Which background knowledge is relevant?*

Stassa Patsantzis and Stephen H. Muggleton[1]
Imperial College London
e.patsantzis17@imperial.ac.uk, s.muggleton@imperial.ac.uk.

Imperial College London

Abstract. Selection of appropriate background knowledge is critical within the application of Inductive Logic Programming. Traditionally such selection is entirely dependent on human beings, who provide primitive predicates to be used by an ILP system for formulating hypotheses. In this paper we consider issues relating to possible automatic selection of background definitions from a large library of pre-existing predicates. In particular, we consider the effect of the *generality* of background definitions on error. In our experiments we introduce randomly defined extensional background predicates with varying levels of generality and measure the effects on omission and commission errors in the family relation domain. Our results indicate increasing generality of background predicates leads to a sharp increase in commission errors with a corresponding rapid decrease in omission errors. In further work we aim to investigate Meta-Interpretive Learning systems which order the selection of background and invented predicates based on their generality.

1 Introduction

A key distinguishing feature of Inductive Logic Programming[3, 6], in terms of Machine Learning, is the ability for users to provide learners with domain knowledge. However, the selection of such background knowledge can be time-consuming for users and also potentially error prone for the learning system. Insufficient provision of background knowledge leads to the learner failing to identify consistent hypotheses. Over-provision of background predicates leads to excessive search and overfitting to the training data. In this paper we consider a setting in which a large library of background predicates is made available. Learning is assumed to consist of two phases: 1) selection of a minimal set of background predicates B deemed relevant to the given examples E and 2) construction and testing of a hypothesis H from B and E.

This paper is organised as follows. Section 2 describes related work on identifying relevant features and background knowledge. In Section 3 we introduce the Meta-Interpretive Learning (MIL) setting and the definition of generality used

* The second author acknowledges support from his Royal Academy of Engineering/Syngenta Research Chair at the Department of Computing at Imperial College London.

in the experiments. Section 4 describes experiments involving varying the generality of randomly generated extensional background predicates. Lastly, Section 5 provides conclusions and suggestions for further work.

2 Related work

Feature subset selection In statistical machine learning and data mining, relevance is primarily discussed in the context of *Feature (Subset) Selection* (FSS) [2], whereby a subset of the features in a dataset are selected for training. The ultimate purpose of this selection is to minimise the dimensionality of the data in order to increase performance and improve training times. This is typically achieved by looking for correlations between features and the correct value of a predicted variable, then selecting those features with the highest such correlation, deemed to be the most relevant to the predicted variable. Additionally, it is possible to look for correlations between relevant features in order to detect and reduce *redundancy*.

Relevance selection in ILP In [7] Srinivasan conjectures that ordering background knowledge predicates according to their relevance to the learning task at hand can improve performance, measured by predictive accuracy and training time. The authors do not attempt a formal definition of "relevance" and instead enlist domain experts to provide a "hand-crafted" partial ordering of background predicates from two datasets, *mutagenesis* and *carcinogenicity*. The partial orderings selected this way are used to generate total orderings, of which one (of only two in each case) is selected for each domain. An incremental procedure is then described by which, during training, predicates are added to an initially empty background set according to the selected total ordering. The results of this "informed" training procedure are compared to a) a procedure using the entire background knowledge at once and b) an incremental procedure that randomly adds predicates to the background (ie without respecting the selected relevance ordering).

3 Framework

MIL framework We assume a Meta-Interpretive Learning [5] setting in which the learning algorithm is provided with Metarules M in the form of second-order definite clauses, first-order Background Knowledge B consisting of definite clause definitions and Examples E consisting of ground unit literals. Based on B and E the learner generates a hypothesis H in the form of a definite program such that $M \models H$ and $B, H \models E$.

Generality Assume Q^n to be the definition of a first-order predicate Q of arity n. We define the generality of Q^n as follows.

Definition 1. Generality of Q^n. *The generality of Q^n is*

$$g(Q^n) = Pr(Q(x^n)|random\ x^n)$$

4 Experiment results

In this section we present the results of experiments performed to determine the effect of generality on Errors of Commission and Errors of Omission.

4.1 Materials

A Prolog program was written to generate a list of ground atoms for each background predicate Q^n by including each possible atom in Q^n with probability $g(Q^n)$. The value of $g(Q^n)$ was assigned according to the following sequence of *generality classes*.

$$\langle 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 \rangle$$

Experiments were performed with Metagol [4, 5] as the learning system and a small dataset of family relations as training data, including atoms of father/2 and mother/2 as background knowledge and positive and negative examples of the "grandfather" relation.

4.2 Methods

Each experiment consisted of ten steps, one for each generality class listed in section 4.1. In each step a new background predicate was generated by the program described in section 4.1, with a random string as a symbol and constants taken from the Herbrand universe of the positive and negative examples of the target concept. In each step, atoms from the Herbrand base of the generated predicate were selected with probability equal to the corresponding generality class, thus controlling the generality of background knowledge.

In each step, 1000 cycles of training and evaluation were performed and the results of evaluation averaged over all cycles, to yield the error value for the generality class corresponding to that step. Hypotheses in each cycle were learned on a random sample of 2% of the positive examples and evaluated on the remaining positive and all negative examples. Evaluation therefore amounted to Monte Carlo cross-validation by random subsampling over the positive examples.

4.3 Results

The results shown in Figure 1 show that a) Errors of Commission increase and b) Errors of Omission decrease, as the generality of background predicates increases.

5 Conclusions and further work

This paper initiates a discussion of the problems with learning from large and possibly partially irrelevant background knowledge and presents some early results suggesting a relation between the validation error of hypotheses and the generality of background predicates used to construct them.

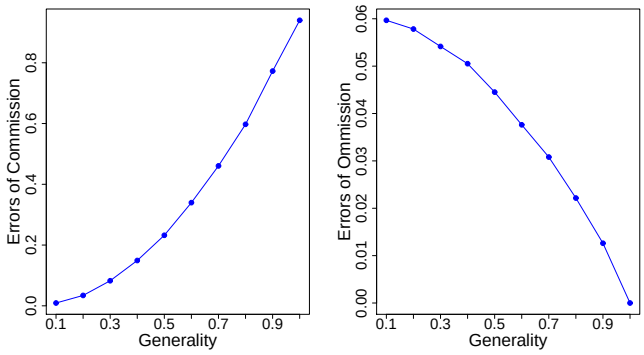

Fig. 1. Effect of varying the generality of randomly chosen background predicates on a) commission and b) omission errors.

5.1 Discussion

Informally, the generality of a predicate represents the probability that it is true for any atom chosen uniformly at random from its Herbrand base. A predicate that is always true, given any set of arguments, is maximally general, whereas one that is never true is minimally general. In the context of an ILP learning session a maximally general hypothesis would never be valid: although it would accept every positive example, it would also accept all negative ones! Conversely, a hypothesis with minimal generality would reject all positive examples, even though it would correctly reject all negative ones also. Further from these extremes of generality, more general hypotheses would correctly accept more positive examples but also incorrectly accept more negative examples and vice-versa. This inverse relation between generality and the correct or incorrect coverage of positive and negative examples, respectively, should manifest as an out-of-sample rate of Errors of Commission increasing with generality. This effect should be the most evident in a cross-validation setting when training with positive examples only: including negative training examples would cause some overly-general hypotheses to be rejected as invalid during training, before they can be evaluated on held-out data in a validation step.

The experiment results listed in section 4 are consistent with the above intuition regarding generality, particularly in the case of positive-only training examples. In figure 1 as the generality of hypotheses increases, more and more negative examples are falsely accepted, increasing the Errors of Commission on the out-of-sample data (which includes negative examples). At the same time, fewer and fewer positive examples are falsely rejected, reducing the Errors of Omission.

The relation between generality and both types of error is striking, but note that in figure 1 the Errors of Omission are never very high whereas Errors of

Commission essentially saturate. It's possible that this difference in the magnitude of the two types of error is a result of the imbalance in the numbers of positive vs. negative examples- there were only 6 positive examples of the target concept, but 78 negative ones in the dataset, so there were more chances for false positives than false negatives. Datasets with more balanced examples may yield higher rates of Errors of Omission.

5.2 Future work

Our experimental results suggest that a valid hypothesis would have to be somewhere between minimum and maximum generality, but closer, indeed *as close as possible*, to the minimum- in other words, sufficiently general to cover all positive examples but not as general as to cover any negative examples. This intuition suggests an ordering of the hypothesis space where the least general hypotheses are visited before the most general ones. Combined with an Occamist bias, as in Metagol's iterative deepening search, whereby shorter hypotheses are visited before longer ones, this would yield a procedure where the shortest, least general hypotheses are examined earlier in the search. We are currently working on an implementation of such a generality-ordered search facility for Metagol.

The question now naturally arises of how to measure the generality of a hypothesis, in order to impose such an ordering. One way is of course to estimate the generality of a hypothesis by sampling from its Herbrand base, but this would require a costly step inserted into an already computationally expensive learning procedure. A more attractive alternative is to calculate the generality of a hypothesis from the generality of its literals, in other words, from the predicates in the background knowledge-base. The generality of each background predicate need only be determined once by sampling from its Herbrand base in a preprocessing step, outside of any critical loops of the learning procedure. These values can then be stored in memory and used for a closed-form calculation during training. Additionally, background predicates can be ordered by their generality to ensure that the least general ones are examined first during training.

But how to calculate the generality of a hypothesis? Fortunately, MIL imposes a strong language bias on learned hypotheses, in the form of second-order metarules entailing hypotheses. Therefore, only a calculation for each metarule clause is needed, rather than for arbitrary clause structures. The necessary calculations can be reduced even further by concentrating on the Turing-complete H^2_2 class of hypotheses and only considering the two metarules that are known to be sufficient to construct any hypothesis in this class, according to [1], the *Inverse* and *Chain* metarules:

$$P(A,B) \leftarrow Q(B,A) \qquad \text{(Inverse)}$$
$$P(A,B) \leftarrow Q(A,C), R(C,B) \qquad \text{(Chain)}$$

Suppose that a hypothesis, H consists of a single clause, of the Inverse metarule. In that case, the generality of H must be the generality of its single body literal. The generality of a clause of the Chain metarule is harder to define. We are directing our future work towards this latter definition.

References

1. Andrew Cropper and Stephen H Muggleton. Logical minimisation of meta-rules within Meta-Interpretive Learning. In *In Proceedings of the 24th International Conference on Inductive Logic Programming*, pages 65–78, 2015.
2. Pat Langley. Selection of Relevant Features in Machine Learning. *In Proceedings of the AAAI Fall Symposium on Relevance*, pages 140–144, 1994.
3. S.H. Muggleton. Inductive Logic Programming. *New Generation Computing*, 8(4):295–318, 1991.
4. S.H. Muggleton, D. Lin, N. Pahlavi, and A. Tamaddoni-Nezhad. Meta-interpretive learning: application to grammatical inference. *Machine Learning*, 94:25–49, 2014.
5. S.H. Muggleton, D. Lin, and A. Tamaddoni-Nezhad. Meta-interpretive learning of higher-order dyadic datalog: Predicate invention revisited. *Machine Learning*, 100(1):49–73, 2015.
6. S.H. Muggleton, L. De Raedt, D. Poole, I. Bratko, P. Flach, and K. Inoue. ILP turns 20: biography and future challenges. *Machine Learning*, 86(1):3–23, 2011.
7. Ashwin Srinivasan, Ross D King, and Michael E Bain. An Empirical Study of the Use of Relevance Information in Inductive Logic Programming. *Journal of Machine Learning Research*, 4:369–383, 2003.