Reasoning with Classifiers*

Dan Roth

Department of Computer Science
University of Illinois at Urbana-Champaign
danr@cs.uiuc.edu

Abstract. Research in machine learning concentrates on the study of learning *single* concepts from examples. In this framework the learner attempts to learn a single hidden function from a collection of examples, assumed to be drawn independently from some unknown probability distribution. However, in many cases – as in most natural language and visual processing situations – decisions depend on the outcomes of several different but mutually dependent classifiers. The classifiers' outcomes need to respect some constraints that could arise from the sequential nature of the data or other domain specific conditions, thus requiring a level of inference on top the predictions.
We will describe research and present challenges related to *Inference with Classifiers* – a paradigm in which we address the problem of using the outcomes of several different classifiers in making coherent inferences – those that respect constraints on the outcome of the classifiers. Examples will be given from the natural language domain.

The emphasis of the research in machine learning has been on the study of learning *single* concepts from examples. In this framework the learner attempts to learn a single hidden function from a collection of examples, assumed to be drawn independently from some unknown probability distribution, and its performance is measured when classifying future examples.

In the context of natural language, for example, work in this direction has allowed researchers and practitioners to address the robust learnability of predicates such as "the part-of-speech of the word `can` in the given sentence is `noun`", "the semantic sense of the word "plant" in the given sentence is "an industrial plant", or determine, in a given sentence, the word that starts a noun phrase. In fact, a large number of disambiguation problems such as part-of speech tagging, word-sense disambiguation, prepositional phrase attachment, accent restoration, word choice selection in machine translation, context-sensitive spelling correction, word selection in speech recognition and identifying discourse markers have been addressed using machine learning techniques – in each of these problems it is necessary to disambiguate two or more [semantically, syntactically or structurally]-distinct forms which have been fused together into the same representation in some medium; a stand alone classifier can be learned to perform these task quite successfully [10].

* Paper written to accompany an invited talk at ECML'02. This research is supported by NSF grants IIS-99-84168,ITR-IIS-00-85836 and an ONR MURI award.

T. Elomaa et al. (Eds.): ECML, LNAI 2430, pp. 506-510, 2002.

However, in many cases – as in most natural language and visual processing situations – higher level decisions depend on the outcomes of several different but mutually dependent classifiers. Consider, for example, the problem of *chunking* natural language sentences where the goal is to identify several kinds of phrases (e.g. noun (NP), verb (VP) and prepositional (PP) phrases) in sentences, as in:

> [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only $ 1.8 billion] [PP in] [NP September] .

A task of this sort involves multiple predictions that interact in some way. For example, one way to address the problem is to utilize two classifiers for each type of phrase, one of which recognizes the beginning of the phrase, and the other its end. Clearly, there are constraints over the predictions; for instance, phrases cannot overlap and there may also be probabilistic constraints over the order of phrases and over their lengths. The goal is to minimize some global measure of accuracy, not necessarily to maximize the performance of each individual classifier involved in the decision [8].

As a second example, consider the problem of recognizing the *kill (KFJ, Oswald)* relation in the sentence "`J. V. Oswald was murdered at JFK after his assassin, R. U. KFJ...`". This task requires making several local decisions, such as identifying named entities in the sentence, in order to support the relation identification. For example, it may be useful to identify that Oswald and KFJ are *people*, and JFK is a *location*. In addition, it is necessary to identify that the action *kill* is described in the sentence. All of this information will help to discover the desired relation and identify its arguments. At the same time, the relation *kill* constrains its arguments to be *people* (or at least, not to be *location*s) and, in turn, helps to enforce that Oswald and KFJ are likely to be *people*, while JFK is not.

Finally, consider the challenge of designing a free-style natural language user interface that allows users to request in-depth information from a large collection of on-line articles, the web, or other semi-structured information sources. Specifically, consider the computational processes required in order to "understand" a simple question of the form "`what is the fastest automobile in the world?`", and respond correctly to it. A straight forward key-word search may suggest that the following two passages contain the answer:

> ... will stretch Volkswagen's lead in *the world's fastest* growing *vehicle* market. Demand for *cars* is expected to soar...

> ... the Jaguar XJ220 is the dearest (415,000 pounds), *fastest* (217mph) and most sought after *car in the world*.

However, "understanding" the question and the passages to a level that allows a decision as to which in fact contains the correct answer, and extracting it, is a very challenging task.

Traditionally, the tasks described above have been viewed as inferential tasks [4, 7]; the hope was that stored knowledge about the language and the world will

allow inferring the syntactic and semantic analysis of the question and the candidate answers; background knowledge (e.g., Jaguar is a car company; automobile is synonymous to car) will then be used to choose the correct passage and to extract the answer. However, it has become clear that many of the difficulties in this task involve problems of context-sensitive ambiguities. These are abundant in natural language and occur at various levels of the processing, from syntactic disambiguation (is "demand" a Noun or a Verb?), to sense and semantic class disambiguation (what is a "Jaguar"?), phrase identification (importantly, "the world's fastest growing vehicle market" is a noun phrase in the passage above) and others. Resolving any of these ambiguities require a lot of knowledge about the world and the language, but knowledge that cannot be written "explicitly" ahead of time. It is widely accepted today that any robust computational approach to these problems has to rely on a significant component of statistical learning, used both to acquire knowledge and to perform low level predictions of the type mentioned above.

The inference component is still very challenging. This view suggests, however, that rather than a deterministic collection of "facts" and "rules", the inference challenge stems from the interaction of the large number of learned predictors involved. Inference of this sort is needed at the level of determining an answer to the question. An answer to the abovementioned question needs to be a name of a car company (predictor 1: identify the sought after entity; predictor 2: determine if the string Z represents a name of a car company) but also the subject of a sentence (predictor 3) in which a word equivalent to "fastest" (predictor 4) modifies (predictor 5) a word equivalent to "automobile" (predictor 6). Inferences of this sort are necessary also at other, lower levels of the process, as in the abovementioned problem of identifying noun phrases in a given sentence.

Thus, decisions typically depend on the outcomes of several predictors and they need to be made in ways that provide coherent inferences that satisfy some constraints. These constraints might arise from the sequential nature of the data, from semantic or pragmatic considerations or other domain specific conditions.

The examples described above exemplify the need for a unified theory of learning and inference. The purpose of this talk is to survey research in this direction, present progress and challenges.

Earlier works in this direction have developed the *Learning to Reason* framework - an integrated theory of learning, knowledge representation and reasoning within a unified framework [2, 9, 12]. This framework addresses an important aspect of the fundamental problem of unifying learning and reasoning - it proves the benefits of performing reasoning on top of learned hypotheses. And, by incorporating learning into the inference process it provides a way around some knowledge representation and comprehensibility issues that have traditionally prevented efficient solutions.

The work described here – on *Inference with Classifiers* – can be viewed as a concrete instantiation of the Learning to Reason framework; it addresses a second important aspect of a unified theory of learning and reasoning, the one which stems from the fact that, inherently, inferences in some domains involve

a large number of predictors that interact in different ways. The fundamental issue addressed is that of systematically combine, chain and perform inferences with the outcome of a large number of mutually dependent learned predictors.

We will discuss several well known inference paradigms, and show how to use those for inference with classifiers. Namely, we will use these inference paradigms to develop inference algorithms that take as input outcomes of classifiers and provide coherent inferences that satisfy some domain or problem specific constraints. Some of the inference paradigms used are hidden Markov models (HMMs), conditional probabilistic models [8, 3], loopy Bayesian networks [6, 11], constraint satisfaction [8, 5] and Markov random fields [1].

Research in this direction may offer several benefits over direct use of classifiers or simply using traditional inference models. One benefit is the ability to directly use powerful classifiers to represent domain variables that are of interest in the inference stage. Advantages of this view have been observed in the speech recognition community when neural network based classifiers were combined within an HMM based inference approach, and have been quantified also in [8]. A second key advantage stems from the fact that only a few of the domain variables are actually of any interest at the inference stage. Performing inference with outcomes of classifiers allows for abstracting away a large number of the domain variables (which will be used only to define the classifiers' outcomes) and will be beneficial also computationally.

Research in this direction offers several challenges to AI and Machine Learning researchers. One of the key challenges of this direction from the machine learning perspective is to understand how the presence of constraints on the outcomes of classifiers can be systematically analyzed and exploited in order to derive better learning algorithms and for reducing the number of labeled examples required for learning.

References

1. D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. M. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75, 2000.
2. R. Khardon and D. Roth. Learning to reason. *Journal of the ACM*, 44(5):697–725, Sept. 1997.
3. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, Williamstown, MA, 2001.
4. J. McCarthy. Programs with common sense. In R. Brachman and H. Levesque, editors, *Readings in Knowledge Representation, 1985*. Morgan-Kaufmann, 1958.
5. M. Munoz, V. Punyakanok, D. Roth, and D. Zimak. A learning approach to shallow parsing. In *EMNLP-VLC'99, the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 168–178, June 1999.
6. K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of Uncertainty in AI*, pages 467–475, 1999.

7. N. J. Nilsson. Logic and artificial intelligence. *Artificial Intelligence*, 47:31–56, 1991.
8. V. Punyakanok and D. Roth. The use of classifiers in sequential inference. In *NIPS-13; The 2000 Conference on Advances in Neural Information Processing Systems*, pages 995–1001. MIT Press, 2001.
9. D. Roth. Learning to reason: The non-monotonic case. In *Proc. of the International Joint Conference on Artificial Intelligence*, pages 1178–1184, 1995.
10. D. Roth. Learning to resolve natural language ambiguities: A unified approach. In *Proc. of the American Association of Artificial Intelligence*, pages 806–813, 1998.
11. D. Roth and W.-T. Yih. Probabilistic reasoning for entity and relation recognition. In *COLING 2002, The 19th International Conference on Computational Linguistics*, 2002.
12. L. G. Valiant. Robust logic. In *Proceedings of the Annual ACM Symp. on the Theory of Computing*, 1999.