Transductive Kernel Map Learning and its Application to Image Annotation

Dinh-Phong Vo
vo@telecom-paristech.fr
Hichem Sahbi
sahbi@telecom-paristech.fr

LTCI CNRS Telecom ParisTech
46 rue Barrault, 75013, Paris, France

We introduce in this paper a novel image annotation approach based on maximum margin classification and a new class of kernels. The method goes beyond the naive use of existing kernels and their restricted combinations in order to design "model-free" transductive kernels applicable to interconnected image databases.

Let $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \ldots, \mathbf{x}_m\}$ denote an image database described in an n-dimensional input space. We assume that only the first l ($l \ll m$) vectors of $\mathcal{S}$ are labeled (a.k.a annotated), i.e., $\{\mathbf{y}_1, \ldots, \mathbf{y}_l\}$ are known; here $\mathbf{y}_i \in \{-1, +1\}^r$ and r is the number of possible labels used for annotation.

Our approach considers image annotation as a multi-label classification problem in which a sample $\mathbf{x}_i$ may have more than one label, i.e., $r > 1$, with $\mathbf{y}_{ik} = +1$ iff $\mathbf{x}_i$ has the k^{th} label and $\mathbf{y}_{ik} = -1$ otherwise. Our objective is to build an optimal *kernel map* and a decision criterion f in order to infer the unknown label vectors $\{\mathbf{y}_{l+1}, \ldots, \mathbf{y}_m\}$.

We adopt the max-margin classification [4] approach in order to learn a classifier $f(\mathbf{x}_i) = \mathbf{W}'\phi(\mathbf{x}_i)$ that balances training error and model complexity. This classifier corresponds to

$$\underset{f}{\operatorname{argmin}}\ \mathcal{R}(f) + \gamma_c \sum_{i=1}^{l} \ell(f(\mathbf{x}_i), \mathbf{y}_i), \tag{1}$$

where $\mathcal{R}$ is a regularizer, $\ell(f(\mathbf{x}_i), \mathbf{y}_i)$ is the loss associated with a prediction $f(\mathbf{x}_i)$ when the true output is $\mathbf{y}_i$ and $\gamma_c > 0$ balances these two terms. For nonlinear classification, ϕ maps the input data (in $\mathcal{S}$) into a high dimensional space $\mathcal{H}$ such that $\mathbf{W}$ can separate labeled data $\{\mathbf{x}_i\}_{i=1}^{l}$.

Following the kernel trick [3], the function f may also be expressed as a linear combination of symmetric, continuous and positive (semi) definite kernel functions. A kernel (denoted κ) is defined on two samples $\mathbf{x}_i, \mathbf{x}_j$ as $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. The closed form of $\kappa(\cdot, \cdot)$ may also be defined among a collection of existing kernels including linear, polynomial and histogram intersection; but the underlying mapping $\phi(.) \in \mathcal{H}$ is usually *implicit*, i.e., it does exist but it is not necessarily known and may be infinite dimensional.

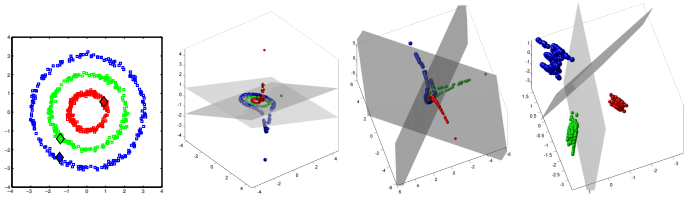
Our proposed method, in contrast to usual kernel methods, finds an *explicit* and finite dimensional kernel map. According to Vapnik's VC-theory [4], a finite dimensional kernel map, with a bounded related VC-dimension, avoids loose generalization bounds and may guarantee better performance.

Our goal is to find hyperplane parameters $\mathbf{W}$ as well as a Gram (kernel) matrix $\mathbf{K} = \Phi'\Phi$ where each column Φ_i corresponds to an explicit mapping of $\mathbf{x}_i$ into a high dimensional space (i.e., $\phi(\mathbf{x}_i) = \Phi_i$). The learned mapping Φ must i) guarantee linear separability of data in $\mathcal{S}$, ii) ensure good generalization performance by maximizing the margin, iii) approximate the input data, and also iv) ensure positive definiteness of $\mathbf{K}$ by construction, i.e., without adding further constraints. Considering $\mathcal{R}(f) = \|\mathbf{W}\|_F^2$ and $\ell(f(\mathbf{x}_i), \mathbf{y}_i) = \|f(\mathbf{x}_i) - \mathbf{y}_i\|_2^2$, the map Φ and the classifier parameters $\mathbf{W}$ are found by solving

$$\begin{array}{ll} \min\limits_{\mathbf{B},\Phi,\mathbf{W}} & \frac{\mu}{2}\|\Phi\|_F^2 + \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{\gamma_c}{2}\left\| \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{B} & \mathbf{0}_{n\times p} \\ \mathbf{0}_{r\times p} & \mathbf{W}' \end{bmatrix} \begin{bmatrix} \Phi \\ \Phi\mathbf{C} \end{bmatrix} \right\|_F^2 \\ \text{s.t} & \|\mathbf{B}_i\|_2^2 = 1, \forall i = 1, \ldots, p \end{array} \tag{2}$$

here $\mathbf{C} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with $\mathbf{C}_{ii} = 1_{\{1\le i\le l\}}$, $\mathbf{0}_{n\times p}$ and $\mathbf{0}_{r\times p}$ are $n \times p$ and $r \times p$ zeros matrices respectively, $\mathbf{X} \approx \mathbf{B}\Phi$ is factorized using an overcomplete basis $\mathbf{B} \in \mathbb{R}^{n\times p}$ (i.e., $p > n$) and a new kernel map $\Phi \in \mathbb{R}^{p\times m}$.

According to [4], the VC-dimension (related to a family of classifiers) depends also on the dimension of the learned kernel map and this may affect generalization, especially if this dimension is very high. Since the actual (intrinsic) dimension of the learned kernel map Φ is unknown, we choose the number of basis p to be sufficiently large such that the factorization term (in right-hand side of Eq. 2) tends to zero for an infinite number of solutions. Then, the actual (intrinsic) dimension is found by regularizing Eq. 2 using the Frobenius norm $\|\Phi\|_F^2$ which has similar effect as the nuclear norm where $\mu \ge 0$ controls the rank of $\mathbf{K}$.



(a) input data (b) $\gamma_s = 10, \gamma_c = 1$ (c) $\gamma_s = 10^2, \gamma_c = 1$ (d) $\gamma_s = 10^3, \gamma_c = 10^2$

Figure 1: (a) This figure shows the input data where different colors stand for different classes; red-colored data are annotated with $(1\ \ -1)'$, blue-colored data with $(-1\ 1)'$ and green-colored data with $(1\ \ 1)'$. Note that just one training sample per class (diamond-shaped) is labeled while others are unlabeled. Figures (b,c,d) are the learned kernel maps (shown in 3d) and the obtained decision hyperplanes for different setting of the parameters γ_c and γ_s.

For a better conditioning of Eq. 2, we adopt transductive inference [1, 5] which assumes that close data in a high-density area of the input space should have similar labels [2]. This assumption, therefore, enables label diffusion from training to the test data (see toy example in Fig. 1).

By representing $\mathbf{x}_i$'s as vertices $\{v_i\}$ and their pairwise similarities as edges $\{e_{ij}\}$, the smoothness assumption between v_i and v_j is modeled by the differences between $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$, i.e.,

$$\frac{1}{4}\sum_{i=1,j=1}^{m} \|\mathbf{W}'\Phi_i - \mathbf{W}'\Phi_j\|^2 \mathbf{A}_{ij} \Longleftrightarrow \frac{1}{2}\operatorname{tr}\left(\mathbf{W}'\Phi\mathbf{L}\Phi'\mathbf{W}\right), \tag{3}$$

where the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is defined by the affinity matrix $\mathbf{A}$ whose elements $\mathbf{A}_{ij} = 1_{\{v_j \in \mathcal{N}_k(v_i)\}} \cdot s(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{D} = \operatorname{diag}(\mathbf{A1})$ with $\mathbf{1}$ being the all-one vector of length m. Here $s(\cdot, \cdot)$ is a visual similarity and $\mathcal{N}_k(v_i)$ is the set of the k-nearest neighbors of v_i.

Now, we obtain the complete form of our transductive learning problem as

$$\begin{array}{ll} \min\limits_{\mathbf{B},\Phi,\mathbf{W}} & \frac{\mu}{2}\|\Phi\|_F^2 + \frac{1}{2}\operatorname{tr}\left(\mathbf{W}'\left(\mathbf{I}_p + \gamma_s\Phi\mathbf{L}\Phi'\right)\mathbf{W}\right) + \\ & \quad + \frac{\gamma_c}{2}\left\| \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{B} & \mathbf{0}_{n\times p} \\ \mathbf{0}_{r\times p} & \mathbf{W}' \end{bmatrix} \begin{bmatrix} \Phi \\ \Phi\mathbf{C} \end{bmatrix} \right\|_F^2, \\ \text{s.t} & \|\mathbf{B}_i\|_2^2 = 1, \forall i = 1, \ldots, p \end{array} \tag{4}$$

with $\mathbf{I}_p$ the $p \times p$ identity matrix and again $\mathbf{C}$ 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.

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 (see Fig. 1). Experiments conducted on image annotation, show that, indeed, our obtained kernel achieves at least comparable results with related state of the art methods on the MSRC and the Corel5k databases.

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

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

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

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

[5] V. Vapnik and A. Sterin. On structural risk minimization or overall risk in a problem of pattern recognition. *Automation and Remote Control*, 10(3):1495–1503, 1977.