x}') = \frac{1-\lambda_h}{1-\lambda_h((\|x - x'\|^2 + c^2)(\|x'' - x'''\|^2 + c^2))^{-1/2}} \quad (11)$$

C. Multi-class ν-SVM in a HRKHS

In this section, we detail how to learn the kernel for a ν-Support Vector Machine in a HRKHS. First, we will introduce the optimization problem and next, we will explain shortly how to solve it using a 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 pseudo-inverse 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 $K = reshape(\underline{K}\beta)$(reshaping an m^2 by 1 vector, $\underline{K}\beta$, to an $m \times m$ matrix), Y = diag(y) (a matrix with y on the diagonal and zero otherwise), $G(\beta) = YKY$ (the dependence on β is made explicit) and $\mathbf{1}$ a vector of ones.

The ν-SVM considered in this paper uses an l_1 soft margin, where $l(x_i, y_i, f(x_i)) = \max(0, 1 - y_i f(x_i))$. Let ξ_i be the slack variables that allow for errors in the training set. Substituting in equation (6) Q_{emp} by the one optimized by ν-SVM (2) the regularized quality functional in a HRKHS can be written as:

$$\begin{aligned} \min_{k \in \underline{H}} \min_{\boldsymbol{w} \in \mathcal{H}_k} \quad & \frac{1}{m}\sum_{i=1}^{m}\xi_i + \frac{1}{2}\|\boldsymbol{w}\|^2_{\mathcal{H}} - \nu\rho + \frac{\lambda_Q}{2}\|k\|^2_{\underline{H}} \\ \text{subject to} \quad & y_i f(x_i) \geq \rho - \xi_i \\ & \xi_i \geq 0 \quad i = 1, \ldots, m \end{aligned} \quad (12)$$

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 [18].

$$\min_{\beta, \gamma, \eta, \xi, \chi} \quad \frac{1}{2}t_1 - \chi\nu + \frac{1}{m}\xi^T\mathbf{1} + \frac{\lambda_Q}{2}t_2 \quad (13)$$

$$\text{subject to} \quad \chi \geq 0,\ \eta \geq 0,\ \xi \geq 0,\ \beta \geq 0 \quad (14)$$

$$\|\underline{K}^{\frac{1}{2}}\beta\| \leq t_2,\ \mathbf{1}^T\beta = 1 \quad (15)$$

$$\begin{bmatrix} G(\beta) & z \\ z^T & t_1 \end{bmatrix} \succeq 0 \quad (16)$$

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

The value of α which optimizes the corresponding

TABLE I

EMPIRICAL RESULTS FOR THE ν-SVM USING A LINEAR COMBINATION OF NON EUCLIDEAN DISSIMILARITIES IN A HRKHS. THE ν-SVM BASED ON THE BEST DISTANCE AND THE CLASSICAL ν-SVM 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%
Infinite family of distances	6%	5.33%	16%

Lagrange function is $G(\beta)^{\dagger}z$, and the classification function, $f = sign(K(\alpha \circ y) - b_{offset})$, is given by $f = sign(KG(\beta)^{\dagger}(y \circ z) - \gamma)$. $\underline{K}$ is the hyperkernel defined in section III-B which represents the combination of dissimilarities considered.

Now, as we mentioned in section III-A, the algorithm proposed can be easily extended to deal with multiple classes via a one-against-one approach (OVO).

IV. EXPERIMENTAL RESULTS

In this section, the method proposed is applied to the identification of several cancer human samples using microarray gene expression data.

Three benchmark gene expression datasets have been considered. The first problem consists of 98 samples of Breast Cancer generated using a two channel microarray. The second and third datasets we address consist of 58 and a129 samples from Diffuse large B cell lymphoma with survival data. The technology applied to obtain all datasets was Affymetrix [20]. The problems addressed in this paper, consider multiple tumor classes which is a more difficult problem than just the identification of cancer samples. It is expected that the information required to solve this kind of problems will be larger.

Due to the large number of genes, samples are codified in a high dimensional and noisy space. Therefore, the non-linear transformations of the dissimilarities are affected by the 'curse of dimensionality' and the correlation among them becomes large [11]. To avoid this problem and to increase the diversity among the non-linear transformations, we have reduced aggressively the number of genes using the standard F-statistics [8].

All the datasets have been standarized because this transformation help to reduce the misclassification errors for all the methods proposed.

The ν, λ_h and λ_Q are regularization parameters and they have been set up by ten fold-crossvalidation [12]. Similarly, the base kernel parameters have been set up by cross-validation. For each dataset, the optimal value of the parameters and the number of genes have been chosen using a greed search strategy.

In order to reduce the computational burden, we have approximated the hyperkernel matrix using the incomplete Cholesky factorization method [5].

Table I compares the proposed algorithms with ν-SVM based on the best distance and the classical ν-SVM. Our approach considers an infinite family of distances obtained by transforming non linearly the base dissimilarities to feature space. This has been done using a hyperkernel defined in the space of kernels itself (see section III-B). Before computing the kernel of dissimilarities, all the distances have been transformed using the multiquadratic kernel introduced in section II. From the analysis of table I, the following conclusions can be drawn:

- The best distance depends on the dataset considered.
- The combination of non-Euclidean dissimilarities help to improve significantly the SVM based on the best dissimilarity particularly for the two first datasets.
- Our algorithm improves the SVM based on coordinates. We also report that working directly from a dissimilarity matrix helps to reduce the misclassification errors.
  The experimental results suggest that the non-linear transformations of the dissimilarities help to discover certain features of the data. Besides, the regularization parameter λ_Q avoids the choice of too complex kernels and the overfitting of the data.

Finally, notice that our algorithm allow us to work with applications in with only a dissimilarity is defined.

V. CONCLUSIONS

In this paper, we propose a method to incorporate in the multiclass ν-SVM algorithm a linear combination of non-Euclidean dissimilarities . The family of distances is learnt in a HRKHS (Hyper Reproducing Kernel Hilbert Space) using an efficient 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 method proposed improves significantly the misclassification error of the ν-SVM based on the best distance and the classical ν-SVM based on coordinates.

REFERENCES

[1] A. Blanco and M. Martín-Merino and J. De Las Rivas, "Combining dissimilarity based classifiers for cancer prediction using gene expression profiles", *BMC Bioinformatics*, BioMed Central Ltd, London, UK, 2007.

[2] N. Cristianini and J. Kandola and A. Elisseeff and J. Shawe-Taylor, "On the Kernel Target Alignment", *On Kernel Target Alignment*,1,1-31,2002.

[3] S. Drãghici, *Data Analysis Tools for DNA Microarrays*. New York: Chapman & Hall/CRC Press, 2003.

[4] S. Dudoit and J. Fridlyand and T. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data", *Journal of the American Statistical Association*, vol. 97, pp. 77-87, 2002.

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

[6] T. Furey and N. Cristianini and N. Duffy and D. Bednarski and 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.

[7] W. Gang and E. Y. Chang and N. Panda, "Formulating Distance Functions via the Kernel Trick", In *Proc. of the11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illlinois, USA, August, 2005.

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

[9] I. Guyon and J. Weston and S. Barnhill and V. Vapnik, "Gene selection for cancer classification using support vector machines", *Machine Learning*, vol. 46, pp. 389-422, 2002.

[10] J. Kandola and J. Shawe-Taylor and N. Cristianini, "Optimizing kernel alignment over combinations of kernels", *NeuroCOLT*, Technical Report, 2002.

[11] M. Martín-Merino and A. Muñoz, "Self organizing map and sammon mapping for asymmetric proximities", *Neurocomputing*, vol. 63, pp. 171-192, 2005.

[12] A. Molinaro and R. Simon and R. Pfeiffer, "Prediction error estimation: a comparison of resampling methods", *Bioinformatics*, vol. 21, no. 15, pp. 3301-3307, 2005.

[13] G. Lanckriet and N. Cristianini and P. Barlett and L. El Ghaoui and M. Jordan, "Learning the kernel matrix with semidefinite programming",*Journal of Machine Learning Research*,3,27-72,2004.

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

[15] B. Schölkopf and A. Smola, *Learning with Kernels*. MIT Press, USA, 2002.

[16] B. Schölkopf and K. Tsuda and J. P. Vert, "Kernel methods in computational biology", MIT Press, 2004.

[17] V. Vapnik,*Statistical Learning Theory*. New York: John Wiley & Sons, 1998.

[18] Soon Ong, C., Smola, A. and Williamson,R.:Learning the Kernel with Hyperkernels, *Journal of Machine Learning Research*, vol. 6, pp. 1043-1071, 2005.

[19] Ramaswamy, S. et al.: Multiclass cancer diagnosis using tumor gene expression signatures, *PNAS*, vol. 98, no. 26, pp. 15149-15154, 2001.

[20] Hoshida, Y. et al.: Subclass Mapping: identifying common subtypes in independent disease data sets, *PLOS ONE*, vol. 11, pag. 1-8.

[21] Statnikov, A.:A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, *Bioinformatics*, vol. 21,no. 5,pp. 631-643.