Online Stacked Graphical Learning

Zhenzhen Kou
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
zkou@andrew.cmu.edu

Vitor R. Carvalho
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
vitor@cs.cmu.edu

William W. Cohen
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
wcohen@cs.cmu.edu

ABSTRACT

Collective classification has been widely studied to predict class labels simultaneously for relational data, such as hyperlinked webpages, social networks, and data in a relational database. Recently there have been studies on relational models for collective inference, such as relational dependency networks [1], relational Markov networks [2], and Markov logic networks [3]. The existing collective classification methods are usually expensive due to the iterative inference in graphical models and their learning procedures based on iterative optimization. When the dataset is large, the cost of maintaining large graphs or related instances in memory becomes a problem as well.

Stacked graphical learning (SGL) has been proposed for collective classification with efficient inference - as shown in [4], stacked graphical learning is 40 to 80 times faster than Gibbs sampling during inference. Stacked graphical learning augments a base learner by providing the predicted labels of related instances and using relational template to build extended features to capture the dependencies among data. To obtain the predictions for training examples, stacked graphical learning applies a cross-validation-like technique. Hence, the memory and time cost of standard stacked graphical learning is still expensive during training.

Online Stacked Graphical Learning In this paper, we propose online stacked graphical learning. The novel online scheme for stacked graphical learning is based on a combination of stacked graphical learning with a recently-developed single-pass online learning algorithm. During the learning procedure of an online learner, the intermediate predictions for training data are generated to learn the online model. Thus the predictions for training data can be obtained naturally and there is no need to apply the base learner several times to the training data to obtain the predictions. A single-pass online learner, *Modified Balanced Winnow* (MBW), is presented in Carvalho & Cohen's work [5] and has been demonstrated to be able to provide excellent performance - even comparable to batch learners.

One practical difficulty is that, while online learning methods produce satisfactory predictions after learning on the whole training set, the intermediate predictions for the training data in the starting stage can be quite inaccurate. Thus, to obtain fair "predictions" for training examples, we define a burn-in data size b. That is, after training on b examples, we start recording intermediate predictions from the online learner and expanding features with the predictions. The learning procedure of online stacked learning is shown in Figure 1.

One thing we would like to point out is that, in stacked graphical learning for collective classification, given an instance x_i, we need to apply the relational template to retrieve the predicted labels for the related instances to extend features. Assume x_i and its neighbors are contained in a subset, we provide the instances in a subset to the online learner as a group and extend the features after the predictions for instances in the whole subset are made. Therefore in general, we provide the instances in groups to the base learner and the burn-in data size b will be chosen to include a few subsets of instances. In practise, the dataset might not be able to be split into disjoint subsets easily. In our work we will demonstrate how to split the dataset into subsets empirically.

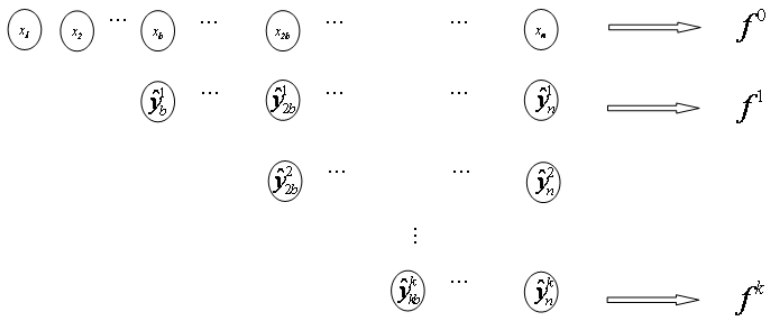

Figure 1: Online Stacked Graphical Learning

Efficiency Analysis Theoretically, when there are infinitely many training examples, i.e., $kb << n$, applying the online stacked graphical learning scheme only requires single-pass training over the training set. We do not need to apply the cross-validation-like trick to get the predictions for training examples. Therefore, online stacked graphical learning can save training time. We will show the speed-up experimentally as well.

In online stacked graphical learning, there are reliable predictions at level k after $(k+1)b$ examples have streamed by, and the learner needs to maintain only k classifiers and does not need to store examples. Therefore, the algorithm can save memory. This becomes extremely important when the size of training data is huge. Also this feature allows online stacked graphical learning to be applied to streaming data.

Experimental Results We evaluated online stacked graphical learning on eleven tasks from three domains - collective classification over relational datasets, sequential partitioning [6], and named entity extraction. In the abstract, we only show the performance. Detailed descriptions about the datasets can be found in previous work [4] or in full version of this paper.

To evaluate the effectiveness of online stacked graphical learning on the collective classification task, we compare local models (i.e., the base learner), stacked models, and a state-of-art competitive model. We evaluated two local models, MaxEnt and MBW, for the collective classification

Table 1: Performance of online stacked graphical learning for relational datasets: accuracy for "Document classification" and F1-accuracy for "SLIF" are reported. We evaluated two local models: MaxEnt and MBW. We also compared to a competitive relational model - relational dependency networks. The standard stacked model used two-fold-cross-validation predictions. The online stacked graphical model is based on MBW. We used 1 level of stacking, i.e., K=1.

	SLIF	Document classification		
		WebKB	Cora	CiteSeer
Local model				
MaxEnt	81.5	58.3	63.9	55.3
MBW	82.3	58.6	63.7	56.1
Competitive relational model				
Relational Dependency Networks	86.7	**74.2**	72.9	58.7
Stacked model				
Standard Stacked model (with MaxEnt, k=1)	90.1	73.2	73.8	59.8
Standard Stacked model (with MBW, k=1)	92.1	**74.2**	**73.5**	**60.3**
Online Stacked model (k=1)	**92.3**	74.1	71.3	-

Table 2: Accuracy comparison of online stacked graphical learning for sequential partitioning. We evaluated two local models: MaxEnt and MBW. We compared to a competitive graphical model - conditional random fields. The standard stacked model used two-fold-cross-validation predictions. The online stacked graphical model is based on MBW. We used 1 level of stacking.

	Sequential Partitioning		
	FAQ	signature	video
Local model			
MaxEnt	67.3	96.3	80.9
MBW	64.9	96.5	78.4
Competitive relational model			
CRFs	85.6	98.1	83.0
Stacked model			
Standard Stacked model (with MaxEnt, k=1)	**87.1**	98.1	**85.8**
Standard Stacked model (with MBW, k=1)	84.1	**98.3**	85.5
Online Stacked model (k=1)	86.3	**98.3**	85.7

tasks. We considered a standard stacked model based on MaxEnt (with two-fold-cross-validation predictions), a standard stacked model based on MBW (with two-fold-cross-validation predictions), and an online stacked graphical model based on MBW. We also compared our stacked graphical model to a state-of-art relational graphical model, *relational dependency networks* [1].

Table 1 shows that on all of the four relational datasets, stacked graphical learning improves the performance of the base learner significantly. The two local models achieved performance of the same level, so did the stacked graphical models based on them. Our comparison to relational dependency networks shows that stacked models can achieve competitive results to the state-of-art model. However, the online stacked graphical model requires much less training time, which will be discussed later.

Table 2 shows the performance of online stacked models on sequence partitioning. The state-of-art models we consider here are conditional random fields (CRFs). CRFs

Table 3: Performance of online stacked graphical learning for Named Entity Extraction, F1 accuracy is reported. "Relational template 1" returns predictions of adjacent tokens only, "relational template 2" returns predictions of adjacent and repeated tokens.

	Named Entity Extraction			
	UT	Yapex	Genia	CSpace
Local model				
MaxEnt	69.1	62.1	66.5	74.2
MBW	67.9	62.3	66.9	75.1
Competitive relational model				
CRFs	73.1	65.7	72.0	80.3
Stacked model				
With relational template 1				
Standard Stacked model (with MaxEnt, k=1)	70.1	63.7	70.8	77.9
Standard Stacked model (with MBW, k=1)	72.1	63.9	71.3	79.9
Online Stacked model (with MBW, k=1)	72.6	64.6	72.3	80.0
With relational template 2				
Standard Stacked model (with MaxEnt, k=1)	77.3	68.2	78.5	82.1
Standard Stacked model (with MBW, k=1)	**76.6**	68.9	**78.9**	83.3
Online Stacked model (with MBW, k=1)	**76.6**	**69.1**	**78.9**	**83.4**

are sequential models that can capture the sequential dependency. On all of the three datasets, stacked graphical learning improves the performance of the base learner significantly. The MaxEnt model did better than MBW on two of three tasks, yet the stacked graphical models based on them achieved performance of the same level.

Table 3 reported the F1 of online stacked graphical learning for Named Entity Extraction. One relational template captures sequential dependency only(denoted as relational template 1 in Table 3), the other one can also capture the dependency among the adjacent and repeated tokens(denoted as relational template 2 in Table 3).

Table 3 shows that on all of the four named entity extraction tasks, stacked graphical learning improves the performance of the base learner. With relational template 1, the stacked graphical models can capture the sequential dependency and achieved comparable results to CRFs. With relational template 2, the stacked graphical models achieved better performance than CRFs. Moreover, the online stacked graphical model requires much less training time.

Efficiency of the Training for Stacked Graphical Learning One big success of online stacked graphical learning is that the learning is an online procedure and thus very efficient. We compared the training time of online stacked graphical models (with one iteration) to that of competitive relational models and the baseline standard stacked graphical model. The baseline algorithm we compare to is the best algorithm in previous work [4], i.e., the standard stacked graphical model based on MaxEnt, with 5-fold-cross-validation to obtain predictions during training.

Table 4 shows the speedup, i.e., in the table "38.1" means the training in standard stacked graphical learning is 38.1 times slower than that of online stacked graphical learning. Table 4 shows that compared to online stacked graphical learning, standard stacked graphical learning based on MaxEnt is approximately 57 times slower in training. We also compared online stacked graphical learning with the competitive relational models. Table 4 shows that online stacked graphical learning is approximately 14 times faster

Table 4: Comparison on training time.

	Standard SGM vs Online SGM	Competitive relational model vs Online SGM
SLIF	38.1	7.9
WebKB	50.0	10.1
Cora	49.7	9.9
Signature	67.4	13.6
FAQ	69.0	14.0
Video	45.0	9.7
UT	68.7	20.3
Yapex	60.6	17.1
Genia	69.4	22.4
CSpace	52.0	15.3
Average speed-up	57.0	14.0

in training.

Moreover, in the previous work [4], it has been shown that during inference stacked graphical learning is 40 to 80 times faster than Gibbs sampling in relational dependency networks.

To summarize, the experimental results demonstrate the effectiveness and efficiency of online stacked graphical models, i.e., the proposed online stacked graphical learning presents accurate and reliable predictions, but with considerably faster training time and smaller memory requirements.

With high accuracy, fast training, and low memory footprint, online stacked graphical learning is very competitive for real world large-scale applications. Furthermore, because the proposed scheme does not need to keep all previous examples in memory, it can effectively handle data in streaming format.

1. REFERENCES

[1] D. Jensen and J. Neville , *Dependency Networks for Relational Data*, Proceedings of ICDM-04, Brighton, UK 2004.

[2] B. Taskar and P. Abbeel and D. Koller , *Discriminative Probabilistic Models for Relational Data*, Proceedings of UAI-02, Edmonton, Canada, 2002.

[3] M. Richardson and P. Domingos, *Markov Logic Networks*, Machine Learning, 62, pp107–136, 2006.

[4] Z. Kou and W. W. Cohen, *Stacked Graphical Learning for Efficient Inference in Markov Random Fields*, Proceedings of SDM 07, Minneapolis, MN, 2007.

[5] V. R. Carvalho and W. W. Cohen, *Single-Pass Online Learning: Performance, Voting Scheme and Online Feature Selection*, Proceedings of KDD-2006, Philadelphia, PA, 2006.

[6] William W. Cohen and Vitor R. Carvalho,*Stacked Sequential Learning*, Proceedings of the IJCAI 2005.