Kernel based support vector machine via semidefinite programming: Application to medical diagnosis

Domenico Conforti, Rosita Guido

DEIS, University of Calabria, Ponte Pietro Bucci, Arcavacata di Rende (CS), Italy

ARTICLE INFO

Available online 5 March 2009

Keywords:
Machine learning
Classification
Support vector machines
Kernel function
Semidefinite programming
Medical diagnosis

ABSTRACT

Support vector machine (SVM) is a well sound learning method and a robust classification procedure. Choosing a suitable kernel function in SVM is crucial for obtaining good performance; the difficulty is how to choose a suitable data transformation for the given problem. To this end, multiple kernel matrices, each of them corresponding to a given similarity measure, can be linearly combined. In this paper, the optimal kernel matrix, obtained as linear combination of known kernel matrices, is generated using a semidefinite programming approach. A suitable model formulation assures that the obtained kernel matrix is positive semidefinite and is optimal with respect to the dataset under consideration. The proposed approach has been applied to some very important medical diagnostic decision making problems and the results obtained by carrying out preliminary numerical experiments demonstrated the effectiveness of the proposed solution approach.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

In this paper we propose an optimization model based approach for learning the best performance kernel function to be embedded into support vector machine (SVM) classifier. In particular, the optimal kernel function is generated by formulating and solving a semidefinite programming (SDP) model. The proposed approach is then effectively applied to the solution of some real-life problems within medical diagnostic decision making domains.

SVM [1–4], developed by Cortes and Vapnik [5] as a method for solving binary classification problem, is currently a hot topic in the machine learning community. Several applications of the SVM improve the results obtained with other methods (such as neural networks), making it rather important. SVM has nice geometrical interpretation: indeed, it finds an optimal separating hyperplane (OSH), that is the hyperplane with larger margin; in this way SVM avoids overfitting and the generalization capability is typically quite good. SVM performance can be further improved by appropriately exploiting the kernel function framework. In fact, kernel based methods derive their high performance from the ability to incorporate prior knowledge via a suitable kernel function, which transforms data points embedding them into a higher dimensional space. This transformation yields a symmetric, positive semidefinite (psd) matrix K, known as kernel matrix, which can be regarded as a matrix of generalized similarity measures among the data points; psd symmetric kernel matrices become thus the performing “engine” of SVM, as well as of other several kernel based algorithms.

In order to get good performance, it is crucial to select the most suitable kernel function among the several possibilities. Indeed, different kernel functions correspond to different embedding of the data and, obviously, it is of paramount importance to identify the “best” embedding of data, hence the “optimal” kernel function, with respect to the best accurate solution of the classification problem. In the literature, the kernel function “optimization” is usually carried out by using a grid-search approach of parameters and the relative performance is valued by cross-validation (CV). In this paper, on the contrary, the SVM model selection is conducted by an automatic “selection” or “construction” of the best kernel function (via SDP), which maximizes the margin of the OSH. Taking into account that the kernel formalism allows combining several kernel matrices, we exploit the idea that the kernel matrix is obtained by the selected kernel function. Thus, we modify and extend the approach proposed by Lancriet et al. [6], with the aim to obtain the optimal kernel function corresponding to a particular feature space in which the margin is maximized [7].

We use such an approach for the inductive SVM model and, for testing purposes, we performed a set of numerical experiments on some remarkable medical diagnostic decision making domains, by specifically considering publicly available patient datasets.
this respect, it is worthwhile to remark that medical decision making is a quite complex process encompassing patient early diagnosis and clinical management. In particular, medical diagnosis procedure is basically a decision making process, along which the physician induces the diagnosis of the relevant case by appropriately acquiring and interpreting a suitable set of signs, symptoms and clinical data. It is well known that medical diagnosis can be made more accurate and objective with the efficient support of advanced data analysis methodologies. These approaches are particularly helpful when medical diagnosis is mainly related to complex clinical problems, typically involving the integrated evaluation of large amounts of heterogeneous data and information.

Among the innovative methodologies that have been proposed to further enhance the development of high performance medical decision making solvers, there is the kernel based SVM classifier. In particular, we have taken into account the early diagnosis of breast cancers, the diagnosis of heart diseases, the evaluation of thyroid conditions, and the diagnosis of cancer by using expression genes. The relevant medical domains (respectively, cancer cytopathology, cardiovascular diseases, endocrinology and cancer stage assessment) are quite important in terms of prevalence and incidence of diseases, and are mainly characterized by complex clinical procedures with the integration and assessment of various and heterogeneous clinical data.

The remainder of this paper is organized as follows. In Section 2, we briefly present a general overview of SVM and SDP, by paying particular attention to its application in SVM model (for more details, we remind to [6]). In Section 3, we report our contribution, by describing the proposed optimization formulation for learning optimal kernel function in SVM using SDP approach. Section 4 reports the application of the proposed methodology to the solution of medical diagnostic decision making problems. In particular, a description of the relevant datasets, the experimental results, and the discussion are presented. Finally, relevant conclusions are given in Section 5.

2. Kernel based SVM and SDP background

To deeply understand the proposed mathematical model, a familiarity with SVM and SDP is required. In this section we give a brief description about kernel based SVM and SDP; for more details we refer to related references.

Let \( X = (x_i,y_i), \quad i = 1, \ldots, m \) be a dataset of \( m \) samples of instance–label pairs, where \( x_i \in \mathbb{R}^n \) represents the input vector and \( y_i \) denotes the corresponding class label. Let \( \mathcal{X} \subseteq \mathbb{R}^n \) be the input space, \( \mathcal{Y} \) the output space and \( h : \mathcal{X} \rightarrow \mathcal{Y} \) the classifier. Roughly speaking, the aim of several classification methods is to predict the unknown label \( y_{\text{new}} \) of new pattern \( x_{\text{new}} \), using the prior knowledge about the problem under investigation and the dataset. In binary classification problem the output space is \( \mathcal{Y} = \{+1,-1\} \). The SVM learning algorithm finds an OSH, with parameters \( \omega \) (gradient vector) and \( \gamma \) (scalar bias), by using the dataset, so as to separate correctly the data points. In particular, OSH is the hyperplane that maximizes the margin defined as the distance to the closest points to the OSH. Most of the real-world problems involve nonlinearly separable data for which there does not exist a hyperplane that successfully separates the two classes. A solution to the linear inseparability problem is to map the data into a higher dimensional space \( \mathcal{F} \) which is called feature space as opposed to the input space \( \mathcal{X} \) [2,8–11]. Kernel methods (see, for instance, [9,10]) work by embedding data items into \( \mathcal{F} \) in which looking for linear relation. The mechanism that defines this mapping process is called kernel function, psd symmetric function \( k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \). By appropriate feature space of sufficient dimensionality, any consistent dataset can be made separable. Defined the following reproducing kernel map as \( \Phi : x \rightarrow \Phi(x) \), a function \( k(x,\cdot) \) is associate to each point \( x \in \mathcal{X} \) in the input space.

The OSH is given as solution of the following convex optimization problem, known as 1-norm soft-margin SVM:

\[
\min \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{m} \xi_i
\]

s.t. \( y_i(w^T \Phi(x_i) + \gamma) \geq 1 - \xi_i, \quad i = 1, \ldots, m \)

\( \xi_i \geq 0, \quad i = 1, \ldots, m \)

(1)

where the trade-off parameter \( C \) puts penalty on patterns that are misclassified or are close to the SVM decision boundary, controlling the complexity of the decision function versus the minimization of the training errors \( \xi_i \).

More useful than (1) is its dual formulation (see [2] for instance)

\[
\max \quad W(x)
\]

\[
W(x) = \sum_{i=1}^{m} \xi_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \xi_i \xi_j y_i y_j K_{ij}
\]

s.t. \( \sum_{i=1}^{m} \xi_i y_i = 0 \)

\( 0 \leq \xi_i \leq C, \quad i = 1, \ldots, m \)

(2)

where \( \xi_i (i = 1, \ldots, m) \) is the Lagrangian multiplier corresponding to \( i \)th constraint of the primal formulation (1). Let \( x^* \) be the optimal solution of (2), it is easy to prove [2,3] that the maximal margin separating hyperplane occurs when \( w^* = \sum_{i=1}^{m} \xi_i^* y_i x_i \). The optimal value of the bias \( \gamma \) of the OSH is computed, consequently, as

\[
\gamma^* = \frac{1}{2} \min_{(w,\xi)} \left( \min_{(y_i)} w^T x_i + \max_{(y_i)} w^T x_i \right).
\]

(3)

A remarkable property of this dual representation is that only a subset of the points belonging to training set are associated with a non-zero value of \( x \); these points are called support vectors. Let \( \mathcal{S} \) be the set of the support vectors, the decision function used to predict the class of an unknown pattern \( x_{\text{new}} \) is

\[
f(x_{\text{new}}) = \text{sign} \left( \sum_{i\in\mathcal{S}} \xi_i^* y_i K(x_i,x_{\text{new}}) + \gamma^* \right).
\]

(4)

SVMs and other kernel methods derive their power from their ability to incorporate prior knowledge of the dataset via the kernel function. If the used kernel function is \( k \), the corresponding kernel matrix, denoted with \( K \) (Gram matrix), is a function of the \( m \) training points:

\[
K_{ij} = k(x_i,x_j), \quad i = 1, \ldots, m.
\]

that is, \( K_{ij} = k(x_i,x_j), \quad i = 1, \ldots, m \). Hence, the kernel matrix contains the value of the kernel function for every couple of data points and every function that satisfies Mercer’s theorem is a valid kernel [12]. The performance of the SVM highly depends both on the preprocessing of data and on the choice of kernel function. Different kernel functions correspond to different embedding of the data, that is, different notion of similarity is captured. Consequently, the kernel function to be used (e.g., polynomial kernel, Gaussian kernel, or sigmoid kernel) and the values of its relative parameters are usually selected by CV, even though it is known that this approach is strongly limited. The process of parameter fitting into SVM (trade-off parameter and kernel parameters) is known as model selection. Due to the unknown distribution of samples, the generalization capability (i.e., the ability to correctly predict unseen patterns) is estimated by using leave-one-out or \( k \)-fold CV, for instance. In most cases, the trial and error approach is used to find the best function. In case SVM is not able to find a separating hyperplane in feature
space, this means that the choose kernel function is a poor similarity measure for relevant training data or data contain mislabeled examples. In recent works, to improve the generalization property of SVM, some kernel functions are combined among them [13] or the optimal kernel matrix is learned by using the SDP [6].

We introduce here the basis of SDP in which the SVM can be cast. SDP [14,15] defines a class of convex optimization problems: a linear objective function is minimized over the intersection of cone of psd matrices and affine sets. An important and attractive property of SDP is convexity: the set of psd symmetric real matrices, denoted as \( \mathcal{P}_n = \{ X \in \mathbb{R}^{n \times n} : X \succeq 0 \} \), is convex since any positive combination of semidefinite matrices is semidefinite (the symbol \( \succeq \) denotes a positive semidefinite matrix). \( \mathcal{P}_n \) is called psd cone and it is a proper cone, i.e., it is closed, convex, pointed, and solid. The closedness property of the cone \( \mathcal{P}_n \) (in the following, the dimension \( n \) of matrices is omitted to simplify the notation) means that the boundary is included (the boundary is the set of the singular psd matrices). Another important property is that, let \( \mathcal{P}_+ \) be the dual cone of \( \mathcal{P}_\circ \), the equality \( \mathcal{P}_n = \mathcal{P}_+ \) holds, i.e., \( \mathcal{P}_n \) is a self-dual cone.

The standard form of SDP formulation [14] is given as follows (for simplicity there are no equality constraints):

\[
\begin{align*}
\min_{x \in \mathbb{R}^n} & \quad z(x) \\
\text{s.t.} & \quad F(x) = F_0 + x_1 F_1 + \cdots + x_n F_n \\
& \quad F(x) \succeq 0,
\end{align*}
\]

where \( x \in \mathbb{R}^n \) denotes the decision variables. The inequality constraint \( F(x) \succeq 0 \), called linear matrix inequality (LMI), means that the matrix \( F \) (function of variable \( x \)) is restricted to be contained into the psd cone. The \((n+1)\) symmetric matrices \( F_i = F_i^T \in \mathbb{R}^{n \times n} \) and the coefficient vector \( c \in \mathbb{R}^n \) of the objective function are the data of the problem. An important result is that a local optimal solution is also the global optimal solution, since both the objective function and the constraints are convex.

In the next section, the SVM model is cast into the SDP formulation aiming at learning a suitable kernel function. Indeed, several kernel functions are combined in linear mode with the constraint that the relative kernel matrix is still psd. This can be interpreted as an improvement of the hyperkernel method introduced by Smola [13]. The goal of such an approach is to capture the “optimal” similarity measure, i.e., the “optimal” kernel function corresponding to a specific feature space in which the maximal margin hyperplane will be found. In this way, a valid kernel function suitable for SVM is given. The positive semidefiniteness of kernel functions translates in practice into psd matrices.

3. Learning the optimal kernel function via the SDP

Lanckriet et al. [6] introduced SDP in a transductive setting with the aim to learn the optimal kernel matrix, whose entries are, in this particular case, both for training and for test sets. In particular, in order to obtain a suitable kernel matrix, they fixed the trace value of the variable kernel matrix to an assigned constant, whose “best” value is typically quite difficult to determine.

Our approach is different with respect to [6]: indeed, we work in the inductive setting and getting rid of any constraint on the trace of the variable kernel matrix, we introduce a constraint on the weights of the linear combination of the kernel functions, in order to obtain the optimal performance kernel function. Further, we experimentally demonstrate that the generalization ability of SVM, embedded with the resulting optimal kernel function, is guaranteed also for inductive classification problems.

Let \( L(x, \beta, \lambda) = x^T \epsilon - \frac{1}{2} x^T G(K)x + \beta^T x - \delta^T (x - Ce) + \lambda y^T x \) be the Lagrangian function of the Wolfe dual SVM model (2). \( \beta, \delta, \lambda \in \mathbb{R}^n \) and \( \epsilon, y \in \mathbb{R} \) are Lagrangian multipliers, \( G(K) \) is defined by \( G_{ij} = y_i y_j \phi(x_i, x_j) \) (\( i, j = 1, \ldots, m \)) and \( e \) is ones vector. By the validity of the strong duality

\[
x^* = \max_{x} \min_{\beta, \delta, \lambda} L(x, \beta, \delta, \lambda) = \min_{\beta, \delta, \lambda} \max_{x} L(x, \beta, \delta, \lambda).
\]

(6)

If \( K \preceq 0 \) then \( G(K) \succeq 0 \); consequently, setting \( \partial L/\partial x = 0 \), the optimal value

\[
x^* = G(K)^{-1}(e + \beta - \delta + \lambda y)
\]

of the dual SVM (2) is obtained. Using this result, (6) is rewritten as

\[
\begin{align*}
\min_{\beta, \delta, \lambda} & \quad \max_{x} L(x, \beta, \delta, \lambda) \\
= & \quad \min_{\beta, \delta, \lambda} \frac{1}{2} (e + \beta - \delta + \lambda y)^T G(K)^{-1} (e + \beta - \delta + \lambda y) + C\delta^T e \\
\text{s.t.} & \quad \beta, \delta, \lambda \geq 0.
\end{align*}
\]

(8)

The constraint \( t \geq \frac{1}{2} (e + \beta - \delta + \lambda y)^T G(K)^{-1} (e + \beta - \delta + \lambda y) + C\delta^T e \) can be rewritten as \( t = 2t \geq (e + \beta - \delta + \lambda y)^T G(K)^{-1} (e + \beta - \delta + \lambda y) + 2C\delta^T e \) which is equivalent to the following LMI constraint (by using Schur’s complement lemma):

\[
\begin{bmatrix}
G(K) \\
(e + \beta - \delta + \lambda y)^T \\
\frac{1}{2} - 2C\delta^T e
\end{bmatrix} \succeq 0.
\]

Let \( \mathcal{K} \) be the cone of psd kernel matrices, our goal is to find the best kernel function which optimizes the function \( z(K) \). The dual SVM soft-margin formulation can be cast then as the following SDP mathematical formulation:

\[
\begin{align*}
\min_{K} & \quad \tilde{t} \\
\text{s.t.} & \quad K \in \mathcal{K} \\
& \quad \frac{G(K)}{\tilde{t} - 2C\delta^T e} \succeq 0, \\
& \quad \beta \succeq 0, \\
& \quad \delta \succeq 0.
\end{align*}
\]

(9)

where \( K \in \mathcal{K} \) is equivalent to the constraint \( K \succeq 0 \). In this way, the optimal kernel matrix \( K_\ast \), belonging to the relative interior cone \( \mathcal{K}_\ast \), is found. Obviously, the non-negative constraints \( \beta \succeq 0, \delta \succeq 0 \) can be expressed as LMI.

In order to formalize the proposed mathematical formulation, we denote with \( \text{KerSet} = \{ K_j, j = 1, \ldots, n_{\text{ker}} \} \) a finite set of \( n_{\text{ker}} \) kernel functions. It is important to remark that \( \mathcal{K} \) has been defined in [6] as linear span of positive semidefinite kernel matrices and bounded trace of the kernel matrix \( K_{\text{tr}, \text{Test}} \) that combines training and test data, that is,

\[
\mathcal{K} = \left\{ K = \sum_{i=1}^{n_{\text{ker}}} \mu_i K_i : \mu_i \geq 0, \mu_i \in \mathbb{R}^{n_{\text{ker}}}, \text{tr}(K_{\text{tr}, \text{Test}}) = \rho \right\},
\]

(11)

where the constant \( \rho \) is user defined. In our proposal, we remove the constraint on the trace of the kernel matrix and introduce an upper
bound on the sum of the weights $\mu_i$. Consequently, $\mathcal{K}$ in our model is defined as

$$\mathcal{K} = \left\{ K = \sum_{i=1}^{n_{ker}} \mu_i K_i \geq 0, \mu_i \in \mathbb{R}^{n_{ker}}, \sum_{i=1}^{n_{ker}} \mu_i \leq r \right\}. \quad (12)$$

The following proposed model SVM-SDP (13) has the aim of finding an optimal kernel function $K^*$ with the relative kernel matrix $K^* = \sum_{i=1}^{n_{ker}} \mu_i^* K_i$, computed as linear combination of $n_{ker}$ known kernel matrices:

$$\begin{align*}
\min_{\bar{K}, \beta, \delta, \mu} & \quad \bar{K} \\
\text{s.t.} & \quad G(\bar{K}) (e + \beta - \delta + 2\delta y) \geq 0, \\
& \quad \bar{K} \in \mathcal{K}, \\
& \quad \beta \geq 0, \\
& \quad \delta \geq 0.
\end{align*} \quad (13)$$

$C$ is the trade-off parameter of model SVM and the parameter $r$ is an upper bound of the linear combination of the weights $\mu_i, i=1, \ldots, n_{ker}$; the values of both parameters are user defined. $\bar{K}$ must be a valid kernel matrix (i.e., psd), thus explicitly the following constraint is required:

$$\bar{K}(\mu) = \mu_1 K_1 + \cdots + \mu_{n_{ker}} K_{n_{ker}} \geq 0. \quad (14)$$

$K(\mu)$ is a block diagonal matrix, where the $i$th block is the given kernel matrix $K_i$.

The optimal solution of (13) is denoted as $(\beta^*, \delta^*, \gamma^*, \mu^*)$. By (7), the decision function (4) is then rewritten as

$$f(\mathbf{x}_{\text{new}}) = \text{sign} \left( \sum_{i=1}^{n_{ker}} \gamma_i^* y_i K^*(\mathbf{x}_i, \mathbf{x}_{\text{new}}) + \gamma^* \right). \quad (15)$$

where $K^*(\mathbf{x}_i, \mathbf{x}_{\text{new}}) = \sum_{j=1}^{n_{ker}} K_i(\mathbf{x}_i, \mathbf{x}_{\text{new}})$ and $\gamma^*$ is computed by (2).

The values $\mu_i (i=1, \ldots, n_{ker})$ weight the kernel functions and the support kernel functions corresponding to $\mu_i \neq 0$ are identified likewise the support vectors. Such an approach assigns a relative weight to each used kernel function and thereby determines a measure to select the most suitable kernel. If more than one weight $\mu$ is non-zero, the kernel matrix $K^*$ is associated to a particular kernel function $k^*$ that we name hybrid kernel function. This approach thus can be viewed as the search of the kernel $K^*$ with reproducing kernel Hilbert space (RKHS$_{k^*}$) that have good generalization properties; that is, we search the best kernel function (hybrid or not) rather than the optimal kernel matrix, as done in [6]. Thus, we optimize the kernel in the hyperkernel and then the regularization on the associated RKHS (defined in [13, 16] as hyper-RKHS). The constraint $K \in \mathcal{K}$ implies that a kernel function $k$ that satisfies Mercer’s theorem exists.

To avoid overfitting, in (13) we restrict the search space $\mathcal{K}$ by considering the following three different classes that are convex subsets of the convex cone $\mathcal{K}$. These classes are the affine, convex, and conic combinations, defined as follows (Fig. 1):

$$\mathcal{K}_{\text{affine}} = \left\{ K = \sum_{i=1}^{n_{ker}} \mu_i K_i \geq 0, \sum_{i=1}^{n_{ker}} \mu_i = 1, \mu_i \geq 0, \forall i = 1, \ldots, n_{ker} \right\}. \quad (16)$$

$$\mathcal{K}_{\text{convex}} = \left\{ K = \sum_{i=1}^{n_{ker}} \mu_i K_i \geq 0, \sum_{i=1}^{n_{ker}} \mu_i = 1, \mu_i \geq 0, \forall i = 1, \ldots, n_{ker} \right\}. \quad (17)$$

$$\mathcal{K}_{\text{conic}} = \left\{ K = \sum_{i=1}^{n_{ker}} \mu_i K_i, \mu_i \geq 0, \forall i = 1, \ldots, n_{ker} \right\}. \quad (18)$$

We remark that the weights $\mu_i (i=1, \ldots, n_{ker})$ are not constrained to be non-negative when $K \in \mathcal{K}_{\text{affine}}$; consequently, whenever $\mu_i^* < 0$ the corresponding kernel matrix $K_i$ is subtrated in linear combination. Moreover, the kernel learning problem (13) can be formulated as a quadratic constraint quadratic programming (QCQP) [15] when the convex and conic combinations are used; indeed, for the closure property, $K$ is still a valid matrix since the constraint $\mu_i \geq 0, \forall i = 1, \ldots, n_{ker}$ implies a valid Gram matrix $K = \sum_{i=1}^{n_{ker}} K_i \geq 0$. The trace of the optimal kernel matrix $K^*$ is equal to the traces of all the kernel matrices linearly combined when the first or the second class is used. Indeed, the constraint $\sum_{i=1}^{n_{ker}} \mu_i = 1$ implies that $Tr(K) = Tr(K_i), i = 1, \ldots, n_{ker}$ (for the linearity property of trace). This means that all matrices belonging to $\mathcal{K}$ have the same spectrum and $K^*$ is the matrix that optimizes the margin of the OSH in $\mathcal{F}$.

To better understand the meaning of the proposed approach and the experimental results, which are reported in the next section, we consider the following example. Let us suppose that two kernel functions $k_1$ and $k_2$ are linearly combined. The resulting kernel matrix $K^*$ is computed by using the optimum value $\mu^*$ obtained solving the SVM-SDP model, i.e., $K^* = \mu_1^* K_1 + \mu_2^* K_2$. The values $\mu_1^*$ and $\mu_2^*$ weight the used kernel functions. Indeed, let $\mu_1^* = 0$ and $\mu_2^* = 1$ be the optimal solution; this means that the margin of the computed OSH by using the kernel function $k_2$ is better than the margin obtained if the kernel function $k_1$ is used. This result also means that every kernel hybrid function is, in this case, worst than the kernel function $k_2$.

### 4. Application to medical diagnosis

In order to experimentally demonstrate the validity of the proposed approach, we applied our model to several real-world medical diagnosis datasets. The detailed information for each dataset is given in Table 1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of instances</th>
<th>Number of attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBDC</td>
<td>682</td>
<td>9</td>
</tr>
<tr>
<td>Heart</td>
<td>270</td>
<td>13</td>
</tr>
<tr>
<td>Thyroid</td>
<td>215</td>
<td>5</td>
</tr>
<tr>
<td>Leukemia</td>
<td>72</td>
<td>7129</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>62</td>
<td>2000</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>216</td>
<td>4000</td>
</tr>
</tbody>
</table>

Three of them are from the UCI Machine Learning Repository [17]: Wisconsin Breast Cancer Database (WBDC), Heart Disease, and Thyroid. WBDC collects data from 682 patients whose cancer status was known; each pattern is characterized by nine attributes obtained via needle aspiration of a tumor. The aim is to classify the breast cancer label (i.e., to diagnose a patient with either benign or malignant cancer) and the prediction accuracy will be measured by ten 3-fold CV. The classification task of the dataset Heart Disease (cleaned version of the UCI Cleveland Heart Disease) refers to the presence of heart disease; the class distribution is the absence of heart disease (150 instances) and the presence of heart disease (120 instances).
The Thyroid dataset is available in 100 partitions into training set and test set (65% and 35%, respectively).

We considered also three microarray gene expression datasets, which are useful to diagnose disease: Ovarian Cancer [18], Leukemia data (from Golub et al. [19]), which involves expressions on samples taken from 72 patients with either acute myeloid leukemia or acute lymphoblastic leukemia; the dataset contains expression levels for 7129 human genes. Finally, the Colon Tumor data (from Alon et al. [20]), where gene expression patterns were measured for 2000 genes in the 62 tissue samples classified as 40 tumor and 22 normal tissues. It is well known that the low number of samples characterizes this type of datasets; for this reason, we combined the original training and test set and conducted a 10-fold CV to evaluate the mean accuracy.

### 4.1. Experimental setting

Our computational approach is organized on the basis of four stages:

1. definition of the KerSet and computation of the relevant kernel matrices;
2. formulation and solution of the SVM-SDP model (13);
3. construction of the decision function (15) by using the obtained optimal solution \((\mathbf{x}^*, \mu^*)\);
4. testing phase and performance evaluation.

The four stages are analyzed in detail in what follows. It is important to notice that the obtained optimal Gram matrix is the matrix of general similarity measures among the data points.

**Stage 1:** Two different kernel functions are used in our experiments:

- polynomial kernel function \(k_{\text{poly}}(x_i, x_j) = ((x_i \cdot x_j) + c)^d\) and
- Gaussian kernel function \(k_{\text{Gauss}}(x_i, x_j) = e^{-\frac{\| x_i - x_j \|^2}{2\sigma^2}}\),

where \(c, d\) and \(\sigma\) are kernel parameters. Only the dataset Heart Disease and Leukemia have been normalized. Moreover, the kernel matrix corresponding to the polynomial kernel function has been normalized as follows:

\[
\tilde{K}(x) = \frac{K(x, z)}{\sqrt{K(x, x)K(z, z)}}
\]

before to combine it linearly. This transformation is needed because the Gaussian kernel is a normalized kernel; its meaning is that all the elements on the principal diagonal \(K(x_i, x_i) = 1 (i = 1, \ldots, m)\), that is, the vectors lie on a unit hyper-sphere into the feature space \(\mathcal{F}\). The polynomial kernel parameter is \(c = 1\) by default.

Among the rather efficient algorithms for solving SDP (as for instance primal–dual interior point methods, dual interior point methods or augmented Lagrangian methods [21–23]), we applied the interior point method for solving the proposed mathematical formulation (13). Indeed, we used SeDuMi free package [24], a general purpose SDP solver, where the primal–dual interior point method is implemented, based on the Nesterov and Todd search direction with a predictor–corrector scheme. The optimal solution is \((\mathbf{x}^*, \mu^*)\).

**Stage 2:** The solution \((\mathbf{x}^*, \mu^*)\) of (13) is a function of both the defined set KerSet and the data points; this explains why the several values of kernel parameters are used in the different datasets (as listed in Table 2). The optimal value of each weight \(\mu_i\) is indicative of the relevance of the corresponding kernel function \(k_i\) for the final classifier within the linear combination (14). The selection or the “construction” of the kernel function is automatically determined by solving the proposed formulation SVM-SDP (13).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dataset</th>
<th>Polynomial kernel</th>
<th>Gaussian kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(C)</td>
<td>(n_{ker})</td>
<td>(d)</td>
</tr>
<tr>
<td>WBCD</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Heart disease</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Thyroid</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Leukemia</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

**Stage 3:** Finally, in order to estimate the performance of the optimal decision function the testing step is performed.

We trained SVM classifier using both the convex \((\mathcal{K}_{\text{convex}})\) and affine combination \((\mathcal{K}_{\text{affine}})\) of kernels. We performed two sets of experiments which differ both for the number of kernel functions linearly combined and for the values of the kernel parameters. In particular, for each dataset a different KerSet has been defined. The assigned values to kernel parameters and the value of the regularization parameter \(C\) are reported in Table 2. We observe that some values of the Gaussian kernel parameter \(\sigma\) correspond to the bad kernel matrix I; thus, these values have been discarded. The obtained experimental results are reported in Table 3 in terms of mean accuracy and standard deviation on the relative test set. In particular, over ten 3-fold CV for WBCD, three 10-fold CV for Heart Disease, over 100 runs for Thyroid dataset, and 10-fold CV for the remaining datasets. All experiments were performed using own code combined with SeDuMi in MATLAB [25], on 2.4 GHz Pentium IV machine with 1 GB memory and running Windows Vista.

We remark that we also considered the adaptation of SVM algorithm in a normalized feature space as proposed in [26], but the suggested correction of the bias \(\gamma\) did not improve the obtained accuracy.

Model performance depends on several factors, besides the learning algorithm, as the class distribution, cost of misclassification, or the size of training and test sets. To evaluate the obtained performance, we have compared our results in terms of predictive accuracy on testing samples with other performances reported on the same datasets in the literature. It is important to notice that the reported results by Lanckriet et al. [6], on the WBCD and Heart Disease datasets, are relative to the transduction setting (which consists in completing the labeling of a partially labeled dataset) that they addressed in their work. To this end, they learned a good embedding of the entire dataset, reporting 95.5% for WBCD and 84.8% for the Heart Disease dataset as best accuracy. To the best of our knowledge the best classifier on Thyroid dataset can be found in [27], which predicts the results with 95.80% ± 2.07 mean accuracy using a kernel Fisher discriminant, whereas they obtained the 95.20% ± 2.19 mean accuracy by using a SVM with RBF kernel. Further, it is worthwhile to remark that the reported results on Leukemia, Colon and Ovarian Cancer datasets listed in Table 3 are obtained by considering all the features. Indeed, the best reported results in the literature in terms of mean accuracy (100%) are relative to a subset of selected genes (see, for instance, [28]).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>(X_{\text{convex}})</th>
<th>(X_{\text{affine}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBCD</td>
<td>96.79 ± 0.20</td>
<td>96.58 ± 0.03</td>
</tr>
<tr>
<td>Heart disease</td>
<td>82.59 ± 7.81</td>
<td>82.59 ± 7.81</td>
</tr>
<tr>
<td>Thyroid</td>
<td>95.83 ± 2.23</td>
<td>96.77 ± 2.03</td>
</tr>
<tr>
<td>Leukemia</td>
<td>98.50 ± 4.51</td>
<td>95.60 ± 7.00</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>90.71 ± 14.77</td>
<td>85.00 ± 12.29</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>97.68 ± 2.44</td>
<td>95.80 ± 4.58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>(X_{\text{convex}})</th>
<th>(X_{\text{affine}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBCD</td>
<td>96.79 ± 0.20</td>
<td>96.58 ± 0.03</td>
</tr>
<tr>
<td>Heart disease</td>
<td>82.59 ± 7.81</td>
<td>82.59 ± 7.81</td>
</tr>
<tr>
<td>Thyroid</td>
<td>95.83 ± 2.23</td>
<td>96.77 ± 2.03</td>
</tr>
<tr>
<td>Leukemia</td>
<td>98.50 ± 4.51</td>
<td>95.60 ± 7.00</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>90.71 ± 14.77</td>
<td>85.00 ± 12.29</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>97.68 ± 2.44</td>
<td>95.80 ± 4.58</td>
</tr>
</tbody>
</table>
5. Conclusion

From the overall results obtained by carrying out the above-mentioned preliminary experiments, we claim that the proposed approach could be quite promising for effectively solving classification problems. Comparing the obtained results by combining kernel functions with respect to those obtained when only one kernel function is used, it is evident an improvement of the performance of final classifier on the same test set. We observe that our model can also be used to select the best value of kernel function parameter and no special effort was made to select the best kernel function. That could result in further improvement of data analysis performance within medical diagnostic decision making problems. Obviously, further experiments should be conducted with different learning algorithms and paradigms to allow performance comparisons with our obtained results. One of our future goals is to minimize the number of SV using SVM-SDP. However, one of the inherent difficulty of SDP/SVM is its computational inefficiency. In theory, the method has time complexity $O(m^3)$, where $m$ is the number of data points or equivalently the dimension of kernel matrix. Thus, in case of very large-scale datasets, learning could not be completed in reasonable time.

References