In *Proceedings of the International Workshop on Statistical Relational Learning (SRL-09)*, Leuven, Belgium, July 2009.

Max-Margin Weight Learning for Markov Logic Networks

Tuyen N. Huynh HNTUYEN@CS.UTEXAS.EDU
Raymond J. Mooney MOONEY@CS.UTEXAS.EDU
The University of Texas at Austin, 1 University Station C0500, Austin, Texas 78712, USA

Keywords: Markov Logic Networks, statistical relational learning, max-margin training

Abstract

Markov logic networks (MLNs) are an expressive representation for statistical relational learning that generalizes both first-order logic and graphical models. Existing discriminative weight learning methods for MLNs all try to learn weights that optimize the Conditional Log Likelihood (CLL) of the training examples. In this work, we present a new discriminative weight learning method for MLNs based on a max-margin framework. This results in a new model, Max-Margin Markov Logic Networks (M3LNs), that combines the expressiveness of MLNs with the predictive accuracy of structural Support Vector Machines (SVMs). To train the proposed model, we design a new approximation algorithm for loss-augmented inference in MLNs based on Linear Programming (LP). The experimental result shows that the proposed approach generally achieves higher F_1 scores than the current best discriminative weight learner for MLNs.

1. Introduction

Existing discriminative training algorithms for learning MLN weights attempt to maximize the conditional log likelihood (CLL) of a set of *target predicates* given evidence provided by a set of *background predicates* (Singla & Domingos, 2005; Lowd & Domingos, 2007; Huynh & Mooney, 2008). If the goal is to predict accurate target-predicate probabilities, this approach is well motivated. However, in many applications, the actual goal is to maximize an alternative performance metric such as classification accuracy or F-measure. Max-margin methods are a competing approach to discriminative training that are well-founded in computational learning theory and have demonstrated empirical success in many applications (Cristianini & Shawe-Taylor, 2000). They also have the advantage that they can be adapted to maximize a variety of performance metrics in addition to classification accuracy (Joachims, 2005). Max-margin methods have been successfully applied to structured prediction problems, such as in Max-Margin Markov Networks (M3Ns) (Taskar et al., 2003) and structural SVMs (Tsochantaridis et al., 2005); however, until now, they have not been applied to an SRL model that generalizes first-order logic such as MLNs.

In this paper, we develop Max-Margin MLNs (M3LNs) by instantiating an existing general framework for max-margin training of structured models (Tsochantaridis et al., 2005). This requires developing a new algorithm for approximating the "loss-augmented" inference in MLNs. Extensive experiments in the two real-world MLN applications demonstrate that M3LNs generally produce improved results when the goal involves maximizing predictive accuracy metrics other than CLL.

Presented at *ILP-MLG-SRL*, Leuven, Belgium, 2009.

2. Max-Margin Weight Learning for MLNs

2.1. Max-Margin Formulation

All of the current discriminative weight learners for MLNs try to find a weight vector $\mathbf{w}$ that optimizes the conditional log-likelihood $P(\mathbf{y}|\mathbf{x})$ of the query atoms $\mathbf{y}$ given the evidence $\mathbf{x}$. However, an alternative approach is to learn a weight vector $\mathbf{w}$ that maximizes the ratio:

$$\frac{P(\mathbf{y}|\mathbf{x},\mathbf{w})}{P(\hat{\mathbf{y}}|\mathbf{x},\mathbf{w})}$$

between the probability of the correct truth assignment $\mathbf{y}$ and the closest competing incorrect truth assignment $\hat{\mathbf{y}} = \arg\max_{\bar{\mathbf{y}} \in \mathbf{Y} \setminus \mathbf{y}} P(\bar{\mathbf{y}}|\mathbf{x})$. For MLNs, this problem translates to the problem of maximizing the following margin:

$$\begin{aligned}\gamma(\mathbf{x},\mathbf{y};\mathbf{w}) &= \mathbf{w}^T\mathbf{n}(\mathbf{x},\mathbf{y}) - \mathbf{w}^T\mathbf{n}(\mathbf{x},\hat{\mathbf{y}}) \\ &= \mathbf{w}^T\mathbf{n}(\mathbf{x},\mathbf{y}) - \max_{\bar{\mathbf{y}} \in \mathbf{Y} \setminus \mathbf{y}} \mathbf{w}^T\mathbf{n}(\mathbf{x},\bar{\mathbf{y}})\end{aligned}$$

where $\mathbf{n}(\mathbf{x},\mathbf{y})$ is a vector in which each component i is the number of true groundings of clause f_i given the truth assignment $(\mathbf{x},\mathbf{y})$ In turn, this max-margin problem can be formulated as a "1-slack" structural SVM (Joachims et al., 2009 to appear) as follows:

Optimization Problem 1 (OP1): Max-Margin Markov Logic Networks

$$\min_{\mathbf{w},\xi\geq 0} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\xi$$
$$s.t.\ \forall \bar{\mathbf{y}} \in Y : \mathbf{w}^T[\mathbf{n}(\mathbf{x},\mathbf{y}) - \mathbf{n}(\mathbf{x},\bar{\mathbf{y}})] \geq \Delta(\mathbf{y},\bar{\mathbf{y}}) - \xi$$

So for MLNs, the number of true groundings of the clauses $\mathbf{n}(\mathbf{x},\mathbf{y})$ plays the role of the feature vector function $\Psi(x,y)$ in the general structural SVM problem. In other words, each clause in an MLN can be viewed as a feature representing a dependency between a subset of inputs and outputs or a relation among several outputs. The optimization problem OP1 can be solved efficiently by the general cutting plane algorithm proposed by Joachims *et al.* (2009-to appear) if we have efficient algorithms to solve the following two $\arg\max$ problems:

Prediction: $\arg\max_{\mathbf{y}\in Y} \mathbf{w}^T\mathbf{n}(\mathbf{x},\mathbf{y})$

Separation Oracle: $\arg\max_{\bar{\mathbf{y}}\in Y}\{\Delta(\mathbf{y},\bar{\mathbf{y}}) + \mathbf{w}^T\mathbf{n}(\mathbf{x},\bar{\mathbf{y}})\}$

It is clear that the prediction problem is just the Most Probable Explanation (MPE) inference problem. For MLNs, this problem is equivalent to the Weighted MAX-SAT problem which is an NP-hard problem (Singla & Domingos, 2005). We can use MaxWalkSAT (Kautz et al., 1997) to get an approximate solution, but we have found that models trained with MaxWalkSAT have very low predictive accuracy. On the other hand, recent work (Finley & Joachims, 2008) has found that fully-connected pairwise Markov random fields, a special class of structural SVMs, trained with overgenerating approximate inference methods (such as relaxation) preserve the theoretical guarantees of structural SVMs trained with exact inference, and exhibit good empirical performance. Based on this result, we sought a relaxation-based approximation for MPE inference. We developed an LP-relaxation algorithm for doing MPE inference and a variant of it which can solve the separation oracle for some specific loss functions.

3. Experimental Evaluation

This section presents experiments comparing M3LNs to the current best discriminative weight learner for MLNs with recursive clauses, *preconditioner scaled conjugate gradient* (PSCG) (Lowd & Domingos, 2007).

3.1. Datasets

We ran experiments on two large, real-world MLN datasets: WebKB for collective web-page classification, and CiteSeer for bibliographic citation segmentation.

The WebKB dataset consists of labeled web pages from the computer science departments of four universities. Different versions of this data have been used in previous work. We used the version from (Lowd & Domingos, 2007), which contains 4165 web pages and 10,935 web links. Each page is labeled with a subset of the categories: person, student, faculty, professor, department, research project, and course. The goal is to predict these categories from the words and links on the web pages. We used the same simple MLN from (Lowd & Domingos, 2007), which only has clauses relating words to page classes, and page classes to the classes of linked pages.

For CiteSeer, we used the dataset and MLN used in (Poon & Domingos, 2007). The dataset has 1,563 citations and each of them is segmented into three fields: Author, Title and Venue. The dataset has four disconnected segments corresponding to four different research topics. We used the simplest MLN in (Poon & Domingos, 2007), which is the isolated segmentation model.

3.2. Metrics

We used F_1, the harmonic mean of recall and precision, to measure the performance of each algorithm. This is the standard evaluation metric in multi-class text categorization and information extraction. For systems that compute marginal probabilities rather than MPEs, we predict that an atom is true iff its probability is at least 0.5.

3.3. Methodology

We ran four-fold cross-validation (i.e. leave one university/topic out) on both datasets. For the max-margin weight learner, we used a simple process for selecting the value of the C parameter. For each train/test split, we trained the algorithm with five different values of C: 1, 10, 100, 1000, and 10000, then selected the one which gave the highest average F_1 score on training. The ϵ parameter was set to 0.001. To solve the QP problems in the cutting plane algorithm and LP problems in the LP-relaxation MPE inference, we used the, we used the MOSEK [1] solver. The PSCG algorithm was carefully tuned by its author. For prediction, we ran MCSAT (Poon & Domingos, 2006) for 100 burn-in and 1000 sampling iterations to get the marginal conditional probability of each query atom, and ran LP-relaxation MPE inference to obtain the most probable truth assignment to all query atoms.

3.4. Results and Discussion

Table 1. F_1 score on WebKB

	AVERAGE F_1
PSCG-MCSAT	0.465 +/- 0.115
PSCG-LPRELAX	0.474 +/- 0.115
MM-LPRELAX	**0.601** +/- **0.100**

Table 1 and 2 present the performance of different systems on the WebKB and Citeseer datasets. Each system is named by the weight learner used and the inference algorithm used

[1] http://www.mosek.com/

Table 2. F_1 score on CiteSeer with different parameter values

	AVERAGE F_1
PSCG-MCSAT-5	0.864 +/- 0.035
PSCG-MCSAT-10	0.939 +/- 0.022
PSCG-MCSAT-15	0.861 +/- 0.047
PSCG-MCSAT-20	0.801 +/- 0.060
PSCG-MCSAT-100	0.656 +/- 0.035
MM-LPRELAX-1	0.934 +/- 0.013
MM-LPRELAX-10	0.932 +/- 0.013
MM-LPRELAX-100	0.932 +/- 0.013
MM-LPRELAX-1000	0.933 +/- 0.015
MM-LPRELAX-10000	0.935 +/- 0.020

in testing. For the max-margin (MM) learner, the inference used in training is the loss-augmented version of the one used in testing. For example, MM-LPRelax is the MM weight learner trained with the loss-augmented (Hamming loss) LP-relaxation MPE inference algorithm and tested with the LP-relaxation MPE inference algorithm.

On WebKB, the max-margin weight learner achieves the best F_1 score (0.601), which is much higher than the 0.465 F_1 score obtained by the current best discriminative weight learner for MLN, PSCG. This improvement is clearly due to the max-margin approach since LP-relaxation MPE inference improves the accuracy of PSCG a bit (the PSCG-LPRelax system), but it is still far from that of the max-margin weight learner.

On the Citeseer dataset, the performance of max-margin methods are very close to those of PSCG. However, its performance is much more stable than that of PSCG. Table 2 shows the performance of MM weight learners and PSCG with different parameter values by varying the C value for MM and the number of iterations for PSCG. The best number of iterations for PSCG is 9 or 10. In principle, we should run PSCG until it converges to get the optimal weight vector. However, in this case, the performance of PSCG drops drastically on both training and testing after a certain number of iterations. For example, from Table 2 we can see that at 10 iterations PSCG achieves the best F_1 score of 0.939, but after 15 iterations, its F_1 score drops to 0.861 which is much worse than the max-margin weight learners. Moreover, if we let it run until 100 iterations, then its F_1 score is only 0.656. On the other hand, the performance of MM only varies a little bit with different values of C and we don't need to tune the number of iterations of MM.

4. Conclusions

We have presented a max-margin weight learning method for MLNs based on the framework of structural SVMs. It resulted in a new model, M3LN, that has the representational expressiveness of MLNs and the predictive performance of SVMs. M3LNs can be trained to optimize different performance measures depending on the needs of the application. To train the proposed model, we developed a new approximation algorithm for loss-augmented MPE inference in MLNs based on LP-relaxation. The experimental results showed that the new max-margin learner generally has better and more stable predictive accuracy (as measured by F_1) than the current best discriminative MLN weight learner.

Acknowledgments

We thank Daniel Lowd and Hoifung Poon for useful discussions and helping with the experiments. This research is sponsored by the DARPA and managed by the AFRL under contract FA8750-05-2-0283. The project is also partly support by ARO grant W911NF-08-1-0242. Most of the experiments were run on the Mastodon Cluster, provided by NSF Grant EIA-0303609.

References

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Finley, T., & Joachims, T. (2008). Training structural SVMs when exact inference is intractable. *ICML-08* (pp. 304–311).

Huynh, T. N., & Mooney, R. J. (2008). Discriminative structure and parameter learning for Markov logic networks. *ICML-08* (pp. 416–423).

Joachims, T. (2005). A support vector method for multivariate performance measures. *ICML-05* (pp. 377–384).

Joachims, T., Finley, T., & Yu, C.-N. (2009-to appear). Cutting-plane training of structural SVMs. *MLJ*. `http://www.cs.cornell.edu/People/tj/publications/joachims_etal_09a.pdf`.

Kautz, H., Selman, B., & Jiang, Y. (1997). A general stochastic approach to solving problems with hard and soft constraints. *The Satisfiability Problem: Theory and Applications* (pp. 573–586). AMS.

Lowd, D., & Domingos, P. (2007). Efficient weight learning for Markov logic networks. *PKDD-2007* (pp. 200–211).

Poon, H., & Domingos, P. (2006). Sound and efficient inference with probabilistic and deterministic dependencies. *AAAI-2006*. Boston, MA.

Poon, H., & Domingos, P. (2007). Joint inference in information extraction. *AAAI-2007* (pp. 913–918).

Singla, P., & Domingos, P. (2005). Discriminative training of Markov logic networks. *AAAI-2005* (pp. 868–873).

Taskar, B., Guestrin, C., & Koller, D. (2003). Max-margin Markov networks. *NIPS-03*.

Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *JMLR*, *6*, 1453–1484.