Online Learning with Knowledge-Based SVMs

Gautam Kunapuli[1], Kristin P. Bennett[2], Richard Maclin[3] and Jude Shavlik[4]

[1]Department of Biostat. & Med. Informatics
University of Wisconsin-Madison
1300 University Avenue Madison WI 53706
kunapuli@wisc.edu

[2]Department of Mathematical Sciences
Rensselaer Polytechnic Institute
110 8th Street Troy NY 12180
bennek@rpi.edu

[3]Department of Computer Science
University of Minnesota-Duluth
1114 Kirby Drive Duluth MN 55812
rmaclin@d.umn.edu

[4]Department of Computer Science
University of Wisconsin-Madison
1210 W Dayton Street Madison WI 53706
shavlik@cs.wisc.edu

We propose a novel approach for incorporating prior knowledge into online learning algorithms. We focus on the supervised binary classification problem via SVMs in an online setting. Here, the learning algorithm receives data points sequentially rather that all at once (as in batch learning) and predicts the label (± 1) at each round. After the prediction, the algorithm receives feedback indicating the correct label, using which the decision function may be updated. The goal is to successively update the decision function taking into account *prior knowledge* in the form of soft polyhedral advice so as to make increasingly accurate predictions on subsequent rounds. The advice helps speed up and bias learning so that better generalization can be obtained with less data.

An existing advice-taking approach, when prior knowledge is in the form of *polyhedral knowledge sets* in the input space of data, is via *knowledge-based support vector machines* (KBSVMs) [2]. The knowledge sets typically characterize an area of input space as belonging to one of the two classes and can be interpreted as a generalization of training examples from points in feature space to regions. More concretely, if we introduce *labels for advice*: $z = \pm 1$, knowledge sets are specified as $D\mathbf{x} \leq \mathbf{d} \Rightarrow z(\mathbf{w}'\mathbf{x} - \gamma) \geq 1$, where the implication means that every point in the polyhedron $D\mathbf{x} \leq \mathbf{d}$ lies either above $\mathbf{w}'\mathbf{x} = b + 1$ (when label is specified as $z = 1$) or below $\mathbf{w}'\mathbf{x} = b - 1$ (when $z = -1$). The advice is in the form of a *logical implication* and cannot be added directly into the SVM formulation. We can use *theorems of the alternative* (see [3]) to derive the following constraints instead, where prior knowledge is characterized by the *knowledge variables* $\mathbf{u}$:

$$D'\mathbf{u} + z\,\mathbf{w} = 0, \;\; -d'\mathbf{u} - z\,b \geq 1, \;\; \mathbf{u} \geq 0.$$

Each knowledge set introduces its own set of constraints which are coupled to the main SVM by their dependence on $(\mathbf{w}, b)$. Soft advice is allowed by relaxing the advice constraints with slack variables.

We adopt the online formalism of *passive-aggressive algorithms* [1]. Given a loss function, the algorithm is *passive* whenever the loss is zero i.e., the data point at the current round t is correctly classified. If misclassified, the algorithm updates the weight vector ($\mathbf{w}^t$) *aggressively*, such that the loss is minimized over the new weights ($\mathbf{w}^{t+1}$). The update rule that achieves this is derived as the optimal solution to a constrained optimization problem comprising two terms: a loss function, and a *proximal term* that requires $\mathbf{w}^{t+1}$ to be as close as possible to $\mathbf{w}^t$. There are several advantages over other approaches: first, it is possible to derive closed-form solutions and consequently, simple update rules. Second, it is possible to formally derive relative loss bounds where the loss suffered by the algorithm is compared to the loss suffered by some arbitrary, fixed hypothesis. Finally, the optimization problem formulations are analogous to their batch versions.

In the online advice algorithm, at round t, given labeled data $(\mathbf{x}^t, y_t)$, and m labeled knowledge sets $(D_i, \mathbf{d}^i, z_i)$, we can propose a general constrained optimization problem that computes

the new weight vector $\mathbf{w}^{t+1}$ and the new advice vectors $\mathbf{u}^{i,t+1}$ from $\mathbf{w}^t$ and $\mathbf{u}^{i,t}$ as

$$
\begin{aligned}
(\mathbf{w}^{t+1}, \mathbf{u}^{i,t+1}) = \underset{\mathbf{w},\mathbf{u}^i,\xi,\boldsymbol{\eta}^i,\zeta_i}{\arg\min} \quad & \frac{1}{2}\|\mathbf{w}-\mathbf{w}^t\|_2^2 + \frac{1}{2}\sum_{i=1}^{m}\|\mathbf{u}^i-\mathbf{u}^{i,t}\|_2^2 + \frac{\lambda}{2}\,\xi^2 + \frac{\mu}{2}\sum_{i=1}^{m}\left(\|\boldsymbol{\eta}^i\|_2^2+\zeta_i^2\right) \\
\text{subject to} \quad & y_t\mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, \\
& \left.\begin{aligned} D_i'\mathbf{u}^i + z_i\mathbf{w} + \boldsymbol{\eta}^i &= 0 \\ -\mathbf{d}^{i\prime}\mathbf{u}^i - 1 + \zeta_i &\geq 0 \\ \mathbf{u}^i &\geq 0 \end{aligned}\right\} \; i = 1,\ldots,m.
\end{aligned}
$$

The algorithm is initialized with $\mathbf{w}^1 = 0$, and $\mathbf{u}^{i,1}$ being learned from a KBSVM that depends on the knowledge sets *alone*. The variables $\boldsymbol{\eta}^i$ and ζ_i are slack variables that relax the hard advice constraints. The objective function consists of proximal terms that ensure that the new weight and knowledge vectors are close to the previous updates; the loss with respect to the data and advice respectively is also minimized. This problem does not have a closed-form solution owing to the complicating constraints $\mathbf{u}^i \geq 0$. This is because these constraints give rise to complementarity conditions that need to be satisfied for the optimality problem; these conditions introduce a combinatorial complexity into the problem addressable only by iterative approaches. If we assume that the advice is reliable, then there is no *advice refinement* and advice variables are fixed: $\mathbf{u}^{i,t} = \mathbf{u}^{i,1}$. This gives the simpler problem (dropping the index t from $\mathbf{u}^{i,t}$, as they are fixed):

$$
\begin{aligned}
\mathbf{w}^{t+1} \;=\; \underset{\mathbf{w},\xi}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}-\mathbf{w}^t\|_2^2 + \frac{\lambda}{2}\,\xi^2 + \frac{\mu}{2}\sum_{i=1}^{m}\|\boldsymbol{\eta}^i\|_2^2 \\
\text{subject to} \quad & y_t\mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, && \textit{(data constraint)} \\
& D_i'\mathbf{u}^i + z_i\mathbf{w} + \boldsymbol{\eta}^i = 0, \quad i = 1,\ldots,m. && \textit{(advice constraints)}
\end{aligned}
$$

Introducing multipliers α and $\boldsymbol{\beta}^i$ for the data and advice constraints respectively, we can express the update rule through the closed-form solution

$$
\mathbf{w}^{t+1} = \mathbf{w}^t + \alpha y_t \mathbf{x}^t + \sum_{i=1}^{m} z_i \boldsymbol{\beta}^i.
$$

We denote $\mathbf{r}^{i,t} = -z_i D_i'\mathbf{u}^i$, which represents information about each advice set as a *point in the hypothesis space*. The centroid of these points, the *average advice*, is $\mathbf{r}^t = \frac{1}{m}\sum_{i=1}^{m}\mathbf{r}^{i,t}$. The update can be computed using[1]

$$
\alpha = \frac{\left(\,1 - \nu\, y_t\mathbf{w}^{t\prime}\mathbf{x}^t - (1-\nu)\, y_t\mathbf{r}^{t\prime}\mathbf{x}^t\,\right)_+}{\dfrac{1}{\lambda} + \nu\|\mathbf{x}^t\|_2^2}, \qquad \frac{z_i\boldsymbol{\beta}^i}{\mu} = \mathbf{r}^i - \frac{\mathbf{w}^t + \alpha_\star\,\lambda\, y_t\mathbf{x}^t + m\mu\,\mathbf{r}^t}{\dfrac{1}{\nu} + \alpha_\star\,\lambda\|\mathbf{x}^t\|_2^2}.
$$

Writing $\nu = 1/(1+m\mu)$, we note that the numerator of α represents the loss function, which depends on both the previous iterate *and* the advice. For any $\mu > 0$, we have $\nu \in [0,1]$. Consequently, the overall loss can be viewed as the hinge loss applied to a *convex combination* of error according to the advice and the current hypothesis. If there is no advice then $\nu = 1$ and the update above becomes exactly identical to online passive-aggressive update derived in [1].

Preliminary experimental results indicate that the approach is promising when compared to online passive-aggressive algorithms. The approach can also be extended to an alternating scheme to update weight and knowledge variables. Further analysis of loss bounds is currently work in progress. Future research directions include incorporation of kernels to extend the current work from the linear to nonlinearly separable data sets.

[1] a_+ denotes componentwise $\max(a,0)$ and $a_\star$ denotes the componentwise Heaviside step function, $step(a)$

References

[1] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

[2] Glenn Fung, Olvi L. Mangasarian, and Jude W. Shavlik. Knowledge-based support vector classifiers. In S. Becker, Sebastian Thrun, and Keith Obermayer, editors, *Neural Information Processing Systems*, volume 15, MIT Press, Cambridge, MA, 2003.

[3] Olvi L. Mangasarian. *Nonlinear Programming.* McGraw-Hill, 1969.

Topic: learning algorithms
Preference: oral/poster
Author presenting: Gautam Kunapuli