Joint Judgments with a Budget: Strategies for Reducing the Cost of Inference

Jay Pujara JAY@CS.UMD.EDU
Hui Miao HUI@CS.UMD.EDU
Lise Getoor GETOOR@CS.UMD.EDU
Department of Computer Science University of Maryland College Park, MD 20742

1. Motivation

Machine learning techniques are often subjected to test-time budgets, in scenarios that range from choosing which advertisements to show a user of a social network to detecting faces with the limited resources available in a digital camera. Solutions to this problem formulate a trade-off between the cost and quality of a model(Viola & Jones, 2001; Saberian & Vasconcelos, 2010; Xu et al., 2013) by formulating a loss function that combines a measure of error with a measure of computational cost. Minimizing this loss function on a training set drawn from the expected distribution of instances produces a model sensitive to computational cost. However, most work in this field has focused on classifiers that make predictions on independent instances, such as a single user or a single image.

We consider a different setting, where judgments are made jointly over a set of instances. Structural or temporal applications can benefit from making such judgments jointly: choosing advertisements for a set of related users(Sharara et al., 2011) or recognizing actors in a sequence of images(Khamis et al., 2012) provide superior results. Performing such joint reasoning under a test-time budget requires a different approach; instead of feature computation, the key contributors to the computational cost in such models are the dependencies between predictions.

We investigate methods to allow joint judgments to be made in a situation where meeting test-time budgets are critical. Specifically, we consider the Never-Ending Language Learner (NELL)(Carlson et al., 2010): a system that is continually extracting information from the web and attempting to use its discoveries to improve extraction. While NELL's extractions are linked by an ontology, a joint approach to reasoning about them requires operating on millions of facts, a scale which can be prohibitively time-consuming. Beginning with a model that fully captures the rich dependencies between these extractions, we degrade the model in two ways: (1) partitioning the extractions and running inference in parallel and (2) removing ontological constraints. We estimate a cost model that incorporates the dependencies between variables and show how this model of cost can be used to reduce the running time of inference.

2. Setting

In the joint inference setting, we seek to estimate the distribution $P(Y|X)$ and determine the most likely values of target variables Y based on observations X. We represent this probability distribution with a Markov random field (MRF), and express variable interactions with a set of logical rule templates. R refers to groundings of these rules generated by atoms in Y and X. The MRF can be decomposed into a set of weighted potentials corresponding to these ground rules, $\phi_r(I)$ having weight w_r, where I is an assignment of the variables Y:

$$f(I) = \frac{1}{Z}\exp\left[\sum_{r\in R} w_r\phi_r(I)\right]$$

Maximizing $f(I)$ corresponds to finding the most likely assignment to the variables, Y.

In our particular setting, the observations X are a set of noisy extractions from the web and the target variables Y are a set of unknown facts. We introduce rules to relate the noisy extractions to potential facts, as well as rules that incorporate ontological information about the relationship between facts. An interpretation, I, is a set of facts considered to be true using the evidence from the extraction system and ontological constraints.

One simple intuition is that the computation of $f(I)$ is related to the number of potential functions, ϕ_r incorporated into our model, and reducing the size of $|R|$ can reduce computation. Reducing the number of ground rules can be accomplished by reducing the number of observations, X or using a simpler model with fewer rules. We consider both strategies and show their impact on a joint inference task.

3. Experiments

We present initial work to show how the test-time cost of using a joint model can be modulated through partitioning and model simplification in experiments on a large and complex joint inference problem: a NELL dataset consisting of 1.7M extractions and 80K ontological constraints.We adopt the formulation of (Jiang et al., 2012) to relate extractions and ontological relationships to facts. Our model (Pujara et al., 2013) uses a a continuous-valued Markov random field, which we formulate using Probabilistic Soft Logic (PSL)(Broecheler et al., 2010). PSL relaxes the truth-value of logical atoms to the $[0, 1]$ domain and formulates Most Probable Explanation (MPE) inference as a convex optimization which scales linearly with the problem in practice(Bach et al., 2012).

The web-extraction data from the NELL project contains candidate relations and labels that can be expressed through logical predicates such as CandRel(`Yankees`, `baseball`, `teamPlaysSport`) and CandLbl(`Yankees`, `sportsteam`). We relate these extractions to the facts using weighted rules:

$$\begin{array}{lcl} \textsc{CandRel}(E_1, E_2, R) & \overset{w_{CR}}{\Rightarrow} & \textsc{Rel}(E_1, E_2, R) \\ \textsc{CandLbl}(E, L) & \overset{w_{CL}}{\Rightarrow} & \textsc{Lbl}(E, L) \end{array}$$

Additionally, facts are related through an ontology that specifies the domain and range of relations, inversely-related relations, mutually-exclusive relations and categories, and subsumed relations and categories. We express this ontology using the logical rules below.

$$\begin{array}{llll} \textsc{Dom}(R, L) & \wedge; \textsc{Rel}(E_1, E_2, R) & \Rightarrow & \textsc{Lbl}(E_1, L) \\ \textsc{Rng}(R, L) & \wedge\ \textsc{Rel}(E_1, E_2, R) & \Rightarrow & \textsc{Lbl}(E_2, L) \\ \textsc{Inv}(R, S) & \wedge\ \textsc{Rel}(E_1, E_2, R) & \Rightarrow & \textsc{Rel}(E_2, E_1, S) \\ \textsc{Sub}(L, P) & \wedge\ \textsc{Lbl}(E, L) & \Rightarrow & \textsc{Lbl}(E, P) \\ \textsc{RSub}(R, S) & \wedge\ \textsc{Rel}(E_1, E_2, R) & \Rightarrow & \textsc{Rel}(E_1, E_2, S) \\ \textsc{Mut}(L_1, L_2) & \wedge\ \textsc{Lbl}(E, L_1) & \Rightarrow & \neg\textsc{Lbl}(E, L_2) \\ \textsc{RMut}(R, S) & \wedge\ \textsc{Rel}(E_1, E_2, R) & \Rightarrow & \neg\textsc{Rel}(E_1, E_2, S) \end{array}$$

We evaluate model performance using the F1-score and area under the P-R curve (AUC), and measure computational performance using running time.

In our first experiments we partition the observations, in this case the set of candidate relations and labels. We perform inference on each partition and combine the results of this distributed inference, averaging values when there is an overlap. One risk of partitioning the observations is separating related evidence, which we reduce by formulating the ontology as a graph and use a graph min-cut algorithm to identify 6 clusters of related relations and labels, which form our partitions. We consider results with 6 partitions corresponding to 6 clusters, and use combinations of these clusters to generate 2 and 3 partitions and compare to the full joint inference (1 partition). As shown in Table 1, F1 and AUC are similar with 1, 2, and 3 partitions but degrade substantially with 6 partitions. The maximum running time across partitions is reported, and partitioning can reduce running time substantially.

Table 1. Results using partitioned model

Partitions	AUC	F1	Time (min.)
6	0.463	0.564	44
3	0.734	0.775	61
2	0.735	0.775	71
1	0.736	0.774	131

Table 2. Results with subsets of ontological constraints

Partitions	AUC	F1	Time (min.)
Dom/ Rng/ Inv	0.714	0.775	35
Mut/ RMut	0.733	0.775	12
Sub/ RSub	0.738	0.775	10
Mut/ Sub	0.763	0.775	80
RMut/ RSub	0.681	0.775	10
ALL	0.736	0.774	131

We also show how limiting the complexity of the joint probabilistic model can change the results of inference. We consider a number of simpler models that use a subset of the ontology. We chose five sets of coordinated ontological rules to emulate simpler models that capture important relationships: Dom/ Rng/ Inv; Mut/ RMut; Sub/ RSub; Mut/ Sub; and RMut/ RSub. As seen in Table 2, performance varies greatly depending on the ontological constraints included in the model but, in some cases, a subset of ontological rules can outperform a model that uses all ontological information while reducing the running time dramatically.

Finally, we combine both techniques, using three partitions and a model that only includes the Mut and Sub constraints. Inference in this setting reduces running time to 32 minutes while maintaining an AUC of .763 and an F1 score of .775. These experiments demonstrate that the test-time cost of joint inference can be controlled to provide a trade-off between computation and accuracy. Many open questions remain about the formal relationship between the computational costs and performance of joint models. Although our initial results offer promising directions, our eventual goal is to produce a system that can use techniques such as partitioning and model simplification to meet arbitrary test-time budgets or trade-offs.

References

Bach, S. H, Broecheler, M, Getoor, L, and O'Leary, D. P. Scaling MPE Inference for Constrained Continuous Markov Random Fields with Consensus Optimization. In *NIPS*, 2012.

Broecheler, M, Mihalkova, L, and Getoor, L. Probabilistic Similarity Logic. In *UAI*, 2010.

Carlson, A, Betteridge, J, Kisiel, B, Settles, B, Hruschka, E. R, and Mitchell, T. M. Toward an Architecture for Never-Ending Language Learning. In *AAAI*, 2010.

Jiang, S, Lowd, D, and Dou, D. Learning to Refine an Automatically Extracted Knowledge Base Using Markov Logic. In *ICDM*, 2012.

Khamis, S, Morariu, V. I, and Davis, L. S. Combining per-frame and per-track cues for multi-person action recognition. In *ECCV*, 2012.

Pujara, J, Miao, H, Getoor, L, and Cohen, W. Knowledge graph identification. In *ICML workshop on Structured Learning*, 2013.

Saberian, M and Vasconcelos, N. Boosting Classifier Cascades. In *NIPS*, 2010.

Sharara, H, Rand, W, and Getoor, L. Differential adaptive diffusion: Understanding diversity and learning whom to trust in viral marketing. In *ICWSM*, 2011.

Viola, P and Jones, M. Rapid object detection using a boosted cascade of simple features. *CVPR*, 2001.

Xu, Z, Kusner, M, Chen, M, and Weinberger, K. Q. Cost-sensitive tree of classifiers. In *ICML*, 2013.