Semi-Supervised Multiclass Kernel Machines with Probabilistic Constraints

Stefano Melacci and Marco Gori

Department of Information Engineering
University of Siena, 53100 - Siena, Italy
{mela,marco}@dii.unisi.it

Abstract. The extension of kernel-based binary classifiers to multiclass problems has been approached with different strategies in the last decades. Nevertheless, the most frequently used schemes simply rely on different criteria to combine the decisions of a set of independently trained binary classifiers. In this paper we propose an approach that aims at establishing a connection in the training stage of the classifiers using an innovative criterion. Motivated by the increasing interest in the semi-supervised learning framework, we describe a soft-constraining scheme that allows us to include probabilistic constraints on the outputs of the classifiers, using the unlabeled training data. Embedding this knowledge in the learning process can improve the generalization capabilities of the multiclass classifier, and it leads to a more accurate approximation of a probabilistic output without an explicit post-processing. We investigate our intuition on a face identification problem with 295 classes.

Key words: Multiclass Support Vector Machines, Probabilistic Constraints, Semi-Supervised Learning.

1 Introduction

In multiclass classification problems we have a set of $k > 2$ classes, and the goal is to construct a classifier which correctly predicts the class to which an input point belongs. Although many real-world classification problems are multiclass, many of the most efficient classifiers are specifically designed for binary problems ($k = 2$), such as Support Vector Machines (SVMs) [16].

The simplest strategy to allow them to handle a larger number of classes is commonly referred to as "one-versus-all" (OVA), and it consists in independently training k classifiers to discriminate each class from the $k - 1$ remaining ones. Given an input instance, the class label corresponding to the classifier which outputs the maximum value is selected [14]. Even if some more sophisticated schemes have been proposed (based on directed acyclic graphs, on error correcting coding theory, or on combination of different strategies [13, 5, 4], for example) the OVA strategy is still one of the most popular approaches, since it has been shown to be as accurate as the most of the other techniques [14].

In this paper we propose to tackle the OVA multiclass problem for regularized kernel machines in an innovative fashion, enforcing the k outputs of the

classification functions to fulfill a probabilistic relationship. In detail, the *probabilistic constraints* represent a domain information on the multiclass problem that we enforce on the available unlabeled training data, in a Semi-Supervised setting. As a matter of fact, the constraints introduce a dependency among the training stages of the k classifiers, encouraging an inductive transfer that may improve the generalization capabilities of the multiclass classifier.

For this reason our work is related to Multi-Task learning [3] and it shares some principles with approaches that post-process the output of an SVM to approximate posterior estimates [12, 17]. It is substantially different from logistic regression models [15, 7] that yield probabilistic outcomes based on a maximum likelihood argument. In particular, we focus on the improvements in terms of generalization performance that the proposed constraining can introduce, and not on the strict definition of a model that is guaranteed to produce a probabilistic output. Nevertheless, we show that it is satisfactorily approximated by our soft-constraining procedure.

We investigate our approach on a face identification problem with 295 classes, using the publicly available XM2VTS multimodal dataset. A detailed experimental analysis shows improvements in the quality of the classifier, successfully exploiting the interaction that is established by the probabilistic constraints.

This paper is organized as follows. In Section 2 the Semi-Supervised binary classifiers on which we focus are introduced. In Section 3 the probabilistic constraints are presented. Section 4 collects our experimental results, and, finally, in Section 5 some conclusions are drawn.

2 Multiclass Learning with Constraints

Given a set of objects in $\mathcal{X}$, let us suppose that each object $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is described by a d-dimensional vector of features. In a generic k-class classification problem, we want to infer the function $c : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is a set of labels. We indicate with $y_i \in \mathcal{Y}$ the label associated to $\boldsymbol{x}_i$. Suppose that there is a probability distribution P on $\mathcal{X} \times \mathcal{Y}$, according to which data are generated.

In a Supervised classification problem, we have a labeled training set $\mathcal{L}$ of l pairs,

$$\mathcal{L} = \{(\boldsymbol{x}_i, y_i) | i = 1, \ldots, l,\ \boldsymbol{x}_i \in \mathcal{X},\ y_i \in \mathcal{Y}\},$$

and the classifier is trained to estimate $c(\cdot)$ using the information in $\mathcal{L}$. A labeled validation set $\mathcal{V}$, if available, is used to tune the classifier parameters, whereas the generalization capabilities are evaluated on an out-of-sample test set $\mathcal{T}$, in a typical inductive setting.

In the Semi-Supervised learning framework, we have also a set $\mathcal{U}$ of u unlabeled training instances,

$$\mathcal{U} = \{\boldsymbol{x}_i | i = 1, \ldots, u,\ \boldsymbol{x}_i \in \mathcal{X}\},$$

that is exploited to improve the quality of the classifier. In a practical context, unlabeled data can be acquired relatively easily, whereas labeling requires the

expensive work of one or more supervisors, so that frequently we have $u >> l$. Unlabeled samples are drawn accordingly to the marginal distribution $P_{\mathcal{X}}$ of P, and the Semi-Supervised framework attempts to incorporate them into the learning process in different ways. Popular Semi-Supervised classifiers make specific assumptions on the geometry of $P_{\mathcal{X}}$, such as, for example, having the structure of a Riemannian manifold [9]. We indicate with n the total number of labeled and unlabeled training points collected in the set $\mathcal{S} = \mathcal{L} \cup \mathcal{U}$, $\mathcal{L} \cap \mathcal{U} = \emptyset$, where the union and intersection are intended to consider only the first element of each pair in $\mathcal{L}$ ($n = l$ in the supervised setting).

SVM-like kernel machines are specifically designed as binary classifiers. Their extension to multiclass classification in a "one-versus-all" (OVA) scheme, consists in the *independent* training of k binary classifiers that discriminate each class from the other $k-1$ ones. We indicate with f_j the function learnt by the j-th classifier, $j = 1, \ldots, k$, and with y_{ij} the target of the sample $\boldsymbol{x}_i$ in the j-th binary problem. Following the Multi-Task learning framework [3], each function represents a specific "task", collected in the vector $\boldsymbol{f} = [f_1, \ldots, f_k]^T$. In the classical multiclass scenario, all the tasks are defined on the same set of points, and the decision function $c(\cdot)$ that determines the overall output of the classifier is

$$c(\boldsymbol{x}) = \arg\max_j f_j(\boldsymbol{x}). \tag{1}$$

When some prior knowledge on the correlation among the tasks is available, we propose to model it with a set of q constraints on $\{f_1(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x})\}$, represented by the functions $\phi_h : \mathbb{R}^k \to \mathbb{R}$:

$$\phi_h(f_1(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x})) \qquad h = 1, \ldots, q \tag{2}$$

that hold $\forall \boldsymbol{x} \in \mathcal{X}$.

Given a positive definite Kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we indicate with $\mathcal{H}$ the Reproducing Kernel Hilbert Space (RKHS) corresponding to it, and with $\|\cdot\|_{\mathcal{H}}$ the norm of $\mathcal{H}$. Each f_j belongs to $\mathcal{H}$ [1], and we formulate the learning problem in the risk minimization scheme, leading to

$$\min_{\boldsymbol{f}} \left(\sum_{j=1}^{k} \sum_{i=1}^{l} V(f_j(\boldsymbol{x}_i), y_{i,j}) + \sum_{j=1}^{k} \lambda_j \cdot \|f_j\|_{\mathcal{H}}^2 + C(\boldsymbol{f}) \right) \tag{3}$$

where

$$C(\boldsymbol{f}) = \sum_{h=1}^{q} \gamma_h \cdot \sum_{i=1}^{n=l+u} L_h(\phi_h(f_1(\boldsymbol{x}_i), \ldots, f_k(\boldsymbol{x}_i))).$$

In detail, the loss function $V(f_j(\boldsymbol{x}_i), y_{i,j})$ measures the fitting quality of each f_j with respect to the targets y_{ij}, and $\|f_j\|_{\mathcal{H}}^2$ is a smoothing factor weighted

[1] More generally, we can define each f_j in its own RKHS. We consider the case of a shared RKHS among the functions just to simplify the description of our idea.

by λ_j, that makes the learning problem well-posed[2]. Unlike the previous terms, $C(\cdot)$ is a penalty function that models a correlation among the tasks during the learning process, expressed by the constraints ϕ_h, $h = 1, \ldots, q$. The parameters $\{\gamma_h\}_{h=1}^{q}$ allow us to weight the contribution of each constraint, and the penalty loss function $L_h(\phi_h)$ is positive when the constraint is violated, otherwise it is zero. To simplify the notation we avoided additional scaling factors on the terms of the summation in Eq. 3.

In this soft-constraining scheme, there are no guarantees of ending up in a classifier that perfectly fulfills the relationships of ϕ_h, $h = 1, \ldots, q$, whereas some violations are tolerated. As a matter of fact, the solution of Eq. 3 is a trade-off among label fitting, smoothness on the entire input space, and problem specific constraints. Note that if $V(f_j(\boldsymbol{x}_i), y_{i,j})$ is a linear hinge loss, and we remove the $C(\cdot)$ term (i.e. $\gamma_h = 0$, $h = 1, \ldots, q$) we get SVM classifiers.

If the loss function V and the term $C(\cdot)$ are convex, the problem of Eq. 3 admits a unique minimizer. The optimal solution of Eq. 3 can be expressed as a kernel expansion, as stated in the following Representer Theorem.

Theorem 1. *Let us consider the minimization problem of Eq. 3, where the function $f_1, \ldots, f_k$ belong to a RKHS $\mathcal{H}$. Then the optimal solution f_j^*, $j = 1, \ldots, k$ is expressed as*

$$f_j^*(\boldsymbol{x}) = \sum_{i=1}^{n=l+u} \alpha_{ij} K(\boldsymbol{x}, \boldsymbol{x}_i), \qquad j = 1, \ldots, k$$

where $K(\cdot, \cdot)$ is the reproducing kernel associated to $\mathcal{H}$, $\boldsymbol{x}_i \in \mathcal{S}$, and α_{ij} are n scalar values.

Proof: Using a simple orthogonality argument, the proof is a straightforward extension of the representer theorem for plain kernel machines [16]. Is is only sufficient to notice that V is measured on the l labeled training points only, whereas the penalty term $C(\cdot)$ involves a set of constraints evaluated on all the $n = l + u$ samples belonging to $\mathcal{S}$, so that the optimal solution lyes in the span of the n training points (both labeled and unlabeled), as in [1]. □

In the next section we will describe the probabilistic constraints using an instance of the described learning framework. Nevertheless this Semi-Supervised scheme is generic and it can be applied to model any kind of interaction among tasks that comes from a problem-dependent prior knowledge.

3 Probabilistic Constraints

For the j-th task, we select $y_{ij} \in \{0, 1\}$, where $y_{ij} = 1$ means that $\boldsymbol{x}_i$ belongs to class j while 0 indicates that it belongs to the other classes, and we penalize the

[2] We are assuming that the *kernel* function is not yielding to interaction among the different tasks, but the essence of what is proposed could be directly extended to the general case of multitask *kernel* functions [2].

label fitting with a squared loss $V(f_j(\boldsymbol{x}_i), y_{ij}) = (f_j(\boldsymbol{x}_i) - y_{ij})^2$. Note that using a hinge loss leads to the same classification accuracies, as investigated in [14], and it would not make any substantial differences with respect to the selected V due to the nature of the constraints that we will introduce in the following (that will enforce f_j in $[0, 1]$).

In its unconstrained and fully Supervised formulation, the OVA scheme does not guarantee that the output values $f_1(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x})$, $\forall \boldsymbol{x} \in I\!R^d$ have the properties of a probability (i.e. that they are in $[0, 1]$ and they sum to one). In other words, the classifier do not fulfill what we refer to as the *probabilistic constraints*, that can be modeled with the following linear system,

$$\begin{cases} \sum_{j=1}^{k} f_j(\boldsymbol{x}) = 1 \\ f_j(\boldsymbol{x}) \geq 0 \quad j = 1, \ldots, k. \end{cases} \tag{4}$$

Clearly, for $\boldsymbol{x}_i \in \mathcal{L}$, this information is implicitly embedded on the targets y_{ij}, $j = 1, \ldots, k$. As a consequence, a hypothetic perfect fitting of labeled points would fulfill Eq. 4 $\forall \boldsymbol{x} \in \mathcal{L}$. On the other hand, each f_j is requested to be smooth in the RKHS, and a perfect fitting is generally not achieved.

Interestingly, Eq. 4 gives us a basic domain information on the problem that is supposed to hold in the entire input space. We exploit this information on the relationship among the functions f_j, $j = 1, \ldots, k$, to introduce an interaction among the tasks within their training stage. As a matter of fact Eq. 4 must hold also for points $\boldsymbol{x} \notin \mathcal{L}$, so that we can cast the problem in the Semi-Supervised setting described in Section 2, enforcing the probabilistic constraints also on the (largely available) unlabeled training data. Differently from approaches that estimate probabilities in a post-processing stage [12, 17] or from kernel logistic regression [15, 7], we do not aim at obtaining a perfectly fulfilled probabilistic output, but at improving the quality of the classifier by task interaction.

We can formulate the probabilistic constraints as a set of $q = k + 1$ linear functions that become zero when they are fulfilled,

$$\begin{cases} \phi_1^{sum}(f_1(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x}))) = \sum_{j=1}^{k} f_j(\boldsymbol{x}) - 1 \\ \phi_h^{pos}(f_{h-1}(\boldsymbol{x})) = \max(-f_{h-1}(\boldsymbol{x}), 0) \quad h = 2, \ldots, k+1. \end{cases} \tag{5}$$

In particular, in this specific problem only ϕ_1^{sum} involves all the k tasks, whereas ϕ_h^{pos}, $h = 2, \ldots, k+1$ model a prior knowledge on the single binary functions. The paired interaction of ϕ_1^{sum} and ϕ_h^{pos} is introduced in the optimization problem of Eq. 3 by the following $C(\cdot)$ term,

$$\begin{aligned} C(\boldsymbol{f}) = \sum_{i=1}^{n} \Big(& \gamma_1 L_1(\phi_1^{sum}(f_1(\boldsymbol{x}_i), \ldots, f_k(\boldsymbol{x}_i))) \\ & + \sum_{h=2}^{k+1} \gamma_h L_h(\phi_h^{pos}(f_{h-1}(\boldsymbol{x}_i))) \Big). \end{aligned} \tag{6}$$

In order to keep intact the squared nature of the problem, we select L_h, $h = 1, \ldots, q$ to be squared loss functions. A constraint violation is then quadratically

penalized. Moreover, the γ_h, $h = 2, \ldots, k+1$ are set to the same value, that we will indicate with γ (without any subscripts), to equivalently weight the ϕ_h^{pos} constraint in each task, whereas γ_1 is set to $k \cdot \gamma$. As a matter of fact we want to emphasize the effect of ϕ_1^{sum} in the minimization procedure, since it encourages the interaction among the binary classifiers. The λ_j, $j = 1, \ldots, k$, coefficients of Eq. 3 are set to λ.

Enforcing the probabilistic constraints with non-linear kernel functions $K(\cdot, \cdot)$ appears the most natural choice. Following the Representer Theorem (Theorem 1), their combination can model highly non-linear f_j, $j = 1, \ldots, k$, allowing the classifier to efficiently alter the shape of each of them accordingly to the interaction with the other ones, to the labeled data fitting and to the smoothness constraint. Popularly used kernels, such as the Gaussian kernel or the polynomial kernel of degree ≥ 2, are well suited for this approach. Modeling the constraints with linear kernels yields to f_j solutions that can easily degenerate towards a constant value, as we experienced, in particular if the dimension of the input space is small. As a matter of fact, enforcing the probabilistic relationship tends to "over constrain" the linear f_j, $j = 1, \ldots, k$. More generally, our approach is well suited for problems with a large number of classes, that emphasize the importance of the task interaction.

3.1 Training the Multiclass Classifier

In order to devise a compact formulation of the minimization problem of Eq. 3, we assume that the n training points of $\mathcal{S}$ are ordered so that the first l are the labeled ones and the remaining u are the unlabeled samples. We overload the notation of K, so that it also indicates the Gram matrix associated to the selected kernel function $K(\cdot, \cdot)$ evaluated on the training data, $K = [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1,\ldots,n}$. Let $A \in \mathbb{R}^{n,k}$ be the matrix where the j-th column collects the n coefficients α_{ij} of the kernel expansion of the j-th task (from Theorem 1). As a result, each column of $KA \in \mathbb{R}^{n,k}$ collects the outputs of a f_j function evaluated on the training points. The subscript is used to refer to a column of a given matrix, so that, for example, A_j indicates the j-th column of A. In $Y \in \{0,1\}^{l,k}$ we collect the task-specific targets for labeled points, i.e. the entry in position (i, j) is y_{ij}.$\mathbf{1} \in \mathbb{R}^n$ is the vector of n elements equal to 1 while $J = [I, O] \in \mathbb{R}^{l,n}$ is a rectangular matrix composed by the identity matrix $I \in \mathbb{R}^{l,l}$ and by a matrix $O \in \mathbb{R}^{l,u}$ of zeros. Finally, the notation $(\boldsymbol{v})_+$ indicates that all the negative elements of the vector $\boldsymbol{v}$ are set to zero.

The instance of the problem of Eq. 3 that we want to minimize is then

$$A^* = \arg\min_A \Big(\sum_{j=1}^k (JKA_j - Y_j)^T (JKA_j - Y_j) + \lambda \sum_{j=1}^k A_j^T K A_j + k\gamma (\sum_{j=1}^k KA_j - \mathbf{1})^T (\sum_{j=1}^k KA_j - \mathbf{1}) + \gamma \sum_{j=1}^k -A_j^T K (-KA_j)_+ \Big). \tag{7}$$

The objective function is continuous, piecewise quadratic (due to the piecewise linear ϕ_h^{pos}, $h = 2, \ldots, k+1$ and the quadratic loss functions V and L_h,

$h = 1, \ldots, k+1$), and strictly convex. As recently investigated for the case of Laplacian SVMs [9], we can efficiently optimize it in its primal formulation using Preconditioned Conjugate Gradient (PCG). In our specific problem, the gradient $\nabla_j \in \mathbb{R}^n$ of Eq. 7 with respect to the A_j coefficients of the j-th function f_j is

$$\nabla_j = 2K \left(J^T(JKA_j - Y_j) + \lambda A_j + k\gamma(\textstyle\sum_{j=1}^{k} KA_j - \mathbf{1}) - \gamma(-KA_j)_+ \right) \quad (8)$$

and preconditioning by the matrix K comes at no additional cost, as discussed in [9].

In order to optimize the multiclass problem, all the ∇_j, $j = 1, \ldots, k$ must be computed. The computational cost of each PCG iteration is $O(kn^2)$, due to the KA product, and the complexity is reduced if K is sparse. Moreover, selecting $\alpha_{ij} = 0$, $i = 1, \ldots, n$, $j = 1, \ldots, k$, as initial point for the optimization procedure, each gradient iteration can be easily parallelized by computing in a separate process each ∇_j, and sharing the KA_j vectors ($j = 1, \ldots, k$) among the parallel processes at the end of the iteration.

4 Experimental Results

Face recognition involves a large number of classes, corresponding to the number of subject to be recognized. We evaluate the performances of the proposed approach in the traditional face identification scenario, where the identity of a given input face must be retrieved among a set of known subjects. If each input face is known to belong to such set, the problem is casted in a *winner-take-all* scenario, where the identity predicted with the highest confidence is selected as overall decision of the classifier, as in the described OVA scheme. SVM-like regularized classifiers have been widely applied to face recognition, focusing on different aspects of the problem [11, 6].

We selected the XM2VTS multimodal database to test our system. It is a publicly available collection of face pictures, video sequences, speech recordings taken of 295 subjects, and distributed by the University of Surrey [10]. In particular, it collects 8 frontal view face pictures for each subject, acquired in four separate sessions uniformly distributed over a period of four months. Face images were acquired in controlled conditions (constant face-camera distance and lighting, uniform background) at the resolution of 720x576.

Data was preprocessed as in many popular eigenface-based face recognition approaches [18]. We cropped each image so that only the face region from eyebrows to the chin was kept; images were converted to gray scale and (uniformly) rescaled to 56x64, using the image height as a reference to compute the scaling factor; pixel values were rescaled to $[0, 1]$; Principal Component Analysis (PCA) was performed, and we kept the first 184 eigenfaces, describing 85% of the total variance. In Fig. 1 some examples of the cropped/scaled images are reported.

Following the data partitioning suggested in the second configuration of the so called "Lausanne" protocol (defined for face verification competitions on the

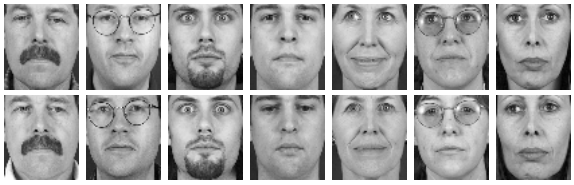

Fig. 1. Some examples of the cropped/scaled XM2VTS faces from two of the four sessions (top and bottom row).

XM2VTS data [8]) we split the available data as described in Table 1, where the details on the XM2VTS data are resumed. Moreover, the training set $\mathcal{S}$ was divided in the sets $\mathcal{L}$ and $\mathcal{U}$ of labeled and unlabeled points, respectively, simulating a Semi-Supervised scenario.

Table 1. The XM2VTS face dataset. For each subject there are 8 images (identified by the numbers $1, \ldots, 8$). The selected data splits follow the Lausanne protocol.

Dataset		**Subjects**	**Total Images**
XM2VTS		295	2360
Data Split		**Image IDs (per subject)**	**Total Images**
Training	$\mathcal{S} = \mathcal{L} \cup \mathcal{U}$	$1, 2, 3, 4$	1180
Validation	$\mathcal{V}$	5,6	590
Test	$\mathcal{T}$	7,8	590

We compared the proposed Multiclass Classifier with Probabilistic Constraints (MC-PC) with an unconstrained OVA Multiclass Support Vector Machines (MSVM) that it is one of the most popular approaches and it has been shown to be as accurate as the most of the other existing techniques [14]. Experiments have also been performed using a K-Nearest Neighbors (KNN) classifier with Euclidean distance, since it is frequently used in face recognition experiments.

For each experiment, and for all the compared algorithms, parameters were tuned by computing the error rate on the validation set $\mathcal{V}$ and selecting the best configuration. In the case of MC-PC, the optimal λ and γ were selected from the set $\{10^{-6}, 10^{-4}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. The same range of values was used for the weight λ of the regularization term in MSVM. The number of neighbors in KNN was changed from 1 to 10.

Our analysis is aimed at showing the behavior of the proposed constraining scheme in variable conditions, using different kernel functions and different configurations of the available supervision. Hence, we selected Gaussian and polynomial kernels, due to their popularity in many classification problems. A Gaussian kernel $k(\boldsymbol{a}, \boldsymbol{b}) = \exp \frac{-\|\boldsymbol{a}-\boldsymbol{b}\|^2}{2\sigma^2}$ was tested with $\sigma \in \{5, 10, 20\}$ to assess the behavior of larger and tighter Gaussian functions (rbf). The polynomial ker-

nel $k(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{a}^T \boldsymbol{b} + 1)^p$ was tested with a degree $p = 2$ and with $p = 3$ (poly). We iteratively increased the fraction of labeled training points to evaluate the behavior of our Semi-Supervised approach as the size of $\mathcal{L}$ increases, and, consequently the amount of unlabeled training points in the set $\mathcal{U}$ decreases. We have $k = 295$ subjects, and for each of them the number of labeled images in $\mathcal{L}$ has been incrementally changed from 1 to 3 whereas $\mathcal{U}$ is reduced from 3 to 1 unlabeled points. The corresponding results are collected in Table 2, where the transductive and inductive configurations are evaluated.

Table 2. The *error rates* on the set $\mathcal{U}$ (transductive) and on the test set $\mathcal{T}$ (inductive). A one-vs-all Multiclass SVM (MSVM) and a Multiclass Classifier with the Probabilistic Constraints (MC-PC) are compared. Two kernel functions and a different numbers of labeled ($|\mathcal{L}|$) and unlabeled ($|\mathcal{U}|$) training points are used ($k = 295$). The variation of correctly identified faces between MSVM and MC-PC is reported in brackets.

Transductive Setting

Kernel	**Classifier**	$\lvert\mathcal{L}\rvert = 1 \cdot k$ $\lvert\mathcal{U}\rvert = 3 \cdot k$	$\lvert\mathcal{L}\rvert = 2 \cdot k$ $\lvert\mathcal{U}\rvert = 2 \cdot k$	$\lvert\mathcal{L}\rvert = 3 \cdot k$ $\lvert\mathcal{U}\rvert = 1 \cdot k$
rbf ($\sigma = 5$)	MSVM MC-PC	34.35 **31.53 (+25)**	28.47 **27.97 (+3)**	8.81 **8.14 (+2)**
rbf ($\sigma = 10$)	MSVM MC-PC	31.98 **31.30 (+6)**	26.10 **25.08 (+6)**	7.46 7.46
rbf ($\sigma = 20$)	MSVM MC-PC	32.09 **31.98 (+1)**	25.59 **25.41 (+1)**	9.15 **8.47 (+2)**
poly ($p = 2$)	MSVM MC-PC	36.95 **34.92 (+18)**	**30** 30.17 (-1)	13.9 **9.83 (+12)**
poly ($p = 3$)	MSVM MC-PC	39.77 **39.66 (+1)**	36.27 **35.93 (+2)**	14.92 **14.24 (+2)**
	KNN	*40.68*	*39.83*	*18.64*

Inductive Setting

Kernel	**Classifier**	$\lvert\mathcal{L}\rvert = 1 \cdot k$ $\lvert\mathcal{U}\rvert = 3 \cdot k$	$\lvert\mathcal{L}\rvert = 2 \cdot k$ $\lvert\mathcal{U}\rvert = 2 \cdot k$	$\lvert\mathcal{L}\rvert = 3 \cdot k$ $\lvert\mathcal{U}\rvert = 1 \cdot k$
rbf ($\sigma = 5$)	MSVM MC-PC	41.36 **39.32 (+12)**	30.17 **29.49 (+4)**	19.32 **18.98 (+2)**
rbf ($\sigma = 10$)	MSVM MC-PC	36.93 **36.44 (+3)**	24.92 **24.41 (+3)**	15.25 15.25
rbf ($\sigma = 20$)	MSVM MC-PC	37.29 **36.95 (+2)**	25.93 25.93	15.59 **15.25 (+2)**
poly ($p = 2$)	MSVM MC-PC	43.56 **42.37 (+7)**	31.19 **30.51 (+4)**	20.34 **19.15 (+7)**
poly ($p = 3$)	MSVM MC-PC	48.81 **47.12 (+10)**	39.83 **37.46 (+14)**	28.64 **26.78 (+11)**
	KNN	*50.64*	*42.03*	*32.71*

The experimental setup of Table 2 with $|\mathcal{L}| = 1 \cdot k$ and $|\mathcal{U}| = 3 \cdot k$ is the one that is closer to a truly Semi-Supervised setting, where a large amount of unlabeled points are available and just a few labels can be fed to the classifier. In roughly all the experiments (and with all the described kernel functions) the introduction of the probabilistic constraints improves the quality of the classifier, both in a transductive framework (on the set $\mathcal{U}$) and in the inductive one (on the set $\mathcal{T}$), showing an increment of the generalization capabilities.

As the number of labeled data increases, we move towards a fully Supervised setting and we reasonably expect a weaken impact of the probabilistic constraints, since the information that they carry is already included on training labels.

In particular, in the case of Gaussian kernel, the error rate on test data is improved mainly when the kernel width is small. As a matter of fact, due to the very local support of the kernel, the information on labeled points is not enough to fulfill the probabilistic constraints on the set $\mathcal{U}$, and the classifier can benefit from its explicit enforcement, even in this close-to-fully Supervised setup. When a polynomial kernel is used, the interaction among the 295 binary classifiers introduced by the probabilistic constraints keeps increasing the quality of the classifier, since they are far from being fulfilled in the whole space, and the action of our soft-constraining can be appreciated.

Those intuitions are confirmed by the graphs in Fig. 2, where we investigate "how strongly" the output values f_j, $j = 1, \ldots, k$ fulfill the probabilistic constraints in our Semi-Supervised scheme, with respect to the unconstrained case. The Mean Squared Error (MSE) of the unitary sum (ϕ_1^{sum}) and non-negativity (ϕ_h^{pos}, $h = 2, \ldots, k+1$) constraints on training and test data is reported. Thanks to our soft constraining procedure, the output values f_j, $j = 1, \ldots, k$ are very close to a probability. When only 1 labeled example per subject is used to train the classifier, the effect can be significantly appreciated, whereas as such number increases, the output of the unconstrained classifier becomes more similar to the constrained one, since labeled training data is the majority portion of $\mathcal{S}$.

The percentage of points for which $f(\boldsymbol{x}) < 0$ (reported over the plots) does not have significant changes when the constraints are introduced, due to the selected squared penalty approach that do not favor sparsity, whereas the fulfillment of such constraints is improved. Finally, we can see that in the case of Gaussian kernel the outputs f_j, $j = 1, \ldots, k$, of the unconstrained classifier are more similar to a probability than in the case of a polynomial kernel, where the importance of the explicit constraining is evident.

5 Conclusions

In this paper we presented an innovative approach to multiclass classification for popular kernel-based binary classifiers. We casted the "one-versus-all" k-class problem in the Semi-Supervised learning framework, where a set of probabilistic constraints is introduced among the outputs of the k classifiers, establishing an interaction in their training stages that biases an inductive transfer of in-

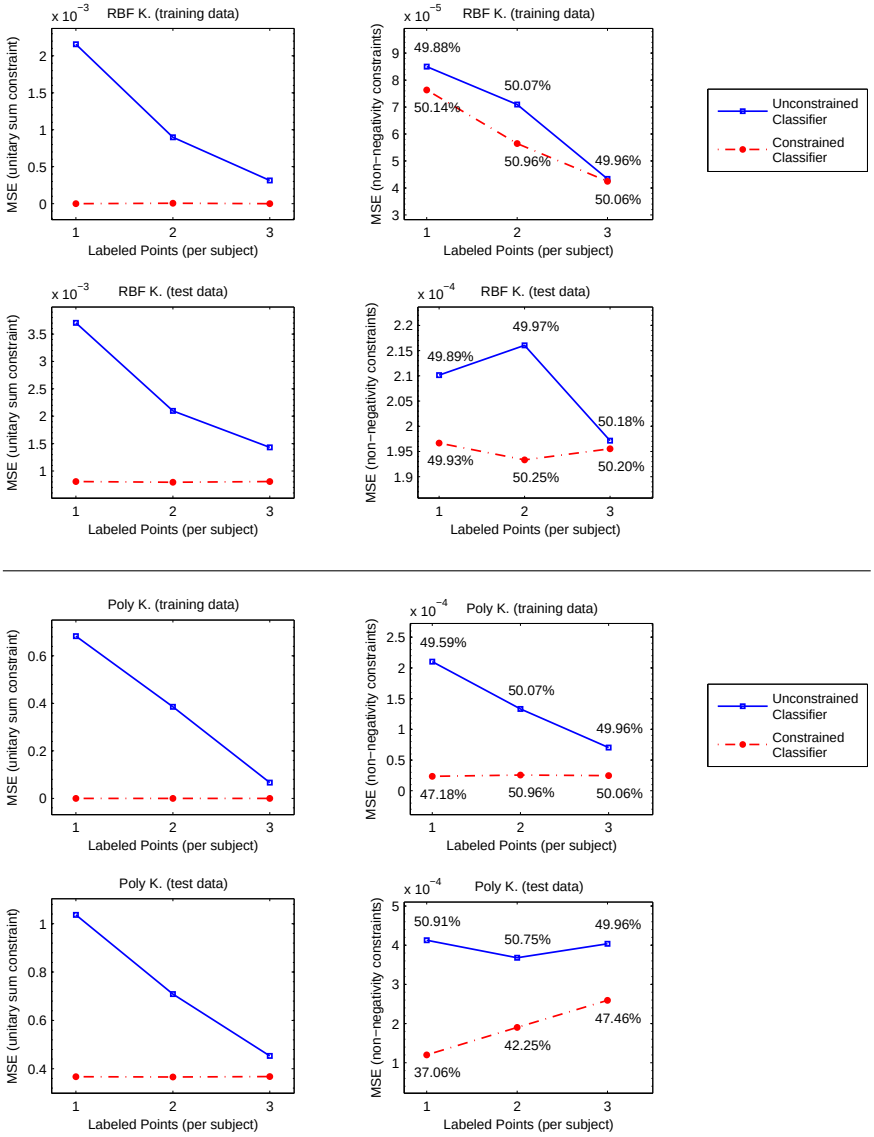

Fig. 2. The Mean Squared Error (MSE) of the unitary sum constraint (ϕ_1^{sum}) and of the non-negativity constraints (ϕ_h^{pos}, $h = 2, \ldots, k+1$, where the reported MSE is averaged over the k measurements) on training and test data, in function of the number of labeled examples per subject. In the latter, the percentage of points for which $f(\boldsymbol{x}) < 0$ is also displayed. The two plots in each graph describe the behavior of a classifier in which such constraints were or were not enforced during the training stage. In the group of graphs on top, a radial basis function kernel (RBF) with $\sigma = 10$ is used, whereas the group on bottom refers to a polynomial kernel (Poly) of degree 3.

formation. The experiments on a face identification problem with 295 classes showed improvements in the generalization capabilities of the multiclass classifier, together with a more accurate approximation of a probabilistic output. Interestingly, the proposed constraining scheme is general, and it also applies to different categories of classifiers.

References

1. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research 7, 2399–2434 (2006)
2. Caponnetto, A., Micchelli, C., Pontil, M., Ying, Y.: Universal multi-task kernels. Journal of Machine Learning Research 9, 1615–1646 (2008)
3. Caruana, R.: Multitask learning. Machine Learning 28(1), 41–75 (1997)
4. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research 2, 265–292 (2002)
5. Crammer, K., Singer, Y.: On the learnability and design of output codes for multiclass problems. Machine Learning 47(2), 201–233 (2002)
6. Heisele, B., Ho, P., Wu, J., Poggio, T.: Face recognition: component-based versus global approaches. Computer Vision and Image Understanding 91(1-2), 6–21 (2003)
7. Karsmakers, P., Pelckmans, K., Suykens, J.: Multi-class kernel logistic regression: a fixed-size implementation. In: Int. Joint Conf. on Neural Networks. pp. 1756–1761. IEEE (2007)
8. Matas, J., Hamouz, M., Jonsson, K., et al.: Comparison of face verification results on the XM2VTS database. In: Int. Conf. on Pattern Recognition. vol. 4, pp. 858–863. IEEE Computer Society (2000)
9. Melacci, S., Belkin, M.: Laplacian Support Vector Machines Trained in the Primal. Journal of Machine Learning Research 12, 1149–1184 (March 2011)
10. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTSDB: The Extended M2VTS Database. In: Proc. of the Int. Conf. on Audio and Video-based Biometric Person Authentication. pp. 72–79 (1999)
11. Phillips, P.: Support vector machines applied to face recognition. Advances in Neural Information Processing Systems pp. 803–809 (1999)
12. Platt, J.: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. Advances in Kernel Methods Support Vector Learning pp. 61–74 (2000)
13. Platt, J., Cristianini, N., Shawe-Taylor, J.: Large margin DAGs for multiclass classification. Advances in NIPS 12(3), 547–553 (2000)
14. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. Journal of Machine Learning Research 5, 101–141 (2004)
15. Roth, V.: Probabilistic discriminative kernel classifiers for multi-class problems. In: Proc. of the 23rd DAGM-Symposium on Pattern Recognition. pp. 246–253. Springer-Verlag, London, UK (2001)
16. Scholkopf, B., Smola, A.: Learning with Kernels. MIT Press (2001)
17. Wu, T., Lin, C., Weng, R.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005 (2004)
18. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys 35(4), 399–458 (2003)