Selecting the Hypothesis Space for Improving the Generalization Ability of Support Vector Machines

Davide Anguita, Member, IEEE, Alessandro Ghio, Member, IEEE, Luca Oneto, Student Member, IEEE, Sandro Ridella, Member, IEEE

Abstract—The Structural Risk Minimization framework has been recently proposed as a practical method for model selection in Support Vector Machines (SVMs). The main idea is to effectively measure the complexity of the hypothesis space, as defined by the set of possible classifiers, and to use this quantity as a penalty term for guiding the model selection process. Unfortunately, the conventional SVM formulation defines a hypothesis space centered at the origin, which can cause undesired effects on the selection of the optimal classifier. We propose here a more flexible SVM formulation, which addresses this drawback, and describe a practical method for selecting more effective hypothesis spaces, leading to the improvement of the generalization ability of the final classifier.

I. INTRODUCTION

The Support Vector Machine (SVM) algorithm [1], whose solid theoretical foundations date back to the development of the Statistical Learning Theory [2], represents one of the state-of-the-art techniques for classification problems and is widely used in real-world applications [3]. The linear SVM is a classifier of the following form:

\[ f(x) = w^T x + b \]  

where the weights \( w \) and the bias \( b \) are the parameters of the model. These parameters are computed during the learning phase using a training set \( D_n \in \mathbb{R}^{n \times d} \), consisting of \( n \) \( d \)-dimensional samples. In addition to the set of parameters \((w, b)\), the SVM algorithm requires the tuning of some hyperparameters, which noticeably affect the classifier performance. This tuning is not part of the training and is known as the model selection phase.

The problem of model selection is generally linked to the estimation of the generalization ability of a classifier, i.e. the ability of a model to correctly classify previously unseen data. For this purpose, the Structural Risk Minimization (SRM) principle [2] defines some basic steps for its estimation: (i) define a centroid \( \tilde{w} \); (ii) choose a (possibly infinite) sequence of hypothesis spaces \( F_k \), \( k = 1, 2, \ldots \), where the classes of functions \( F_k \) describe classifiers of growing complexity and are centered on \( \tilde{w} \); (iii) select the optimal model \( f^* \) among the hypothesis spaces by exploiting the following trade-off between overfitting and underfitting:

\[ f^* = \arg \min_{f \in F_k, k \in 1, 2, \ldots} C(f, F_k) \]  

where

\[ C(f, F_k) = \left[ \hat{L}_n(f) + \text{pen}(F_k) \right] \]  

\[ \hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \delta(y_i, f(x_i)) \]  

\[ \text{pen}(F_k) = \sum_{i=1}^{k} \alpha_i \]  

The first term \( \hat{L}_n(f) \) is the empirical risk or, in other words, the misclassification rate of \( f \) on \( D_n \). It is well-known that the minimization of the empirical risk alone can lead to overfitting, if the hypothesis space is too large, but this tendency can be counterbalanced by penalty term \( \text{pen}(F_k) \), which penalizes complex models (Fig. 1).

In the conventional SVM formulation, as it will be clearer in the next sections, the hypothesis space is usually (and arbitrarily) centered in the origin \( (\tilde{w}_0 = 0) \), because, in general, there is no a-priori information leading to a better choice. This choice, however, severely influences the sequence \( F_k \) and its detrimental effect on the model selection can be understood through the example of Fig. 2, where we suppose to know the optimal classifier \( f^* \). In this case, in fact, the hypothesis space \( F_{k^*} \), which includes \( f^* \), is characterized by a large \( \text{pen}(F_{k^*}) \) term and, thus, is penalized with respect to other models (i.e. \( f^o \neq f^* \)). This wrong choice of the optimal classifier is due to the fact that \( f^* \) is “far” from the a-prioristically (and indiscriminately) chosen centroid \( \tilde{w}_0 \) for the sequence \( F_k \). On the contrary, if we were able to define a sequence of hypothesis spaces \( G_k \), \( k = 1, 2, \ldots \), centered on a point \( \tilde{w} \) sufficiently close to \( f^* \), the penalty term in \( C(f^*, G_k) \) would be noticeably reduced and we would be able to improve the SVM model selection by choosing a model close to the optimal one (i.e. \( f^o \approx f^* \) or, in the best case, \( f^o = f^* \)).
The objective of this work is to present a technique that allows us to select a "good" centroid \( \hat{w} \) and to center a sequence of hypothesis spaces around it, so to better explore the classes of functions for model selection purposes. In order to estimate the complexity of the hypothesis space, captured by the penalty term, we make use of the Rademacher Complexity (RC) [4] and the Maximal Discrepancy (MD) [5] techniques, which are two statistically rigorous tools that can be effectively used in practice [6]. These methods allow us to perform an effective model selection, especially when applied to small sample problems, where the number of patterns is small compared to the dimensionality of the problem [7].

The paper is organized as follows: in Section II, we briefly revise the RC and MD methods for estimating the penalty term for problem (2). In Section III the Support Vector Machine algorithm is summarized and in Section IV we detail the main results of the paper: how the sequence of hypothesis spaces can be centered on a generic centroid and how the latter can be effectively chosen. Finally, Section V presents some experimental results supporting our approach.

II. THE MAXIMAL DISCREPANCY AND THE RADEMACHER COMPLEXITY OF A CLASSIFIER

Let us consider the training set \( D_n \) and a hypothesis space \( \mathcal{F} \). Let us split \( D_n \) in two halves and compute the two empirical errors:

\[
\hat{L}_{n/2}^{(1)}(f) = \frac{2}{n} \sum_{i=1}^{\frac{n}{2}} \ell(f(x_i), y_i) \tag{4}
\]

\[
\hat{L}_{n/2}^{(2)}(f) = \frac{2}{n} \sum_{i=\frac{n}{2}+1}^{n} \ell(f(x_i), y_i) \tag{5}
\]

where \( \ell(f(x_i), y_i) \) is a suitable bounded loss function. The Maximal Discrepancy (MD) of \( \mathcal{F} \) is defined as

\[
\hat{\mathcal{M}}_n(\mathcal{F}) = \max_{f \in \mathcal{F}} \left( \hat{L}_{n/2}^{(1)}(f) - \hat{L}_{n/2}^{(2)}(f) \right) \tag{6}
\]

Since, in practical cases, we want to avoid that a possible unlucky shuffling results in an unreliable MD value, we replicate \( m \) times the splitting procedure: at each iteration, the dataset \( D_n \) is randomly shuffled and, then, \( m \) MD values are averaged. Then, the penalty term for the MD procedure can be computed as:

\[
\text{pen}_{MD}(\mathcal{F}) = \frac{1}{m} \sum_{i=1}^{m} \hat{\mathcal{M}}_n^{(i)}(\mathcal{F}) + 3 \sqrt{\frac{\log \left( \frac{4}{\delta} \right)}{2n}}, \tag{7}
\]

where \( \delta \) is a user–defined confidence value. It is worth noting that the penalization term characterizes the class of functions \( \mathcal{F} \), thus is valid for every classifier \( f \) within this hypothesis space. If the loss function is such that \( \forall f \in \mathcal{F} \)

\[
\ell(f(x_i), y_i) = 1 - \ell(f(x_i), y_i), \tag{8}
\]

then the MD term can be computed through a conventional minimization procedure:

\[
\hat{\mathcal{M}}_n^{(i)}(\mathcal{F}) = 1 - 2 \inf_{f \in \mathcal{F}} \hat{L}_n^{(i)}(f) \tag{9}
\]

where \( \hat{L}_n^{(i)}(f) \) is the error on the training set where half of the labels are flipped.

The Rademacher Complexity (RC) of \( \mathcal{F} \) is defined as:

\[
\hat{\mathcal{R}}_n(\mathcal{F}) = E_{\sigma} \sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell(f(x_i), y_i) \tag{10}
\]

where \( \sigma_1, \ldots, \sigma_n \) are \( n \) independent random variables for which \( P(\sigma_i = +1) = P(\sigma_i = -1) = 1/2 \). The penalty term of Eq. (2), according to the RC approach, becomes:

\[
\text{pen}_{RC}(\mathcal{F}) = \hat{\mathcal{R}}_n(\mathcal{F}) + 3 \sqrt{\frac{\log \left( \frac{4}{\delta} \right)}{2n}}. \tag{11}
\]

Analogously to MD, if the condition of Eq. (8) holds for the considered loss function, we have that:

\[
\hat{\mathcal{R}}_n(\mathcal{F}) = 1 - 2E_{\sigma} \inf_{f \in \mathcal{F}} \hat{L}_n^\sigma(f) \tag{12}
\]

where \( \hat{L}_n^\sigma(f) \) is the error performed on \( D_n^\sigma = \{(x_1, \sigma_1), \ldots, (x_n, \sigma_n)\} \) (i.e. the original dataset, but with random labels).

III. THE SUPPORT VECTOR MACHINE

Throughout this work, we will consider the linear SVM classifier, as in Eq. (1), because we will focus on small sample problems, where the previously summarized MD and RC techniques are particularly effective [7]. The extension to the non–linear case, through the use of kernels, will be detailed in the Appendix.

Given a training set \( D_n \), the weights \( w \in \mathbb{R}^d \) and the bias \( b \in \mathbb{R} \) of a classifier \( f \in \mathcal{F} \), where \( \mathcal{F} \) is centered on \( w_0 = 0 \), are found in the learning phase of SVM by solving the following primal Convex Constrained Quadratic Programming (CCQP) problem based on Tikhonov regularization (T-SVM) [2]:

\[
\min_{w, b, \xi} \quad \frac{1}{2} \|w\|^2 + C e^T \xi \tag{13}
\]

\[
y_i (w^T x_i + b) \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n]
\]

\[
\xi_i \geq 0 \quad \forall i \in [1, \ldots, n]
\]
In order to better control the size of the hypothesis space, an hypothesis space but does not allow to directly determine the size of the hypothesis space of increasing size and centered around an arbitrarily chosen centroid. The conventional model selection procedures use a sequence of hypothesis spaces of increasing size and centered around an arbitrarily chosen centroid $\hat{w}_0 = 0$.

Let us suppose to know that the classifier

$$\hat{f}(x) = \hat{w}^T x + \hat{b}$$

is sufficiently close to the optimal solution $f^*$. Then, we can use $\hat{w}$ as the centroid for the sequence of hypothesis spaces and the I-SVM problem (17) can be reformulated as

$$\min_{w,b,\xi} \quad \frac{1}{2} \|w\|^2 \leq \rho^2 \quad \text{subject to} \quad y_i (w^T \bar{x}_i + b) \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n] \quad \xi_i \geq 0 \quad \forall i \in [1, \ldots, n]$$

where $\rho$ is a hyperparameter. As the two formulations (17) and (13) are equivalent for some value of $C$ [8], the I-SVM problem can be solved through the usual T-SVM problem, for which efficient solvers can be retrieved in literature [9] like the well-known Sequential Minimal Optimization (SMO) algorithm [10], [11], when the space of functions is centered on $\hat{w}_0$. The tool for model selection using the MD and RC approaches is completed by the theoretically rigorous procedure, presented in [6], [7].

IV. IMPROVING THE SVM MODEL SELECTION

A. Using a generic centroid for the hypothesis space

The conventional model selection procedures use a sequence of hypothesis spaces of increasing size and centered around an arbitrarily chosen centroid $\hat{w}_0 = 0$.

Let us suppose to know that the classifier

$$f(x) = w^T x + b$$

is sufficiently close to the optimal solution $f^*$. Then, we can use $\hat{w}$ as the centroid for the sequence of hypothesis spaces and the I-SVM problem (17) can be reformulated as

$$\min_{w,b,\xi} \quad e^T \xi$$

subject to

$$\|w - \hat{w}\|^2 \leq \rho^2$$

$$y_i (w^T x_i + b) \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n]$$

$$\xi_i \geq 0 \quad \forall i \in [1, \ldots, n]$$

where the hypothesis space is defined as the set of functions for which $\|w - \hat{w}\|^2 \leq \rho^2$ and $b \in \mathbb{R}$. The previous I-SVM can be solved through the following T-SVM problem, as they are equivalent for some value of $C$:

$$\min_{w,b,\xi} \quad \frac{1}{2} \|\hat{w}\|^2 + C e^T \xi$$

subject to

$$y_i (\hat{w}^T x_i + \hat{b}) \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n]$$

$$\xi_i \geq 0 \quad \forall i \in [1, \ldots, n]$$

Problem (20) is equivalent to the conventional primal formulation of SVM, thus the results of [8] can be exploited. In fact, noting that the centroid $\hat{w}$ is fixed, it is sufficient to set $\hat{w} = [(w - \hat{w})/1]$ and $\bar{x}_i = [x_i |w^T x_i]$ to obtain:

$$\min_{w,b,\xi} \quad \frac{1}{2} \|\hat{w}\|^2 + C e^T \xi$$

subject to

$$y_i (\hat{w}^T \bar{x}_i + \hat{b}) \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n]$$

$$\xi_i \geq 0 \quad \forall i \in [1, \ldots, n]$$

The dual formulation of problem (19) can be computed as well. The Lagrangian is:

$$L = \frac{1}{2} \|w - \hat{w}\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \beta_i [y_i (w^T x_i + b) - 1 - \xi_i] - \sum_{i=1}^n \delta_i \xi_i$$

from which we obtain the following Karush-Kuhn-Tucker (KKT) conditions:

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \hat{w} + \sum_{i=1}^n \beta_i y_i x_i$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^n \beta_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = 0 \rightarrow C - \beta_i - \delta_i = 0 \rightarrow \beta_i \leq C$$

and obtaining:

$$\min_{\beta} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \beta_i \beta_j y_i y_j x_i^T x_j + \sum_{i=1}^n (y_i \hat{w}^T x_i - 1) \beta_i$$

subject to

$$0 \leq \beta_i \leq C \quad \forall i \in [1, \ldots, n]$$

$$y_i^T \beta = 0$$

which can be solved with the efficient methods used for the conventional SVM problem [9].
Let us consider the classifier of Eq. (18): we exploited \( \mathbf{\hat{w}} \) as a centroid, since we supposed that \( \hat{f} \) was a good choice with respect to the true optimal classifier \( f^* \). As a further step, it is possible to include the information concerning the bias \( b \) of \( \hat{f} \), by defining an augmented centroid \( \mathbf{\hat{w}} = [\mathbf{\hat{w}}| b] \). By also setting \( \mathbf{\hat{w}} = [\mathbf{\hat{w}}| b] \) and \( \mathbf{x}_i = [\mathbf{x}_i|1] \), the hypothesis space is defined as the set of classifiers for which \( \| \mathbf{\hat{w}} - \mathbf{w} \|^2 \leq \rho^2 \). Thus, the I-SVM problem can be reformulated as:

\[
\min_{\mathbf{w}, \xi} \left\{ \mathbf{e}^T \xi \right\}
\]

\[
\| \mathbf{\hat{w}} - \mathbf{w} \|^2 \leq \rho^2
\]

\[
y_i \mathbf{\hat{w}}^T \mathbf{x}_i \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n]
\]

\[
\xi_i \geq 0 \quad \forall i \in [1, \ldots, n]
\]

We can solve the T-SVM problem instead, whose primal formulation is:

\[
\min_{\mathbf{w}, \xi} \frac{1}{2} \| \mathbf{\hat{w}} - \mathbf{w} \|^2 + C \mathbf{e}^T \xi
\]

\[
y_i \mathbf{\hat{w}}^T \mathbf{x}_i \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n]
\]

\[
\xi_i \geq 0 \quad \forall i \in [1, \ldots, n]
\]

The dual formulation becomes:

\[
\min_{\mathbf{\beta}} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_i \beta_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_{i=1}^{n} \left[ y_i \mathbf{\hat{w}}^T \mathbf{x}_i - 1 \right] \beta_i
\]

\[
0 \leq \beta_i \leq C \quad \forall i \in [1, \ldots, n]
\]

which can be solved exploiting the algorithm proposed in [12]. The SVM function assumes the following formulation:

\[
f(\mathbf{x}) = \mathbf{\hat{w}}^T \mathbf{x} = \mathbf{w}^T \mathbf{x} + b.
\]

B. Choosing feasible centroids

In the previous section, we supposed to know a function \( \hat{f} = \mathbf{\hat{w}}^T \mathbf{x} + b \), which represents a good starting point for searching the true optimal classifier \( f^* \). In this section, we propose a heuristic approach for finding at least one \( \hat{f} \), so that \( \mathbf{\hat{w}} \), and eventually \( b \), can be used for defining an effective centroid for the sequence of hypothesis spaces.

For this purpose, let us split the dataset \( D_n \) in two parts:

\[
D_{n_t} = \{ (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n_t}, y_{n_t}) \}
\]

\[
D_{n_c} = D_n \setminus D_{n_t}
\]

where \( n_t \) patterns are used to find the function \( \hat{f} \) and the remaining \( n_c = n - n_t \) samples can be safely exploited for estimating the values of the MD and RC penalty terms (7) and (11) (as \( D_{n_t} \cap D_{n_c} = \emptyset \)).

In order to find \( \hat{f} \), let us consider \( D_{n_c} \) and a class of functions \( \mathcal{F}_c \) centered on \( \mathbf{\hat{w}}_0 = 0 \). When varying the hyperparameter \( C \) and solving problem (13), the following three sets:

\[
\mathcal{E} = \{ i : y_i f(\mathbf{x}_i) = 1 \}
\]

\[
\mathcal{L} = \{ i : y_i f(\mathbf{x}_i) < 1 \}
\]

\[
\mathcal{R} = \{ i : y_i f(\mathbf{x}_i) > 1 \}
\]

can assume only a finite number \( n_s \) of configurations [13]. In other words, the number of possible classifiers, found using \( D_{n_c} \) as training set, is computable and equals \( n_s \) (\( \hat{f}_p \), \( p = 1, \ldots, n_s \)).

When working in the SRM framework, we have to guarantee that the sequence of hypothesis spaces \( \mathcal{G}_k \), \( k = 1, 2, \ldots \), is such that

\[
\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \ldots \subseteq \mathcal{G}_k \subseteq \ldots
\]

For this purpose, we must: (i) create \( n_s \) subsets of functions \( \mathcal{G}_k(p) \), centered on \( \mathbf{\hat{w}}_p \), \( p = 1, \ldots, n_s \); (ii) build \( \mathcal{G}_1 = \bigcup_{p=1}^{n_s} \mathcal{G}_k(p) \); (iii) enlarge all the subsets of functions at the same time and, by consequence, the hypothesis space \( \mathcal{G}_k \). Fig. 3 shows an example with three centroid.

![Fig. 3. A sequence of hypothesis spaces, which exploit \( n_s \) centroids.](image)

We have to modify problem (19) so to contemplate possible hypothesis spaces, consisting of \( n_s \) subspaces with different centroids. We can reformulate the problem as:

\[
\min_{\mathbf{w}, b, \xi} \left\{ \mathbf{e}^T \xi \right\}
\]

\[
\forall \mathbf{w}_p \| \mathbf{w} - \mathbf{w}_p \|^2 \leq \frac{\rho^2}{\| \mathbf{w}_p \|^2}
\]

\[
y_i (\mathbf{w}_p^T \mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n]
\]

\[
\xi_i \geq 0 \quad \forall i \in [1, \ldots, n]
\]

which is equivalent to solving \( n_s \) times problem (19) with different centroids and, then, choosing the solution characterized by the minimum value of \( \mathbf{e}^T \xi \). We can also generalize problem (20) by solving \( n_s \) problems like the following one

\[
\min_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \| \mathbf{w} - \mathbf{\hat{w}} \|^2 + C \mathbf{e}^T \xi \right\}
\]

\[
y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n]
\]

\[
\xi_i \geq 0 \quad \forall i \in [1, \ldots, n]
\]

and choosing the best solution out of the \( n_s \) available.
We can also exploit the information concerning the biases of the $\hat{f}_p$ classifiers $b_p$, $p = 1, \ldots , n_s$, for defining augmented centroids. Thus, analogously to Section IV-A, problem (27) is reformulated as:

$$\min_{\vec{w}, \xi} \quad e^T \xi \quad \text{subject to} \quad \sum_{p=1}^{n_s} \|\vec{w} - \hat{w}_p\|^2 \leq \frac{\rho^2}{\|\hat{w}_p\|^2} \\forall \xi \geq 0 \quad \forall i \in [1, \ldots , n]$$

and problem (28) assumes the following formulation for $p = 1, \ldots , n_s$:

$$\min_{\vec{w}, \xi} \quad \frac{1}{2}\|\vec{w} - \hat{w}_p\|^2 + Ce^T \xi \quad \text{subject to} \quad y_i\hat{w}_p^T x_i \geq 1 - \xi_i \quad \forall i \in [1, \ldots , n] \\xi_i \geq 0 \quad \forall i \in [1, \ldots , n].$$

C. The improved SVM model selection

In this section, we exploit the previously obtained results to sketch the method for an improved SVM model selection. In the following, we refer to the SVM formulation based on Ivanov regularization of problems (37) and (39), for which we have to tune the hyperparameter $\rho$ and which can be solved through the Tikhonov based formulations (20) and (28) (or, alternatively, (26) and (29)) [8].

The first step consists in splitting the dataset $D_n$ into two subsets, $D_{n_1}$ and $D_{n_2}$, which will be used, respectively, to find a set of feasible centroids (or augmented centroids) and to properly select the classifier $f^o$ (see Eq. (2)). We do not know any general rule for choosing the cardinality of $D_{n_1}$ and $D_{n_2}$, but we found that, in practice, using 20%, 30% of the $n$ data for $D_{n_1}$, and the remaining patterns for $D_{n_2}$, often represents an effective choice.

Then, we focus our attention on $D_{n_1}$: by applying the method described in [13], we find $n_s$ possible models $\hat{f}_p = \hat{w}_p^T x + b_p$, $p = 1, \ldots , n_s$, which can be used for defining the centroids for the next steps. In practice, as $n_s$ often assumes large values, we can select only a subset of the possible centroids ($n_s \leq n_s$); we found experimentally that between 3 and 5 centroids are sufficient for exploring the space of possible SVM solutions.

In the next steps, we finally perform the SVM model selection. Theoretically speaking, we would have to explore all possible values of $\rho \in (0, +\infty)$; however, in practice, the optimal $\rho$ is found among $n_{GS}$ equally-spaced values in a finite interval $[\rho_{\min}, \rho_{\max}]$ [14]. Let us define $\rho_k$ the $k$-th value of the hyperparameter, then the corresponding $k$-th hypothesis space in the sequence for model selection is

$$\mathcal{G}_k = \bigcup_{p=1}^{n_s} \mathcal{G}_k^{(p)}$$

Algorithm 1 The IMS algorithm.

Input: The dataset $D_n$, the number of patterns for computing the centroid $n_s$, the maximum number of centroids $\hat{n}_s$, the hyperparameter search grid of $n_{GS}$ elements $\{\rho_{\min}, \ldots , \rho_{\max}\}$, a procedure $P_{pen}$ for computing the penalization term (either MD or RC), a procedure $P_{sol}$ for solving problem (37).

Output: The optimal classifier $f^o$.

1. $\{D_{n_1}, D_{n_2}\} = \text{SplitDataset}(D_n, n_s)$
2. $\{\hat{w}_p\} = \text{FindCentroid}(D_{n_1}, n_s)$, $p = 1, \ldots , \hat{n}_s$
3. $C^o = +\infty$
4. For all $\rho_k \in \{\rho_{\min}, \ldots , \rho_{\max}\}$, $k = 1, \ldots , n_{GS}$
5. $\mathcal{G}_k = \text{CreateHypothesisSpace}(\rho_k, \hat{w}_p)$
6. $\text{pen}(\mathcal{G}_k) = P_{pen}(\mathcal{G}_k, D_{n_1}, n_{GS}, \text{problem (37)}, \mathcal{G}_k, D_{n_1}))$
7. $f = P_{sol}(\text{problem (37)}, \mathcal{G}_k, D_{n_1})$
8. $C(f, \mathcal{G}_k) = \text{ComputeCost}(f, D_{n_1}, \text{pen}(\mathcal{G}_k))$
9. If $C(f, \mathcal{G}_k) < C^o$ then
10. $C^o = C(f, \mathcal{G}_k)$
11. $f^o = f$
12. End for
13. Return $f^o$

Algorithm 2 The IMSA algorithm.

Input: The dataset $D_n$, the number of patterns for computing the centroid $n_s$, the maximum number of centroids $\hat{n}_s$, the hyperparameter search grid of $n_{GS}$ elements $\{\rho_{\min}, \ldots , \rho_{\max}\}$, a procedure $P_{pen}$ for computing the penalization term (either MD or RC), a procedure $P_{sol}$ for solving problem (39).

Output: The optimal classifier $f^o$.

1. $\{D_{n_1}, D_{n_2}\} = \text{SplitDataset}(D_n, n_s)$
2. $\{\hat{w}_p\} = \text{FindCentroid}(D_{n_1}, n_s)$, $p = 1, \ldots , \hat{n}_s$
3. $C^o = +\infty$
4. For all $\rho_k \in \{\rho_{\min}, \ldots , \rho_{\max}\}$, $k = 1, \ldots , n_{GS}$
5. $\mathcal{G}_k = \text{CreateHypothesisSpace}(\rho_k, \hat{w}_p)$
6. $\text{pen}(\mathcal{G}_k) = P_{pen}(\mathcal{G}_k, D_{n_1}, n_{GS}, \text{problem (39)}, \mathcal{G}_k, D_{n_1}))$
7. $f = P_{sol}(\text{problem (39)}, \mathcal{G}_k, D_{n_1})$
8. $C(f, \mathcal{G}_k) = \text{ComputeCost}(f, D_{n_1}, \text{pen}(\mathcal{G}_k))$
9. If $C(f, \mathcal{G}_k) < C^o$ then
10. $C^o = C(f, \mathcal{G}_k)$
11. $f^o = f$
12. End for
13. Return $f^o$

where $\mathcal{G}_k^{(p)}$ are the hypothesis subspaces, for which either $\|\vec{w} - \hat{w}_p\|^2 \leq \frac{\rho^2}{\|\hat{w}_p\|^2}$ and $b \in \mathbb{R}$ or, when augmented centroids are used, $\|\vec{w} - \hat{w}_p\|^2 \leq \frac{\rho^2}{\|\hat{w}_p\|^2}$. Then, by solving either problem (37) or (39) using $D_{n_1}$ as training set, we compute the value of the penalty term for the Maximal Discrepancy (Eq. (7)) or the Rademacher Complexity (Eq. (11)) procedures, we train the model $f \in \mathcal{G}_k$ and compute the empirical error $L_{n_1}(f)$ on the $n_1$ data (see Eq. (2)).

The previously described procedures, where $\hat{w}_p$ or the augmented $\hat{w}_p$ are used as centroids, are presented respectively in Algorithms 1 (Improved Model Selection, IMS) and 2 (Improved Model Selection using Augmented centroids, IMSA).

V. EXPERIMENTAL RESULTS

In the following experiments, we compare the conventional model selection procedure with the IMS and IMSA methods.
The experimental setup is the following:

- the data are normalized in the range \([0, 1]\);
- the model selection is performed by searching for the optimal value of \(\rho\), as described in Section IV-C;
- 30\% of the data are used for selecting the centroids;
- the remaining samples are used for model selection purposes through the MD and RC approaches:
  - in order to avoid “unlucky” values when computing the MD penalization term, we set \(m = 30\) in Eq. (7);
  - the expectation in Eq. (10) is approximated through a Monte Carlo procedure where 200 random shufflings of the labels are used;
- the error rates of the optimal models, chosen by the different approaches, are then computed on a separate test set, where available;
- when a separate test set is not available, the approach of [15] is used by generating different training/test pairs for the comparison.

A. The MNIST Dataset

The MD and RC procedures are targeted toward small sample problems, where these techniques are usually more effective than other model selection methods (e.g., k-Fold Cross Validation) [7], [16], [17]. In order to fairly compare the performance of the different model selection techniques, we select a real-world application, the MNIST dataset [18], consisting of a large number of samples, and use only a small amount of the available data as training set. The remaining samples can be used as a large-cardinality test set for comparison purposes.

The MNIST dataset consists of 62000 images, representing the numbers from 0 to 9: in particular, we consider the 13074 patterns containing 0’s and 1’s, allowing us to deal with a binary classification problem. We build the training set by randomly sampling a small number of patterns, varying from \(n = 10\) to \(n = 300\), while the remaining 13074 – \(n\) images are used as a test set. In order to build statistically relevant results we build a set of 30 replicates using a random sampling technique. Note that the dimensionality of the dataset is 784, which is much higher than the number of samples in each of the training sets and, therefore, defines a typical small sample setting.

We present in Table I the error rates, obtained on the test sets, when the conventional (i.e., using the \(w_0 = 0\) centroid), the IMS (see Algorithm 1) and the IMSA (see Algorithm 2) are exploited for model selection using the Maximal Discrepancy approach. The last row indicates the number of times that a method results to be the best performer. The results clearly show that the IMS and, especially, the IMSA procedures allow to obtain better performing models than the ones chosen by the conventional model selection. These claims are also supported by the results obtained using the Rademacher Complexity approach, as shown in Table II.

B. Human Gene Expression Datasets

In the experiments described in the previous section, we extracted small sample sets in order to fairly compare the theoretical model selection techniques on a large cardinality test set. In a real-world small sample setting, a test set of such size is not available: then, we reproduce the methodology used by [15], which consists in generating \(v\) different training/test pairs using a cross validation approach. In particular, we set \(v = 5\) for our experiments. If the number of patterns of a dataset is not exactly a multiple of \(v\), some patterns are left out of the training set: however, they are not neglected (as in many other applications) and they are simply added to every test set. Analogously to the analysis of the MNIST dataset, we create the training/test splitting using a random sampling approach in order to verify the performance of the different model selection techniques.

In particular, we make use of a series of human gene expression classification problems, presented in Table III. A first group of dataset is taken from the well-known GEMS repository [19]: Brain Tumor 1, Brain Tumor 2, DLBCL, Leukemia 1, Leukemia 2, Lung Cancer, Prostate Tumor, and SRBCT. In addition to these sets, we also make use of: Leukemia [20], Colon Cancer 1 [21], Colon Cancer 2 [22], Duke Breast Cancer [23] and Myeloma [24]. As, in
TABLE III  
CHARACTERISTICS OF THE HUMAN GENE EXPRESSIONS DATASETS.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Reference</th>
<th>$d$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain Tumor 1</td>
<td>[19]</td>
<td>5920</td>
<td>90</td>
</tr>
<tr>
<td>Brain Tumor 2</td>
<td>[19]</td>
<td>10367</td>
<td>50</td>
</tr>
<tr>
<td>Colon Cancer 1</td>
<td>[21]</td>
<td>22283</td>
<td>47</td>
</tr>
<tr>
<td>Colon Cancer 2</td>
<td>[23]</td>
<td>2000</td>
<td>62</td>
</tr>
<tr>
<td>DLBCL</td>
<td>[19]</td>
<td>5469</td>
<td>77</td>
</tr>
<tr>
<td>Duke Breast Cancer</td>
<td>[23]</td>
<td>7129</td>
<td>44</td>
</tr>
<tr>
<td>Leukemia</td>
<td>[20]</td>
<td>7129</td>
<td>72</td>
</tr>
<tr>
<td>Leukemia 1</td>
<td>[19]</td>
<td>5327</td>
<td>72</td>
</tr>
<tr>
<td>Leukemia 2</td>
<td>[19]</td>
<td>11225</td>
<td>72</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>[19]</td>
<td>12600</td>
<td>203</td>
</tr>
<tr>
<td>Myeloma</td>
<td>[24]</td>
<td>20932</td>
<td>105</td>
</tr>
<tr>
<td>Prostate Tumor</td>
<td>[19]</td>
<td>10509</td>
<td>102</td>
</tr>
<tr>
<td>SRBCT</td>
<td>[19]</td>
<td>2308</td>
<td>83</td>
</tr>
</tbody>
</table>

TABLE IV  
HUMAN GENE EXPRESSIONS DATASETS - NUMBER OF ERRORS ON THE TEST SETS USING THE MAXIMAL DISCREPANCY APPROACH. BEST RESULTS ARE IN BOLD FACE.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Conventional</th>
<th>IMS</th>
<th>IMSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain Tumor1</td>
<td>6.60 ± 0.00</td>
<td>1.00 ± 0.81</td>
<td>1.00 ± 0.81</td>
</tr>
<tr>
<td>Brain Tumor2</td>
<td>0.40 ± 0.63</td>
<td>0.40 ± 0.63</td>
<td>0.40 ± 0.63</td>
</tr>
<tr>
<td>Colon Cancer1</td>
<td>6.00 ± 0.00</td>
<td>2.00 ± 1.82</td>
<td>2.00 ± 1.82</td>
</tr>
<tr>
<td>Colon Cancer2</td>
<td>6.00 ± 0.00</td>
<td>2.60 ± 1.54</td>
<td>2.60 ± 1.54</td>
</tr>
<tr>
<td>DLBCL</td>
<td>7.00 ± 0.00</td>
<td>1.80 ± 0.96</td>
<td>1.80 ± 0.96</td>
</tr>
<tr>
<td>Duke Breast Cancer</td>
<td>5.80 ± 1.50</td>
<td>2.20 ± 1.50</td>
<td></td>
</tr>
<tr>
<td>Leukemia</td>
<td>5.00 ± 0.00</td>
<td>1.20 ± 0.96</td>
<td>1.20 ± 0.96</td>
</tr>
<tr>
<td>Leukemia1</td>
<td>9.80 ± 0.51</td>
<td>0.20 ± 0.51</td>
<td>0.20 ± 0.51</td>
</tr>
<tr>
<td>Leukemia2</td>
<td>8.00 ± 0.63</td>
<td>0.60 ± 0.63</td>
<td>0.60 ± 0.63</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>16.00 ± 0.00</td>
<td>2.80 ± 1.50</td>
<td>2.80 ± 1.50</td>
</tr>
<tr>
<td>Myeloma</td>
<td>7.00 ± 0.00</td>
<td>0.00 ± 0.00</td>
<td>0.00 ± 0.00</td>
</tr>
<tr>
<td>Prostate Tumor</td>
<td>7.60 ± 2.65</td>
<td>2.40 ± 1.03</td>
<td>2.40 ± 1.03</td>
</tr>
<tr>
<td>SRBCT</td>
<td>9.00 ± 0.00</td>
<td>0.40 ± 0.63</td>
<td>0.40 ± 0.63</td>
</tr>
<tr>
<td># best</td>
<td>1</td>
<td>12</td>
<td>13</td>
</tr>
</tbody>
</table>

In this paper, we are targeting bi-class classification, we map multi-class datasets into bi-class ones by simply grouping classes so to obtain almost balanced problems.

Analogously to the MNIST dataset, we show in Tables IV and V the average number of errors, performed by the different model selection techniques when the Maximal Discrepancy and the Rademacher Complexity approaches are used. In particular, IMS and especially IMSA outperform the conventional procedure and are characterized by a similar performance.

VI. CONCLUSION

This work addresses two issues: the first one is the possibility of selecting a different hypothesis space respect to the one used by the conventional SVM formulation; the second one is the exploitation of this greater flexibility to improve the generalization ability of the trained classifier. While the first problem could be seen as a theoretical curiosity, we showed that its solution leads to better results in practice, at least in the small–sample setting. Further research is currently underway to understand how to exploit the new formulation in other ways. An example is the choice of the alternative centroid(s), through some a–priori information about the classification problem, instead of deriving it from a portion of the training set.

APPENDIX

In this appendix, we propose the non-linear kernel extensions for the SVM formulations, presented in Section IV. We can always apply a non-linear mapping such that

\[ x \mapsto \phi(x). \]  

Problem (19) then can be reformulated as:

\[
\begin{align*}
\min_{w, b, \xi} & \quad e^T \xi \\
\text{subject to} & \quad \|w - \hat{w}\|^2 \leq \rho^2 \\
& \quad y_i \left( w^T \phi(x_i) + b \right) \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n] \\
& \quad \xi_i \geq 0 \quad \forall i \in [1, \ldots, n],
\end{align*}
\]

or, alternatively, by exploiting the results of [8]

\[
\begin{align*}
\min_{w, b, \xi} & \quad \frac{1}{2} \|w - \hat{w}\|^2 + C e^T \xi \\
\text{subject to} & \quad y_i \left( w^T \phi(x_i) + b \right) \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n] \\
& \quad \xi_i \geq 0 \quad \forall i \in [1, \ldots, n].
\end{align*}
\]

Analogously to Section IV-A, we can compute the dual formulation as well:

\[
\begin{align*}
\min_{\beta} & \quad \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_i \beta_j y_i y_j K(x_i, x_j) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_i \beta_j y_i y_j K(x_i, \hat{x}_j) - 1) \\
& \quad 0 \leq \beta_i \leq C \quad \forall i \in [1, \ldots, n] \\
& \quad y^T \beta = 0,
\end{align*}
\]

where \(K(\cdot, \cdot) = \phi(\cdot)^T \phi(\cdot)\) is a suitable kernel function [1]. In the previous problem, we set, without any loss of generality
\[
\hat{w} = \sum_{j=1}^{n_c} \hat{y}_j \hat{\beta}_j \phi(\hat{x}_j),
\]
(46)

where \( \hat{y}_j, \hat{\beta}_j \) and \( \hat{x}_j \in \mathbb{R}^d \), \( j = 1, \ldots, n_c \), are found while calculating the centroid for the hypothesis spaces (see Section IV-B).

Finally, we can compute the non-linear formulations for problems (27), (28) and (29) (i.e. where augmented centroids are used). For this purpose, we need to re-define some problems (27), (28) and (29) (i.e. where augmented centroids calculations are used). From these formulations, it is straightforward to derive the extension to the case of multiple centroids.

\[ \min_{w, \xi} e^T \xi \]
\[ \sum_{i=1}^{n} \xi_i \geq 0 \quad \forall i \in [1, \ldots, n] \]
\[ y_i \hat{w}^T \phi(x_i) \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n] \]
\[ \xi_i \geq 0 \quad \forall i \in [1, \ldots, n] \]

Problem (28) is reformulated as:

\[ \min_{w, \xi} \frac{1}{2} \|w - \hat{w}\|^2 + C e^T \xi \]
\[ y_i \hat{w}^T \phi(x_i) \geq 1 - \xi_i \quad \forall i \in [1, \ldots, n] \]
\[ \xi_i \geq 0 \quad \forall i \in [1, \ldots, n] \]

Finally, the dual formulation (29) becomes:

\[ \min_{\beta} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\beta}_j \hat{y}_j \hat{y}_j^T \phi(x_i)^T \phi(x_j) + 1 \]
\[ + \sum_{i=1}^{n} \left[ \sum_{j=1}^{n} \hat{y}_j \hat{y}_j \phi(x_i)^T \phi(x_j) \right] \hat{\beta}_i \]
\[ 0 \leq \hat{\beta}_i \leq C \quad \forall i \in [1, \ldots, n] \]

where we exploited Eq. (46) for the centroid \( \hat{w} \). From these formulations, it is straightforward to derive the extension to the case of multiple centroids.

REFERENCES