Knowledge-Driven Learning and Discovery

Benjamin Lambert & Scott E. Fahlman

Language Technologies Institute, Carnegie Mellon University
Pittsburgh, PA 15213
benlambert@cmu.edu, sef@cs.cmu.edu

Abstract

The goal of our current research is machine learning with the help and guidance of a knowledge base (KB). Rather than learning numerical models, our approach generates explicit symbolic hypotheses. These hypotheses are subject to the constraints of the KB and are easily human-readable and verifiable. Toward this end, we have implemented algorithms that hypothesize new relations and new types of entities in a KB by examining structural regularities in the KB that represent implicit knowledge. We evaluate these algorithms on a publications KB and a zoology KB.

Introduction

Our current research is learning and discovery with the help and guidance of a knowledge base (KB). Rather than learning a numerical model, such as a feature vector, our system proposes specific symbolic hypotheses. Hypotheses are then checked for consistency with the knowledge in the KB to ensure that they do not conflict with background knowledge.

We can discover interesting regularities within a KB, on the Web, or in the world, using knowledge of the task plus background knowledge to keep the set of hypotheses from growing too large. We express these observed regularities as symbolic hypotheses. We then try to confirm them: by looking for evidence, pro or con, either in our KB or in the world, or by asking the user.

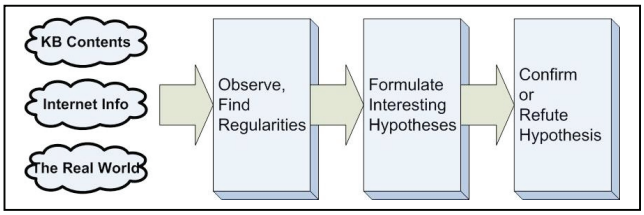
Generating hypotheses is (more or less) statistical in nature. We look for regularities (things that happen more than chance) and evidence. But by interposing a symbolic model, we can do a lot of sanity checking, relevance checking, and we have a clear symbolic statement that we can ask a user about. This research work is an example of learning that combines statistical and symbolic techniques.

We have begun to explore two instantiations of this general approach. In particular, we have looked at hypothesizing new relations among entities in a KB and hypothesizing new categories or types of entities in a KB.

Approach

Our technique specifies several steps in the learning process. The first is to recognize patterns that may be useful in an application. For example, in a conference planning application, we may note the pattern "Barbara has rejected 3 Monday meetings in a row" – a potentially useful observation in this domain. Then, we must recognize that we do not know what might explain this pattern ("for some reason Mondays are bad for Barbara"). Next, we formulate an explicit hypothesis that may explain this pattern ("Barbara does not like Monday meetings").

Now we can look in the KB to see if this hypothesis agrees with or conflicts with any known facts. We can also check if analogous hypotheses have been confirmed ("John and Fred have confirmed that they do not like Monday meetings"). Finally, we attempt to confirm our hypothesis that Barbara doesn't like Monday meetings by simply asking the user or perhaps by searching the Web. We may find that the hypothesis is incorrect or needs to be refined (e.g. "Barbara doesn't like to meet on Monday *mornings*").



There are many ways to formulate hypotheses. Our current implementation focuses on generating hypotheses by introspection: looking within the KB for regularities that correspond to *implicit* knowledge. In particular, we look for regularities that will be useful if they are made explicit. The two types of implicit knowledge we currently identify are implicit *relations* between entities and implicit *types* of entities.

To hypothesize implicit relations, we calculate statistics about the entities and relations in the KB, and use naïve Bayes to find the likelihood that two entities are related by a particular relation. To hypothesize new types of entities (concept formation), we identify elements that share certain characteristics and propose a description of the new type in a manner similar to work on learning symbolic concept hierarchies (Fisher 1987).

Experiments

To evaluate this approach we used two datasets: a sample of a computer science publication database (DBLP) (1500 examples), and the UCI 'zoo' dataset. We use DBLP to evaluate relation learning and use the zoology data to evaluate concept formation. The knowledge representation system is Scone (Fahlman 2006). For more information on

the data and KB system, please refer to the first author's Web site.

To evaluate our relational learner we use a leave-one-out technique where a relation is momentarily "forgotten" about. Then we try to reconstruct the missing knowledge. To evaluate our concept formation technique we remove the provided hierarchy and attempt to rebuild it from only the instances. We report accuracies for when the correct entity is identified in the top-1, top-2, top-3, etc. of the likelihood-ranked list of hypotheses. The "count" column is the number of instances of each relation.

Relation	Top1	Top2	Top3	Top5	Top10	Top20	Count
Overall	13.41	25.24	34.47	42.17	46.54	48.37	4209
authorOf()	12.82	24.15	33.86	41.33	44.07	45.90	3830
editorOf()	15.44	21.48	28.19	40.94	60.40	60.40	149
publisherOf()	15.38	15.38	15.38	15.38	100	100	65
schoolOf()	44.16	44.16	44.16	44.16	46.75	46.75	77
publSeries()	1.79	100	100	100	100	100	56
journalOf()	16.67	16.67	36.67	100	100	100	30

Table 1: Results of learning relations (percent)

These results show that often the correct answers do show up near the top of the rankings even though there may be hundreds of alternatives. We would expect to see better precision if the KB were larger thus providing more evidence (and better probability estimates). These are preliminary proof-of-concept results that we will be comparing to purely statistical approaches in future work.

An example of a correct prediction made, the paper *Experiments with Dispatching in a Distributed Object System* by *Farshad Nayeri* and *Benjamin Hurwitz* published at *GTE Laboratories Incorporated* as a tech report in July, 1993. After author *Hurwitz* was dissociated from the paper, we identified him as a likely co-author because he has published other papers with Nayeri.

For concept formation in the zoology domain, the data consists of 101 species along with 17 attributes of each. The data classifies these into seven groups: mammal, bird, reptile, fish, amphibian, insect, and mollusk.

Interpreting the results of this task is tricky because we allow multiple inheritance. A fairly strict interpretation of the result shows this method to be correct 86% of the time. This is competitive with other clustering techniques.

The advantage to doing this in a symbolic framework is that we can now apply background knowledge and reasoning to improve our initial clustering. Also, we can now present our initial clustering, along with the features that characterize each class, to a human collaborator for editing and fix up. For example, the learned mammal class includes "doesn't have fins" which we can easily remove to include dolphin, porpoise, seal and sea lion.

Related Research

This research is closely related to earlier work in the data mining community that attempted to use a KB to help a user to interactively mine databases, for example INLEN (Michalski et al., 1992). That research led to more recent work on inductive databases which is summarized by Kaufman (2005).

This earlier work did not have the Web as a resource for mining, seeking evidence, and confirming hypotheses. We intend to use some of the Web techniques that have been demonstrated by Etzioni (2005). Our work on inferring new types in a KB is also closely related to work by Fisher (1987) and more recent clustering algorithms.

Conclusion

This research shows how we can learn new types of entities and new relations entirely within a KB, and so take advantages of the built in reasoning and expressive power of the knowledge representation. We have shown how statistical and symbolic learning can be combined in a KB, and tested our ideas, in a preliminary way, on two datasets.

With this research we intend to demonstrate how background knowledge along with a reasoning system benefit learning and discovery for problems where purely numerical techniques have reached a plateau. We intend to test this claim by comparing our technique with zero-knowledge versions. We also plan to study what kinds of background knowledge help certain types of problems.

Achieving this goal will be relevant to the community because we will be able to surpass current performance on tasks that benefit from additional knowledge. It will also simplify the process of attacking new problems by allowing us to using existing techniques with a little hand-crafted knowledge rather than having to develop new techniques for each problem.

Acknowledgements

This work is supported in part by the Defense Advanced Research Projects Agency (DARPA) under contract number NBCHD030010, and a generous research grant from Cisco Systems Inc.

References

Etzioni, Oren, et al. (2005). Comprehensive Overview of KnowItAll. *Artificial Intelligence*, 165(1):91-134.

Fahlman, Scott E. (2006). Marker-Passing Inference in the Scone Knowledge-Base System. In *KSEM*. Guilin, China.

Fisher, D. H. (1987). Knowledge Acquisition via Incremental Conceptual Clustering. Machine Learning 1987 Volume 2, pp. 139-172.

Kaufman, K. A.; and Michalski, R. S. From Data Mining to Knowledge Mining. *Handbook in Statistics* 24: 47-75.

Michalski, R. S.; Kerschberg, L.; Kaufman, K. A.; and Ribeiro, J. S. 1992. Mining for Knowledge in Databases: the Inlen Architecture, Initial Implementation and First Results. *Journal of Intelligent Information Systems* 1: 85-113.