TRANSDUCTIVE INFERENCE & KERNEL DESIGN FOR OBJECT CLASS SEGMENTATION

Phong D. Vo, Hichem Sahbi

TRANSDUCTIVE INFERENCE & KERNEL DESIGN FOR OBJECT CLASS SEGMENTATION

Dinh-Phong Vo, Hichem Sahbi

CNRS Telecom ParisTech

ABSTRACT

Transductive inference techniques are nowadays becoming standard in machine learning due to their relative success in solving many real-world applications. Among them, kernel-based methods are particularly interesting but their success remains highly dependent on the choice of kernels. The latter are usually handcrafted or designed in order to capture better similarity in training data. In this paper, we introduce a novel transductive learning algorithm for kernel design and classification. Our approach is based on the minimization of an energy function mixing i) a reconstruction term that factorizes a matrix of input data as a product of a learned dictionary and a learned kernel map ii) a fidelity term that ensures consistent label predictions with those provided in a ground-truth and iii) a smoothness term which guarantees similar labels for neighboring data and allows us to iteratively diffuse kernel maps and labels from labeled to unlabeled data. Solving this minimization problem makes it possible to learn both a decision criterion and a kernel map that guarantee linear separability in a high dimensional space and good generalization performance. Experiments conducted on object class segmentation, show improvements with respect to baseline as well as related work on the challenging VOC database.

1. INTRODUCTION

Existing machine inference techniques may be categorized into *inductive* and *transductive* [1]. The former consists in finding a decision function from a labeled training set, and uses that function in order to generalize across unlabeled data. Among popular inductive techniques support vector machines (SVMs) [1, 2] are well studied and proved to be performant in many real-world applications including object recognition, text analysis and bioinformatics [3, 4, 5]. The success of SVMs is highly dependent on the choice of kernels; existing ones include the linear, the gaussian and the histogram intersection. However, usual kernels may not be appropriate in order to capture the actual and the "semantic" similarity between data for some specific tasks. Variants known as multiple kernels (MKL) [6, 7, 8] consider convex (and possibly sparse) linear combinations of elementary kernels and proved to be more suitable.

Even-though performant, the success of these methods, also depends on cardinality of the labeled data. In many applications such as object class segmentation [9], labeled data is rare and expensive; only a very small fraction of training data is labeled and the unlabeled data may not follow the same distribution as the labeled one, so learning kernels using inductive inference techniques is clearly not appropriate. Alternative approaches [10] may include the unlabeled data as a part of the learning process and this is known as transductive inference. The concept of transductive inference, or transduction, was pioneered by Vapnik [1]. It relates to semi-supervised learning and relies on the i) smoothness assumption which states that close data in a high-density area of the input space, should have similar labels [10, 11] and ii) the cluster assumption which finds decision rules in low density areas of the input space [12, 9]. In that context, transductive versions of SVMs were also introduced [13, 11]; they build decision functions by optimizing the parameters of a learning model together with the labels of the unlabeled data. This turned out to be very useful in order to overcome the limited cardinality of the labeled data w.r.t the number of training parameters.

In this paper we introduce a novel transductive learning algorithm, for classification and kernel learning, based on constrained matrix factorization. Our factorization produces *a kernel map* that takes data from the input space into a high dimensional space in order to guarantee their linear separability while maximizing their margin. This margin property, however, does not necessarily guarantee good generalization performance on the unlabeled set, if the latter is drawn from a different probability distribution compared to the labeled data [1, 9]. Therefore and beside maximizing the margin, our transductive approach includes a regularization term that enforces smoothness in the resulting kernel map in order to correctly diffuse labels to the unlabeled data. Following our formulation, and in contrast to MKL, our learning model is not restricted to only convex linear combinations of existing kernels; indeed it is model-free. Furthermore, it also takes advantage of both labeled and unlabeled data and this results into better generalization performances as corroborated by our experiments.

The remainder of this paper is organized as follows. We introduce our transductive learning approach and kernel design in Section 2 and the implementation of our optimization procedure in Section 3. We illustrate the application of our method to object class segmentation in Section 4. We conclude the paper in Section 5 while providing a possible extension for a future work.

2. INFERENCE AND KERNEL DESIGN

Define $\mathcal{X} \subseteq \mathbb{R}^n$ as an input space corresponding to all the possible image features and let $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell, \ldots, \mathbf{x}_m\}$ be a finite subset of $\mathcal{X}$ with a particular order. This order is defined so only the first ℓ labels of $\mathcal{S}$, denoted $\{y_1, \ldots, y_\ell\}$ (with $y_i \in \{-1, +1\}$), are known. In many real-world applications only a few data is labeled (i.e., $\ell \ll m$) and its distribution may be different from the unlabeled data. We can view $\mathcal{S}$ as a matrix $\mathbf{X}$ in which the i^{th} row corresponds to $\mathbf{x}_i$. Our objective is to build both a decision criterion and an optimal *kernel map* in order to infer the unknown labels $\{y_{\ell+1}, \ldots, y_m\}$.

2.1. Max-margin Inference and Kernel Design

Inductive learning aims to build a decision function f that predicts a label y for any given input data $\mathbf{x}$; this function is trained on $\mathcal{S}' = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ and used in order to infer labels on $\mathcal{S} \backslash \mathcal{S}'$. In the max-margin classification [1], we consider ϕ as a mapping of the input data (in $\mathcal{X}$) into a high dimensional space $\mathcal{H}$. The dimension of $\mathcal{H}$ is usually sufficiently large (possibly infinite) in order to guarantee linear separability of data.

Assuming data linearly separable in $\mathcal{H}$, the max-margin inductive learning finds a hyperplane f (with a normal $\mathbf{w}$ and shift b) that separates ℓ training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ while maximizing their margin. The margin is defined as twice the distance between the closest training samples w.r.t f and the optimal $(\hat{\mathbf{w}}, \hat{b})$ correspond to

$$\underset{\mathbf{w},b}{\text{argmin}} \ \tfrac{1}{2}\|\mathbf{w}\|_2^2 \ \text{ s.t } \ y_i\left(\mathbf{w}'\phi(\mathbf{x_i}) + b\right) \geq 1, \quad i = 1, \ldots, \ell, \tag{1}$$

which is the primal form of the hard margin support vector machine [1], $\|.\|_2^2$ is the L_2 norm and $\mathbf{w}'$ is the transpose of $\mathbf{w}$. Given $x_i \in \mathcal{S}\backslash\mathcal{S}'$, the class of $\mathbf{x}_i$ in $\{-1, +1\}$ is decided by the sign of $f(\mathbf{x}_i) = \mathbf{w}'\phi(\mathbf{x}_i) + b$. Following the kernel trick [1], one may show that $f(\mathbf{x}_i)$ can also be expressed as $\sum_{j=1}^{\ell} \alpha_j y_j \ \kappa(\mathbf{x}_i, \mathbf{x}_j) + b$, here $(\alpha_1 \ldots \alpha_\ell)'$ is a vector of positive real-valued training parameters and $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) \rangle$ is a symmetric, continuous, positive (semi-definite) kernel function [2]. The closed form of $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ is defined among a collection of existing kernels including linear, gaussian and histogram intersection; but the underlying mapping $\phi(\mathbf{x}) \in \mathcal{H}$ is usually *implicit*, i.e., it does exist but it is not necessarily known and may be infinite dimensional.

We propose in the remainder of this section a new approach that builds *explicit* and finite dimensional kernel map. In contrast to usual kernels, such as the gaussian, the VC-dimension [1], related to a finite dimensional kernel map, is finite[1]. According to the Vapnik's VC-theory [14], the finiteness of the VC-dimension avoids loose generalization bounds and may guarantee better performance.

Now, we turn the problem into finding the hyperplane f as well as a Gram (kernel) matrix $\mathbf{K} = \mathbf{\Phi}'\mathbf{\Phi}$ where each column $\mathbf{\Phi}_i$ corresponds to an explicit mapping of $\mathbf{x}_i$ into a high dimensional space (i.e., $\phi(\mathbf{x}_i) = \mathbf{\Phi}_i$). This mapping is designed in order to i) guarantee linear separability of data in $\mathcal{S}$, ii) to ensure good generalization performance by maximizing the margin, iii) to approximate the input data, and also iv) to ensure positive definiteness of $\mathbf{K}$ by construction, i.e., without adding further constraints. This results into the following constrained minimization problem

$$\begin{array}{ll} \underset{\mathbf{B},\mathbf{\Phi},\mathbf{w}}{\min} & \tfrac{1}{2}\|\mathbf{w}\|_2^2 + \tfrac{1}{2}\|\mathbf{X} - \mathbf{B}\mathbf{\Phi}\|_F^2 \\ \text{s.t} & y_i \mathbf{w}'\mathbf{\Phi}_i \geq 1, \quad i = 1, \ldots, \ell, \end{array} \tag{2}$$

here $\|\mathbf{A}\|_F^2 = \mathbf{tr}(\mathbf{A}\mathbf{A}')$ stands for the square of the Frobenius norm and $\mathbf{X} \approx \mathbf{B}\mathbf{\Phi}$ is factorized using an overcomplete basis $\mathbf{B} \in \mathbb{R}^{n \times p}$ (i.e., $p > n$) and a new kernel map $\mathbf{\Phi} \in \mathbb{R}^{p \times m}$. Without a loss of generality b is omitted in the above expression as it can be induced from $\mathbf{w}$ and the mapping $\mathbf{\Phi}$.

By choosing the dimension p sufficiently large; for instance $\max(\ell, n) + 1$, one may show that inequality constraints can be satisfied and the right-hand side term tends to zero for an infinite number of solutions, so the above constrained minimization problem is guaranteed to have a solution; nevertheless it is under-conditioned.

2.2. Transductive Setting

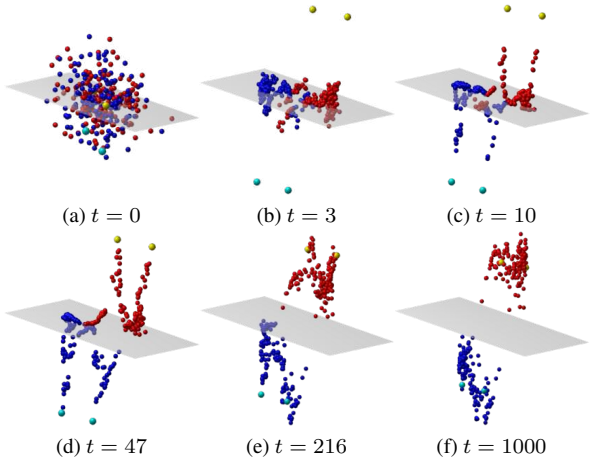
For a better conditioning of (2), we implement in this section the smoothness assumption discussed in Section 1. This makes it possible to design smooth kernel maps and to assign similar predictions to neighboring data for a better generalization on the unlabeled ones (see toy example in Fig. 1).

We model the input data $\mathcal{S}$ using an adjacency graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes $\mathcal{V} = \{v_1, \ldots, v_m\}$ correspond to samples $\{\mathbf{x}_i\}$ and edges $\mathcal{E} = \{e_{ij}\}$ are the set of weighted links of $\mathcal{G}$. In the above definition, $\mathbf{x}_i \in \mathbb{R}^n$ is a feature vector (color, texture, etc.) while $e_{ij} = (v_i, v_j, \mathbf{A}_{ij})$ defines a connection between v_i, v_j weighted by $\mathbf{A}_{ij}$. The latter is defined as $\mathbf{A}_{ij} = 1_{\{v_j \in \mathcal{N}_k(v_i)\}} \cdot \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / \sigma^2\right)$, here the neighborhood $\mathcal{N}_k(v_i)$ of a given node v_i, includes the set of the k-nearest neighbors of v_i. Notice that this neighborhood system is designed in order to guarantee that $\forall v_i, v_j \in \mathcal{V}$, $v_j \in \mathcal{N}_k(v_i)$ implies $v_i \in \mathcal{N}_k(v_j)$ and vice-versa.

Considering $f(\mathbf{x}_i) = \mathbf{w}'\mathbf{\Phi}_i$ and $f(\mathbf{x}_j) = \mathbf{w}'\mathbf{\Phi}_j$, we define our regularizer as $\frac{\gamma_s}{4}\sum_{i,j=1}^{m}\left(\mathbf{w}'\mathbf{\Phi}_i - \mathbf{w}'\mathbf{\Phi}_j\right)^2 \mathbf{A}_{ij}$, which may be rewritten as $\frac{\gamma_s}{2}\mathbf{w}'\mathbf{\Phi}\mathbf{L}\mathbf{\Phi}'\mathbf{w}$, here $\gamma_s \geq 0$ and $\mathbf{L}$ is the graph Laplacian defined by $\mathbf{L} = \mathbf{D} - \mathbf{A}$ and $\mathbf{D} = \text{diag}(\mathbf{A}\mathbf{1})$ where $\mathbf{1}$ is all-ones vector. When adding this regularizer in objective function (2) and replacing inequality constraints with the squared loss $\frac{\gamma_c}{2}(\mathbf{Y} - \mathbf{\Phi}'\mathbf{w})'\mathbf{C}(\mathbf{Y} - \mathbf{\Phi}'\mathbf{w})$, we obtain the complete form of our transductive learning problem

$$\underset{\mathbf{B},\mathbf{\Phi},\mathbf{w}}{\min} \quad \tfrac{1}{2}\mathbf{w}'\left(\mathbf{I} + \mathbf{\Phi}\tilde{\mathbf{L}}\mathbf{\Phi}'\right)\mathbf{w} + \tfrac{1}{2}\|\mathbf{X} - \mathbf{B}\mathbf{\Phi}\|_F^2 - \gamma_c \mathbf{Y}'\mathbf{C}\mathbf{\Phi}'\mathbf{w}, \tag{3}$$

where $\mathbf{I}$ is $p \times p$ identity matrix, $\tilde{\mathbf{L}} = (\gamma_c\mathbf{C} + \gamma_s\mathbf{L})$, $\mathbf{C}$ is the diagonal $m \times m$ matrix for which the i^{th} diagonal element is fixed to 1 for a labeled sample and 0 for an unlabeled one, $\mathbf{Y}$ is the m-dimensional vector for which the i-th element is y_i for a labeled data and 0 for an unlabeled one.



(a) $t = 0$ (b) $t = 3$ (c) $t = 10$

(d) $t = 47$ (e) $t = 216$ (f) $t = 1000$

Fig. 1: This figure shows the evolution of the learned kernel map through different iterations of our method (see Algorithm 1). This map is found for the popular "two moon" example in [11]. The underlying 2D input data are not linearly separable, while the learned kernel map makes them linearly separable in a 3D space. In these experiments, only $\ell = 4$ samples were labeled (shown in blue and yellow resp. for the positive and the negative classes).

3. OPTIMIZATION

It is clear that the constrained minimization problem in (3) is not convex jointly w.r.t $\mathbf{B}, \mathbf{\Phi}, \mathbf{w}$. We consider an alternating optimization procedure by solving three subproblems: we first maximize the margin $2/\|\mathbf{w}\|^2$ w.r.t $\mathbf{w}$, then we minimize the regularization criterion and the reconstruction error w.r.t $\mathbf{\Phi}$ and finally, we update the basis $\mathbf{B}$. This process is repeated until convergence; i.e., all the unknowns remain unchanged from one iteration to another. Different steps of the algorithm are shown in Algorithm (1); the superscript (t) is added to $\mathbf{w}$, $\mathbf{B}$ and $\mathbf{\Phi}$ in order to show the evolution of their

[1]The VC-dimension is the maximum number of data samples, that can be shattered, whatever their labels.

Algorithm 1 TransRMF

Input: labeled $\{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ and unlabeled data $\{\mathbf{x}_i\}_{i=\ell+1}^{m}$

Initialization: set the adjacency matrix $\mathbf{A}$, $t \leftarrow 0$ and set $\mathbf{\Phi}^{(0)}$, $\mathbf{B}^{(0)}$ to random full rank matrices.

Repeat steps (1+2) until convergence

1. Update $\mathbf{w}^{(t+1)}$ and $\mathbf{B}^{(t+1)}$ using (4), (7) resp.
2. Update $\mathbf{\Phi}^{(t+1)}$ iteratively using (8).

Output: kernel maps $\{\mathbf{\Phi}_i\}$ and labels $\{y_i\}$ with $y_i = \mathbf{w}'\mathbf{\Phi}_i$.

values through different iterations of the learning process.

Updating basis and classifier. considering iteration $t+1$ of the optimization algorithm, we assume fixed $\mathbf{B}^{(t)}$ and $\mathbf{\Phi}^{(t)}$. Now enforcing the gradient of (3) to vanish (w.r.t $\mathbf{w}$) leads to

$$\mathbf{w}^{(t+1)} = \gamma_c \left(\mathbf{I} + \mathbf{\Phi}^{(t)}\tilde{\mathbf{L}}\mathbf{\Phi}^{(t)'}\right)^{-1} \mathbf{Y}'\mathbf{C}\mathbf{\Phi}^{(t)'}. \tag{4}$$

Similarly, assuming $\mathbf{\Phi}^{(t)}$ fixed, we find $\mathbf{B}^{(t+1)}$ as

$$\underset{\mathbf{B}}{\operatorname{argmin}} \frac{1}{2} \left\| \mathbf{X} - \mathbf{B}\mathbf{\Phi}^{(t)} \right\|_F^2. \tag{5}$$

Expressing the Frobenius norm using the trace operator

$$\begin{aligned} & \tfrac{1}{2} \left\| \mathbf{X} - \mathbf{B}\mathbf{\Phi}^{(t)} \right\|_F^2 \\ = \; & \tfrac{1}{2}\mathrm{tr}\left(\left(\mathbf{X} - \mathbf{B}\mathbf{\Phi}^{(t)}\right) \left(\mathbf{X} - \mathbf{B}\mathbf{\Phi}^{(t)}\right)' \right), \end{aligned} \tag{6}$$

and enforcing the gradient of (6) to vanish (w.r.t $\mathbf{B}$) leads to

$$\mathbf{B}^{(t+1)} \;=\; \mathbf{X}\mathbf{\Phi}^{(t)'} \left(\mathbf{\Phi}^{(t)}\mathbf{\Phi}^{(t)'}\right)^{-1}. \tag{7}$$

Learning kernel map. considering fixed $\mathbf{B}$ and $\mathbf{w}$, (3) is solved w.r.t $\{\mathbf{\Phi_i}\}$ by fixing the m kernel maps $\mathbf{\Phi}^{(t)}$ at step t and finding the new maps $\mathbf{\Phi}^{(t+1)}$. Following this reasoning, conditions for optimality lead to a solution $\mathbf{\Phi}^{(t+1)} = \tilde{\mathbf{\Psi}}$ as the limit of

$$\begin{aligned} \mathbf{\Psi}_i^{(k+1)} = \; & \left(\mathbf{B}'\mathbf{B} + (\gamma_s \mathbf{D}_{ii} + \gamma_c \mathbf{C}_{ii})\mathbf{w}\mathbf{w}'\right)^{-1} \\ & \cdot \left[\mathbf{B}'\mathbf{X} + \gamma_c \mathbf{w}\mathbf{Y}'\mathbf{C} + \gamma_s \mathbf{w}\mathbf{w}'\mathbf{\Psi}^{(k)}\mathbf{A}\right]_i, \end{aligned} \tag{8}$$

for $i = 1, \ldots, m$, with $\mathbf{\Psi}^{(0)} = \mathbf{\Phi}^{(t)}$ and $[.]_i$ stands for the i^{th} column of a matrix. Here we omit the superscript $(t+1)$ for variables $\mathbf{B}$ and $\mathbf{w}$ for brief. The process described in the above equation allows us to recursively diffuse the kernel maps from the labeled to the unlabeled data, through the neighborhood system defined in the graph $\mathcal{G}$. This process is iterative and may require many steps before convergence. The latter is reached when $\|\mathbf{\Phi}^{(k+1)} - \mathbf{\Phi}^{(k)}\| \leq \epsilon$; (in practice, $\epsilon = 10^{-2}$, and convergence usually happens in less than $k = 100$ iterations, see Fig. 2, bottom-right).

4. EXPERIMENTS

4.1. Database and Setting

We use the Pascal VOC 2011 dataset[2] in order to evaluate the performance of our transductive inference method on object class segmentation (OCS). For that purpose, we use 556 images from this dataset belonging to 21 categories; given an image, the goal is to assign each group of pixels (referred to as superpixel) to one of these 21 categories. In practice, a given image is subdivided into an irregular grid (neighborhood system) of 700 superpixels, each one is processed in order to extract various features [15] including texture (texton histograms), color (3D RGB mean, standard deviation and 3D RGB histograms) and bag-of-word SIFT descriptors, resulting into a final feature vector of 305 dimensions.

For each image in VOC, we turn OCS into a transductive inference problem where only a small fraction of its underlying superpixels is labeled (see Fig. 3, third column). We train one transductive classifier (referred to as TransRMF) for each category and we combine these classifiers using the "winner-take-all" strategy in order to infer the category of a given unlabeled superpixel.

Following the evaluation protocol of Pascal VOC 2011, we use the standard segmentation accuracy for assessment. This measure is defined for each category $\mathcal{C}$ using intersection/union score, defined as the number of correctly labeled pixels of $\mathcal{C}$, divided by the number of pixels labeled with that category into different images and the ground truth. This accuracy is expressed as

$$\text{accuracy} = \frac{\text{true pos.}}{\text{true pos.} + \text{false pos.} + \text{false neg.}} \tag{9}$$

A mean accuracy is also considered as the expectation through different categories.

4.2. Performance and Comparison

Fig. (2, top-left) reports average accuracy for different values of the regularization parameter γ_s; note that $\gamma_s = 0$ corresponds to the *baseline* inductive setting (i.e., no regularization is applied). Fig. 3, shows the evolution of the underlying segmentation results w.r.t γ_s. According to these results, an underestimated γ_s results into noisy segmentation while an overestimated γ_s makes the segmentation results very smooth (with possibly lost details). It is also clear that the transductive setting of our method (i.e., $\gamma_s > 0$) outperforms the inductive one (i.e., $\gamma_s \to 0$). Fig. 2 (top-right) also reports the average accuracy as an increasing function of γ_c (almost quasi-constant for larger values of γ_c). According to these experiments, we found that the best performance are achieved when $\gamma_s = 0.1$ and $\gamma_c = 10$.

Finally, we compare our TransRMF approach w.r.t to inductive as well as transductive approaches. Fig. (2, middle) shows the average accuracies and comparison. Inductive approaches include SVM classifiers with different kernels (linear, RBF, χ^2, and histogram intersection) and their combination using multiple kernel learning[3] (Fig. 2, bottom-left). Transductive approaches include Laplacian-SVM[4] and transductive SVM[5]. Our TransRMF method shows a clear and a consistent gain w.r.t both inductive and transductive methods as well as multiple kernel learning (see also Fig. 4).

5. CONCLUSION

We introduced in this paper, a new transductive learning approach for kernel design and classification. The strength of our contribution resides in the variational framework that allows us to explicitly design an optimal kernel map as a part of the learning process. When compared to baseline inductive methods, multiple kernel learning and also transductive methods, our approach shows superior accuracy on the challenging object class segmentation task.

As a future extension of this work, we will investigate the application of this method to other tasks including interactive image retrieval.

[2] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2011/index.html

[3] http://asi.insa-rouen.fr/enseignants/ arakotom/code/mklindex.html

[4] http://www.dii.unisi.it/ melacci/lapsvmp/

[5] http://svmlight.joachims.org/

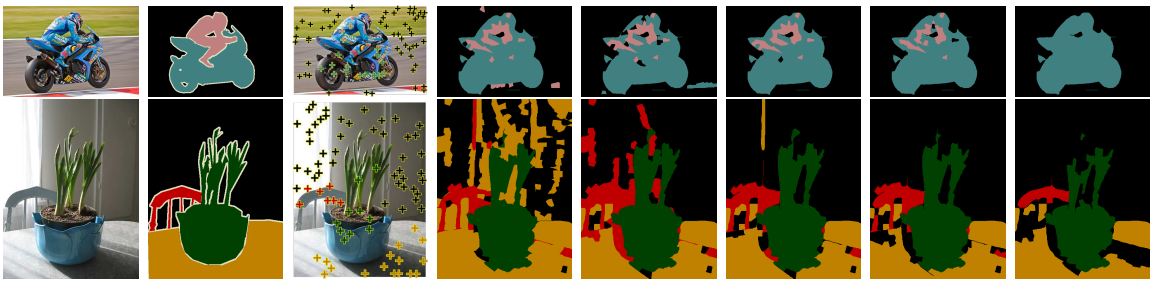


Fig. 3: **Left to right:** original image; ground truth; labeled data; segmentation results with $\gamma_s = 10^{-3}, 10^{-2}, 10^{-1}, 1, 10$.

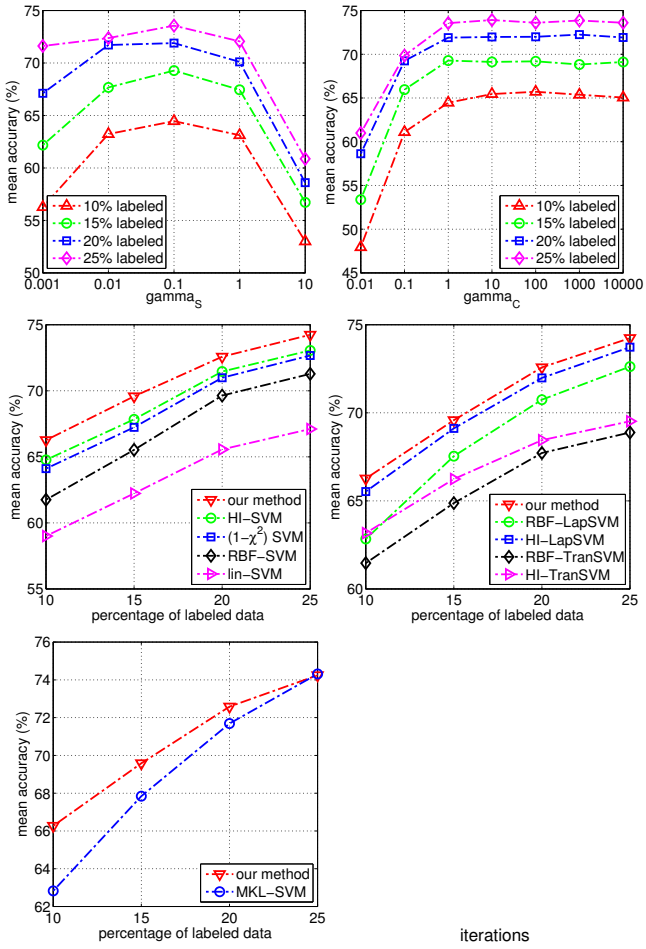


Fig. 2: Diagrams in top show the evolution of the average accuracy w.r.t the regularization term (i.e. γ_s) and the fidelity term (i.e. γ_c) respectively. Diagrams in middle and (bottom, left) show comparison of TransRMF w.r.t inductive, transductive and MKL learning respectively. Note that all the performances are shown for different percentages of labeled data. Figure in (bottom, right) illustrates the convergence process, i.e., the difference between current and previous estimate of kernel maps through different iterations.

Fig. 4: **1st col:** original image; **2nd col:** ground truth; **3rd col:** TransRMF; **4th col:** MKLSVM; **5th col:** LapSVM.

References

[1] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[2] B. Schölkopf and AJ. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, Dec. 2001.

[3] S. Maji, A-C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *CVPR*, 2008.

[4] T. Joachims, *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms*, Kluwer/Springer, 2002.

[5] Asa, Ong Cheng Soon, Sonnenburg Sören, Schölkopf Bernhard, and Rätsch Gunnar Ben-Hur, "Support vector machines and kernels for computational biology," *PLoS Comput Biol*, vol. 4, no. 10, pp. e1000173, 10 2008.

[6] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *JMLR*, vol. 9, pp. 2491–2521, 2008.

[7] G-R Lanckriet, P. Bartlett, and M. Jordan, "Learning the Kernel Matrix with Semidefinite Programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[8] M. Varma and D. Ray, "Learning The Discriminative Power-Invariance Trade-Off," *2007 IEEE 11th ICCV*, pp. 1–8, 2007.

[9] O. Duchenne, J.-Y. Audibert, R. Keriven, J. Ponce, and F. Segonne, "Segmentation by transduction," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[10] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*, MIT Press, Cambridge, MA, 2006.

[11] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, December 2006.

[12] M. Seeger, "Learning with labeled and unlabeled data," *Technical Report, University of Edinburgh*, 2001.

[13] T. Joachims, "Transductive inference for text classification using support vector machines," in *ICML*, 1999, pp. 200–209.

[14] V. Vapnik and A. Sterin, "On structural risk minimization or overall risk in a problem of pattern recognition," *Automation and Remote Control*, vol. 10, no. 3, pp. 1495–1503, 1977.

[15] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *ECCV (5)*, 2010, pp. 352–365.