Multiple Instance Learning with Generalized Support Vector Machines

Stuart Andrews, Thomas Hofmann and Ioannis Tsochantaridis
Department of Computer Science, Brown University
Providence, Rhode Island 02912, {stu,th,it}@cs.brown.edu

Multiple Instance Learning

In pattern classification it is usually assumed that a training set of labeled patterns is available. Multiple-Instance Learning (MIL) generalizes this problem setting by making weaker assumptions about the labeling information. While each pattern is still believed to possess a true label, training labels are associated with sets or *bags* of patterns rather than individual patterns.

More formally, given is a set of patterns $\mathbf{x}_1, ..., \mathbf{x}_n$ grouped into bags $X_1, ..., X_m$, with $X_j = \{\mathbf{x}_i : i \in I_j\}$ and $I_j \subseteq \{1, ..., n\}$. With each bag X_j is associated a label $Y_j \in \{-1, 1\}$. These labels are interpreted in the following way: if a bag has a negative label $Y_j = -1$, all patterns in that bag inherit the negative label. If on the other hand, $Y_j = 1$, then at least one pattern $\mathbf{x}_i \in X_j$ is a positive example of the underlying concept.

The MIL scenario has many interesting applications: One prominent application is the classification of molecules in the context of drug design (Dietterich, Lathrop, & Lozano-Perez 1997). Here, each molecule is represented by a bag of possible conformations. Another application is in image retrieval where images can be viewed as bags of local image patches (Maron & Ratan 1998) or image regions.

Algorithms for the MIL problem were first presented in (Dietterich, Lathrop, & Lozano-Perez 1997; Auer 1997; Long & Tan 1996). These methods (and analytical results) are based on hypothesis classes consisting of axis-aligned rectangles. Similarly, methods developed subsequently (e.g., (Maron & Lozano-Pérez 1998; Zhang & Goldman 2002)) have focused on specially tailored machine learning algorithms that do not compare favorably in the limiting case of bags of size 1 (the standard classification setting). A notable exception is (Ramon & Raedt 2000).

Generalized Support Vector Machines

We propose to generalize Support Vector Machines (SVMs) (Vapnik 1998) to take into account weak labeling information of the type found in MIL.

SVMs are based on the theory of linear classifiers, more precisely the idea of the *maximum margin hyperplane*. For linearly separable data, the maximum margin hyperplane is defined by parameters $\mathbf{w}^*, b^*$ with

$$(\mathbf{w}^*, b^*) = \underset{(\mathbf{w},b),\|\mathbf{w}\|=1}{\arg\max} \min_i \gamma_i, \quad \gamma_i \equiv y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) \quad (1)$$

The minimum $\gamma^* = \min_i \gamma_i$ is called the (geometric) *margin* and the patterns $\mathbf{x}_i$ with $\gamma_i = \gamma^*$ are called *support vectors*. The so-called soft-margin generalization of SVMs with L_1 penalties on margin violations amounts to solving the following convex quadratic program:

$$\begin{aligned} &\text{minimize} \quad H(w, b, \eta) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \eta_i \qquad (2) \\ &\text{s.t. } \forall i \qquad y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \eta_i, \quad \eta_i \geq 0 \end{aligned}$$

where the scalar C controls the trade-off between margin violation and regularization. What makes SVMs particularly powerful is the generalization to arbitrary kernel functions K. A kernel function implicitly maps patterns to a new high dimensional feature space in which an inner product is computed. Since this mapping needs not to be performed explicitly, this results in a very efficient non-linear classification algorithm.

To generalize SVMs for MIL, labels of patterns that only occur in positive bags are treated as unknown integer variables. Each bag with a positive label imposes an inequality constraint on the labels of the contained patterns; for negative bags, the pattern labels are known to be negative. These constraints can be incorporated in a generalized version of SVM learning as follows:

$$\begin{aligned} &\text{if} \quad Y_j = 1, \quad \text{then} \sum_{i \in I_j} \frac{1 + y_i}{2} \geq 1 \qquad (3) \\ &\text{if} \quad Y_j = -1, \quad \text{then } y_i = -1, \quad \forall i \in I_j \end{aligned}$$

The resulting problem, MIL-SVM, is a mixed integer program that bears some similarity to the transductive version of SVMs (Joachims 1999; Demirez & Bennett 2000). The goal is thus to minimize (2) jointly over the continuous parameters $(\mathbf{w}, b)$ and over the integer variables (labels of patterns in positive bags).

We propose a heuristic approach in order to find an approximation to this mixed integer program which cannot be solved exactly with current optimization methods for large problem sizes. After initializing all positive bag pattern labels to $+1$, one alternates solving the quadratic program in

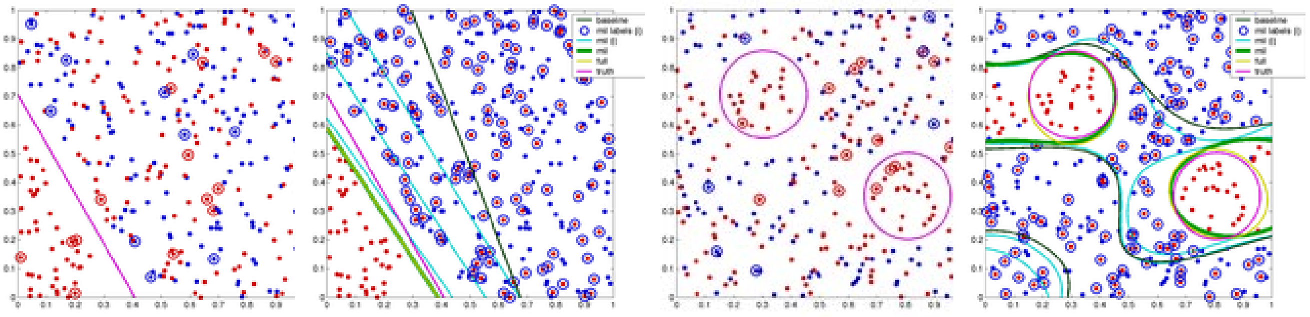

Figure 1: SVM (black), SVM with true labels (yellow), intermediate MIL-SVM (blue), MIL-SVM final (green) and correct (magenta) solutions on two synthetic data sets. Red and blue circles depict one positive bag and one negative bag on the left-hand images. On the right, blue circles indicate examples from positive bags that have been re-labeled $y_i = -1$.

(2) using the given labels, with a re-labeling step where the labels of patterns in positive bags are updated. Alternating these two steps defines a convergent procedure which will lead to a local optimum. Our current implementation swaps the label of a positive bag pattern that leads to the largest decrease in the objective (2) while not violating constraints in (3).

Results

We have experimentally verified the proposed generalization of SVMs on synthetic data, by comparing it with a naive baseline application of SVMs (labeling all patterns with the label of the bag they belong to) and with an optimal application of SVMs (labeling all patterns with the true concept label). A proof of concept on synthetic data is shown in Fig. 1 which shows that the MIL generalization of SVMs is able to identify superior discriminant functions, which is also reflected in a significantly reduced error rate.

A second preliminary series of experiments has been performed on a data set of 1000 images from the Corel image data base, preprocessed with the Blobworld system (Carson *et al.* 1999). In this representation, an image consists of a set of segments (or blobs), each characterized by color, texture and shape descriptors. Although standard SVMs already perform quite well, we have been able to achieve relative improvements in average precision in the range of 10% (e.g., from 25.3% to 30.3% for the “tiger” and from 43.2% to 46.0% for the “elephant” category). We are currently investigating ways to find better optimization heuristics and are conducting benchmark experiments on a larger scale.

Acknowledgments

Thanks to Chad Carson and the Blobworld team for making the Blobworld Matlab code publically available.

References

Auer, P. 1997. On learning from multi-instance examples: Empirical evaluation of a theoretical approach. In *Proc. 14th International Conference on Machine Learning*, 21–29. Morgan Kaufmann.

Carson, C.; Thomas, M.; Belongie, S.; Hellerstein, J. M.; and Malik, J. 1999. Blobworld: A system for region-based image indexing and retrieval. In *Third International Conference on Visual Information Systems*. Springer.

Demirez, A., and Bennett, K. 2000. Optimization approaches to semisupervised learning. In Ferris, M.; Mangasarian, O.; and Pang, J., eds., *Applications and Algorithms of Complementarity*. Kluwer Academic Publishers, Boston.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Perez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* 89(1-2):31–71.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *Proc. 16th International Conf. on Machine Learning*, 200–209. Morgan Kaufmann, San Francisco, CA.

Long, P., and Tan, L. 1996. PAC learning axis aligned rectangles with respect to product distributions from multiple-instance examples. In *Proceedings of the Conference on Computational Learning Theory*, 228–234.

Maron, O., and Lozano-Pérez, T. 1998. A framework for multiple-instance learning. In Jordan, M. I.; Kearns, M. J.; and Solla, S. A., eds., *Advances in Neural Information Processing Systems*, volume 10. The MIT Press.

Maron, O., and Ratan, A. L. 1998. Multiple-instance learning for natural scene classification. In *Proc. 15th International Conf. on Machine Learning*, 341–349. Morgan Kaufmann, San Francisco, CA.

Ramon, J., and Raedt, L. D. 2000. Multi instance neural networks. In *Proceedings of IMCL-2000 Workshop on Attribute-Value and Relational Learning.*

Vapnik, V. 1998. *Statistical Learning Theory*. Wiley.

Zhang, Q., and Goldman, S. A. 2002. EM-DD: An improved multiple-instance learning technique. In *Advances in Neural Information Processing Systems*.