Guest editorial: special issue on structured prediction

Charles Parker · Yasemin Altun · Prasad Tadepalli

Received: 8 September 2009 / Revised: 8 September 2009 / Accepted: 15 September 2009 /
Published online: 29 September 2009
The Author(s) 2009

1 Overview of this special issue

Structured Prediction or *Structured Classification* (Bakir et al. 2007) is the task of predicting a collection of related variables given some input. The relationship between the variables to be predicted is often complex. An example of such complex dependencies is machine translation, where the input is a sequence of words in the source natural language and the output is a sequence of words in the target natural language. Here, each word in the target language relates not only to the words in the source language, but also to the other (arbitrarily far) words in the target sequence.

As a field, structured prediction has some unique challenges, several of which are addressed by the papers in this issue. One of the most obvious of these is that the output spaces in question are often exponential in size, and so the complexity, both of these spaces and the learned models, can result in computationally infeasible learning and inference algorithms. In many cases, therefore, efficient optimization remains an open problem in structured prediction. Two papers in this special issue (Hsu et al. 2009; Sutton and McCallum 2009) address this problem. An important source of complexity in structured prediction algorithms is in the iterative nature of the training step: Often, training is done EM-style, where model parameters are estimated at each step, and then inference is performed based on these parameters. In models with complex structure, this inference step

C. Parker (✉)
Eastman Kodak Company, Rochester, NY, USA
e-mail: charles.parker1@kodak.com

Y. Altun
Max Planck Institute for Biological Cybernetics, Tübingen 72070, Germany
e-mail: altun@tuebingen.mpg.de

P. Tadepalli
Oregon State University, Corvallis, OR 97331, USA
e-mail: tadepall@eeecs.oregonstate.edu

can be prohibitively expensive. Furthermore, if the variables to be predicted have large cardinality, inference may not be feasible even if the model structure is tractable. These problems are addressed in the paper “Piecewise training for structured prediction” by Sutton and McCallum (2009). The authors propose two techniques. The first, called *piecewise training*, is an approximate training method in which complex graphs are broken into tractable *pieces*, which are then trained separately. The second, *piecewise pseudolikelihood* uses pseudolikelihood in combination with piecewise training to arrive at a training process that scales well with both variable cardinality and model complexity. An analysis of the techniques on a real-world natural language processing data set shows significant speed improvements over traditional training via maximum likelihood, and favorable comparisons with other methods of approximate training.

An alternative approach is to improve convex optimization methods that are commonly employed by structured prediction methods. For general convex optimization problems, Newton methods are known to converge faster than gradient based methods. Unfortunately, computing the Hessian for loss functions on structured objects involves correlations between labels in different cliques, which is prohibitively expensive in most cases. The paper, “Periodic step-size adaptation in second-order gradient descent for single-pass online structured learning” by Hsu et al. (2009) proposes a second-order stochastic gradient descent method in an online setting. The method approximates the Hessian by exploring a linear relation between the Hessian and the Jacobian such that the computation can be performed very efficiently for each online update. Comparative experimental analysis of the method on Conditional Random Fields (CRFs) show the validity of the approach.

In addition to the problem of efficient optimization, there are several difficulties that have been dealt with effectively in the general machine learning literature, such as noisy training data and the incorporation of prior knowledge, but pose special difficulties in the case of structured prediction. As to the latter of these, it is well-known that learning a model from finite samples is an ill-posed problem. While it is possible to overcome this by imposing general smoothness assumptions, e.g. via regularization terms, prior knowledge can be more effective. Hence, incorporating prior knowledge into learning is an important problem of machine learning and has been studied in depth for classification (e.g. Lauer and Bloch 2008 and the references thereof). Surprisingly, prior knowledge has not received much attention in the context of Structured Prediction. The paper “Generalized isotonic conditional random fields” by Mao and Lebanon (2009) is one of the few papers addressing this issue. The authors propose a novel approach for incorporating domain knowledge when training CRFs (Lafferty et al. 2001), by imposing generalized isotonic constraints on model parameters. These constraints are formulated as a taxonomy of probability ranks and/or ratios based on expert knowledge. The authors derive a parametric representation of the problem and provide an efficient algorithm. Experiments on natural language processing problems, such as sentiment analysis and text summarization, indicate that incorporating prior knowledge via this method consistently improves the performance over standard CRFs, whether the domain knowledge is gathered from an expert or from auxiliary data.

Label noise is also an important issue in machine learning that has received little attention in the structure prediction literature. In many current structured prediction algorithms, the presence of complete and correct training data is assumed. It is often further assumed that the correct predictions given in the training data are the only correct predictions for the given input, and all others are not desirable. The paper “Structured prediction by joint kernel support estimation” by Lampert and Blaschko (2009) presents a method for learning the support of the joint distribution over inputs and outputs using the one-class support vector machine. Because the distribution learned is generative in nature, it is robust to noisy training data and to the case where more than one prediction for a given input is correct. The

method also has a significant speed advantage over other methods of structured prediction as there is no inference step in the learning algorithm. The authors compare this method to other methods of structured prediction analytically. Experimental results using the task of object localization on a benchmark dataset show that the predictors obtained using their method are far more robust to noise than SVMStruct (Tsochantaridis et al. 2005) on this task using the same kernel. The results also show that the speed of the training method allows the incorporation of a kernel (the χ^2 kernel) that would be infeasible under the iterative training procedure used in SVMStruct. With the improved kernel, the authors show significant improvement on a precision/recall basis.

Finally, the relative youth of structured prediction as a machine learning subdiscipline means that many important connections between the structured prediction literature and general machine learning have yet to be made. This special issue presents two papers that build bridges between reinforcement learning and structured prediction. In the first of these papers, we are again reminded of the limitations that the complexity of the output space imposes on structured prediction algorithms. In order to overcome this limitation, the paper “Structured prediction with reinforcement learning” by Maes et al. (2009) casts structured prediction as a sequential decision problem and applies Reinforcement Learning methods for solving structured prediction problems. The advantage of this method over global models such as CRFs (Lafferty et al. 2001), SVMStruct (Tsochantaridis et al. 2005) and M^3N (Taskar et al. 2004) is that it makes no assumptions on the decomposability of output structure or the cost function. Furthermore, it improves over previous sequential decision formulations of SP (Daumé and Marcu 2005; Daumé et al. 2009), since it does not require an Optimal Learning Trajectory or an Optimal Learning Policy, whose computation is intractable for complex structures. Experimental evaluation of the method on sequence prediction and tree transformation problem shows that the method is competitive with existing methods on simple structures and superior on complex structures.

More connections between reinforcement learning and Structured Prediction are drawn in the paper “Training parsers via inverse reinforcement learning” by Neu and Szepesvári (2009). In this paper, the authors explore the research in grammatical parsing, inverse reinforcement learning, and structured prediction, and find a remarkable amount of overlap between the three fields. Along the way, the authors present a system of notation that unifies a collection of important inverse reinforcement learning algorithms and maps them onto the problem of training structured predictors. In particular, the problem of training parsers is treated in detail. An analysis of the performance of all algorithms on a benchmark dataset shows important differences between the performance of the algorithms and their sensitivity under varying parameter choices.

Structured Prediction is a relatively new subdiscipline of machine learning. We believe these papers provide a useful snapshot of some important aspects of structured prediction. The editors thank the reviewers for their valuable contributions in producing this special issue, as well as Foster Provost and Melissa Fearon of Machine Learning for their unlimited support.

References

- Bakir, G. H., Hofmann, T., Schölkopf, B., Smola, A. J., Taskar, B., & Vishwanathan, S. V. N. (2007). *Neural information processing. Predicting structured data*. Cambridge: MIT Press.
- Daumé III, H., & Marcu, D. (2005). Learning as search optimization: approximate large margin methods for structured prediction. In *Proceedings of the international conference on machine learning* (pp. 169–176).

- Daumé III, H., Langford, J., & Marcu, D. (2009). Search-based structured prediction. *Machine Learning*, 75(3), 297–325.
- Hsu, C.-N., Huang, H.-S., & Chang, Y.-M. (2009). Periodic step-size adaptation in second-order gradient descent for single-pass on-line structured learning. *Machine Learning*, 77(2–3), 195–224.
- Lafferty, J. D., McCallum, A., & Pereira, F. (2001). Conditional random fields: probabilistic modeling for segmenting and labeling sequence data. In *Proceedings of the international conference on machine learning* (pp. 282–289).
- Lampert, C. H., & Blaschko, M. B. (2009). Structured prediction by joint kernel support estimation. *Machine Learning*, 77(2–3), 249–269.
- Lauer, F., & Bloch, G. (2008). Incorporating prior knowledge in support vector machines for classification: a review. *Neurocomputing*, 71(7–9), 1578–1594.
- Maes, F., Denoyer, L., & Gallinari, P. (2009). Structured prediction with reinforcement learning. *Machine Learning*, 77(2–3), 271–301.
- Mao, Y., & Lebanon, G. (2009). Generalized isotonic conditional random fields. *Machine Learning*, 77(2–3), 225–248.
- Neu, G., & Szepesvári, C. (2009). Training parsers by inverse reinforcement learning. *Machine Learning*, 77(2–3), 303–337.
- Sutton, C., & McCallum, A. (2009). Piecewise training for structured prediction. *Machine Learning*, 77(2–3), 165–194.
- Taskar, B., Guestrin, C., & Koller, D. (2004). Max-margin Markov networks. In *NIPS*.
- Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6, 1453–1484.