Learning Higher-Order Logic Programs Through Abstraction and Invention

Andrew Cropper and Stephen H. Muggleton
Imperial College London
United Kingdom
{a.cropper13,s.muggleton}@imperial.ac.uk

Abstract

Many tasks in AI require the design of complex programs and representations, whether for programming robots, designing game-playing programs, or conducting textual or visual transformations. This paper explores a novel inductive logic programming approach to learn such programs from examples. To reduce the complexity of the learned programs, and thus the search for such a program, we introduce higher-order operations involving an alternation of Abstraction and Invention. Abstractions are described using logic program definitions containing higher-order predicate variables. Inventions involve the construction of definitions for the predicate variables used in the Abstractions. The use of Abstractions extends the Meta-Interpretive Learning framework and is supported by the use of a user-extendable set of higher-order operators, such as map, until, and ifthenelse. Using these operators reduces the textual complexity required to express target classes of programs. We provide sample complexity results which indicate that the approach leads to reductions in the numbers of examples required to reach high predictive accuracy, as well as significant reductions in overall learning time. Our experiments demonstrate increased accuracy and reduced learning times in all cases. We believe that this paper is the first in the literature to demonstrate the efficiency and accuracy advantages involved in the use of higher-order abstractions.

1 Introduction

Inductive Programming (IP) [Gulwani et al., 2015] is a form of machine learning which aims to learn programs from examples given background knowledge (BK). To illustrate this form of machine learning, consider teaching a robot to pour tea and coffee for all place settings at a table. For each setting there is an indication of whether the associated guest prefers tea or coffee. Figure 1 shows an example in terms of an initial state (Figure 1a) and final state (Figure 1b).

Now consider learning a general strategy for the task from a set of such examples. Given that there may be an arbitrary number of place settings, existing approaches to IP, such as Meta-Interpretive Learning (MIL) [Muggleton et al., 2015; Cropper and Muggleton, 2015a], would learn a recursive strategy, such as that shown in Figure 1c. In this paper, we extend the MIL framework to support learning theories with higher-order constructs, such as map, until, and ifthenelse. In this approach, an equivalent yet more compact strategy can be learned, as in Figure 1d. This is implemented in a system called MetagolAI which uses a form of interpreted BK to learn programs through a sequence of interleaved Abstraction

[The first author thanks the BBSRC and Syngenta for his PhD Case studentship. The second author acknowledges support from his Royal Academy of Engineering/Syngenta Research Chair at the Department of Computing at Imperial College London.]
and Invention steps (see Figure 1e). We show that the compactness of such definitions leads to substantially improved predictive accuracy and significantly reduced learning time.

The paper is organised as follows. Section 2 discusses related work. Section 3 describes the theoretical framework for the augmented form of MIL involving Abstraction and Invention, together with a sample complexity result for the new representation. Section 4 describes Metagol_AI, including changes to the meta-interpretive learner required to support Abstraction and Invention. Section 5 details three experiments in which predictive accuracies and learning times for Metagol_AI are compared with and without higher-order BK. In each case, a substantial increase in predictive accuracy is achieved when the higher-order BK is included, in accordance with the sample complexity result from Section 3. Finally, Section 6 summarises the outcomes and discusses further work.

2 Related work

Interest in IP has grown recently, partially due to successful applications in real-world problems, such as end-user programming [Gulwani, 2014a] and computer education [Gulwani, 2014b]. IP approaches can be classified as either task-specific or general-purpose. Task-specific approaches focus on learning programs for a specific domain and are often restricted to specific data types, such as numbers [Singh and Gulwani, 2012] and strings [Gulwani, 2011; Wu and Knoblock, 2015]. By contrast, the MIL framework is general-purpose, and has been used in a variety of problems including grammar induction [Muggleton et al., 2014b], string transformations [Lin et al., 2014], and extracting information from markup files [Cropper et al., 2015].

MagicHaskeller [Katayama, 2008] is a general-purpose IP system which learns Haskell functions by selecting and instantiating higher-order functions from a pre-defined vocabulary. In contrast to MagicHaskeller, MIL supports predicate invention and learning explicitly recursive programs. Igor2 [Kitzelmann, 2007] also learns recursive Haskell programs and supports auxiliary function invention but is restricted in that it requires the first k examples of a target theory to generalise over a whole class. Esher [Albarghouthi et al., 2013] learns recursive programs but needs to query an oracle each time a recursive call is encountered to ask for examples. The L2 system [Feser et al., 2015] synthesises recursive functional algorithms, but the hypotheses learned by L2 are not directly executable. By contrast, Metagol_AI learns Prolog programs.

Section 5 includes experiments in learning robot strategies [Cropper and Muggleton, 2015a]. Various machine learning architectures support the construction of strategies, including the SOAR architecture [Laird, 2008], reinforcement learning [Sutton and Barto, 1998], and action learning in inductive logic programming (ILP) [Moyle and Muggleton, 1997; Otero, 2005]. This work differs from most of these approaches in that the Metagol_AI learns human-readable Prolog programs.

Early work in ILP [Flener and Yilmaz, 1999] considered using schema to specify the overall form of recursive programs to be learned. By contrast, the use of abstraction described in this paper involves higher-order definitions which treat predicate symbols as first-class citizens. This approach supports a form of abstraction which goes beyond typical first-order predicate invention [Saitta and Zucker, 2013] in that the use of higher-order definitions combined with meta-interpretation drives both the search for a hypothesis and predicate invention, leading to more accurate and compact programs.

Lloyd [Lloyd, 2003] advocates using higher-order logic in the learning process, though the approach was more strongly allied to learning functional programs, and did not support predicate invention.

3 Theoretical framework

The sets of constants, predicate symbols and first and second-order variables are denoted C, P, V1 and V2. Elements of V1 and V2 can bind to elements of C and P respectively.

3.1 Higher-order definitions

**Definition 1 (Higher-order definite clause)** A higher-order definite clause is a well-formed formulae \( \forall \tau \exists (s_1, \ldots, s_m) \leftarrow \ldots, Q_i(t_1, \ldots, t_n) \ldots \) where \( \tau \subseteq V_1 \cup V_2 \) and \( P, Q_i, s_j, t_k \in P \cup C \cup V_1 \cup V_2 \).

**Definition 2 (Higher-order definite definition)** A higher-order definite definition is a set of higher-order clauses which all have the form \( \forall \tau p(s_1, \ldots, s_m) \leftarrow \ldots \) where \( \tau \subseteq V_1 \cup V_2 \) and \( p \in P \).

The clauses in Figure 2a comprise a higher-order definition.

3.2 Abstractions and inventions

**Definition 3 (Abstraction)** An abstraction is a higher-order definite clause having the form \( \forall \tau p(s_1, \ldots, s_m) \leftarrow q(v_1, \ldots, v_n, r_1, \ldots, r_o) \ldots \) where \( \tau \subseteq V_1 \cup V_2 \) and \( p, q, r_1, \ldots, r_o \in P \) and \( v_1, \ldots, v_n \in V_1 \).

Within Computer Science code abstraction [Cardelli and Wegner, 1985] involves hiding complex code to provide a simplified interface for users to select key details. In this paper Abstractions contain one atom in the body which references a higher-order predicate, as shown in Figure 2b. The second-order arguments of until are grounded to predicate symbols.

**Definition 4 (Invention)** In the case background knowledge \( B \) is extended to \( B \cup H \), where \( H \) is a set of higher-order definite definitions, we call predicate \( p \) an Invention if \( p \) is defined in \( H \) but not in \( B \).

Within this paper Abstractions are used by a meta-interpreter to generate Inventions (Figure 2c).
3.3 Meta-Interpretive Learning

Given background knowledge \( B \) and examples \( E \) the aim of a MIL system is to learn a hypothesis \( H \) such that \( B, H \models E \), where \( B = B_p \cup M \), \( B_p \) is a set of compiled Prolog definitions and \( M \) is a set of metarules (see Figure 3). MIL [Muggleton et al., 2014b; 2015; Cropper and Muggleton, 2015b; Muggleton et al., 2014a] is a form of ILP based on an adapted Prolog meta-interpreter. A standard Prolog meta-interpreter proves goals by repeatedly fetching first-order clauses whose heads unify with the goal. By contrast, a MIL learner proves goals by fetching higher-order metarules (Figure 3) whose heads unify with the goal. The resulting meta-substitutions are saved, allowing them to be used as background knowledge by substituting them into corresponding metarules.

We now consider the ratio of these bounds in the case \( n \gg p \).

Proposition 1 (Ratio of unabstracted and abstracted bounds) Given \( m, m_A \) are the bounds on the number of training examples required to achieve error less than \( \epsilon \) with probability at least \( 1 - \delta \) and \( n_A \) are the numbers of clauses in the minimum expression of the target theories in these cases then the ratio \( m : m_A \) approaches \( n : n_A \) in the case \( n \gg p \).

Proof. Since \( n \gg p \) it follows \( m : m_A \approx \frac{n}{n_A} \), where \( n_A \) are the bounds on the number of clauses required to express abstracted hypotheses. For instance, in Figure 1 the use of un\( \text{til} \) and if\( \text{then} \) reduces the hypothesis size by one clause each. Thus the minimal hypothesis reduces from six clauses to four leading to a sample complexity reduction of \( 3 : 2 \). Figure 4 tabulates higher-order predicates with corresponding clause reductions.

<table>
<thead>
<tr>
<th>HO predicate</th>
<th>Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>until</td>
<td>1</td>
</tr>
<tr>
<td>if( \text{then} )</td>
<td>1</td>
</tr>
<tr>
<td>map</td>
<td>1</td>
</tr>
<tr>
<td>filter</td>
<td>2</td>
</tr>
</tbody>
</table>

3.4 Abstracted Meta-Interpretive Learning

We extend the MIL framework by assuming the background knowledge \( B = B_p \cup B_t \cup M \), where \( B_p \) consists of compiled Prolog code (compiled BK), \( B_t \) consists of higher-order definitions (interpreted BK), and \( M \) is a set of metarules. The existence of \( B_p \) supports efficient execution of background knowledge, but makes the substitution of meta-variables inaccessible to the meta-interpreter for inventing new predicates. By contrast, the existence of \( B_t \) allows the meta-interpreter to efficiently interleave Abstraction and Invention.

3.5 Language classes, expressivity and complexity

Metarules limit the standard class for the hypothesis space. For instance, the Chain rule in Figure 3 restricts clauses to be definite with two body atoms and a predicate arity of two. This corresponds to the language class \( H_2^2 \). In [Lin et al., 2014] it was shown the number of \( H_2^2 \) programs expressible with \( n \) clauses is \( O(|M|^n p^{3n}) \). The result below updates this bound for the abstracted MIL framework.

**Lemma 1 (Number of abstracted \( H_2^2 \) programs of size \( n \)).**
Given \( p \) predicate symbols, \( |M| \) metarules, and abstractions each with at most \( k \) second-order variables the number of \( H_2^2 \) programs expressible with \( n \) clauses is \( O(|M|^n p^{2+k} n) \).

**Proof.** Since each abstraction has at most \( k \) second-order variables the number of clauses \( S_p \) which can be constructed from an \( H_2^2 \) metarule given \( p \) predicate symbols is at most \( \max(p^k, p^{2+k}) = p^{2+k}. \) The set of such clauses \( S_{m,p} \) has cardinality at most \( |M|^p p^{2+k}. \) It follows the number of logic programs constructed from a selection of \( n \) rules chosen from \( S_{|M|,p} \) is at most \( (|M|^p p^{2+k})^n \). The bound follows.

We use this result to develop sample complexity results for unabstracted versus abstracted MIL.

**Theorem 1 (Sample complexity of unabstracted MIL).**
Unabstracted MIL has a polynomial sample complexity of \( m \geq n \ln(|M|)+p \ln(3n)+\ln \frac{1}{\delta} \).

**Proof.** According to the Blumer bound [Blumer et al., 1989] the error of consistent hypotheses is bounded by \( \epsilon \) with probability at least \( 1 - (1 - \delta) \) once \( m \geq \frac{\ln|H|+\ln \frac{1}{\delta} + \ln(c)}{\epsilon^2} \), where \( |H| \) is the size of the hypothesis space. From [Lin et al., 2014] \( |H| = c(|M|^n p^{3n} + d \right) where \( c, d \) are constants. Applying logs and substituting gives \( m \geq \frac{\ln(|M|)+p \ln(3n)+\ln \frac{1}{\delta}}{\epsilon^2} \).

**Theorem 2 (Sample complexity of abstracted MIL).**
Abstracted MIL has a polynomial sample complexity of \( m \geq n \ln(|M|)+p \ln(2k+1)n+\ln \frac{1}{\delta} \).

**Proof.** Analogous to Theorem 1.

We now consider the ratio of these bounds in the case \( n \gg p \).

**Proposition 1 (Ratio of unabstracted and abstracted bounds).**

Given \( m, m_A \) are the bounds on the number of training examples required to achieve error less than \( \epsilon \) with probability at least \( 1 - \delta \) and \( n, n_A \) are the numbers of clauses in the minimum expression of the target theories in these cases then the ratio \( m : m_A \) approaches \( n : n_A \) in the case \( n \gg p \).

**Proof.** Since \( n \gg p \) it follows \( m : m_A \approx \frac{n}{n_A} \), where \( n_A \) are the bounds on the number of clauses required to express abstracted hypotheses. For instance, in Figure 1 the use of un\( \text{til} \) and if\( \text{then} \) reduces the hypothesis size by one clause each. Thus the minimal hypothesis reduces from six clauses to four leading to a sample complexity reduction of \( 3 : 2 \). Figure 4 tabulates higher-order predicates with corresponding clause reductions.

<table>
<thead>
<tr>
<th>HO predicate</th>
<th>Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>until</td>
<td>1</td>
</tr>
<tr>
<td>if( \text{then} )</td>
<td>1</td>
</tr>
<tr>
<td>map</td>
<td>1</td>
</tr>
<tr>
<td>filter</td>
<td>2</td>
</tr>
</tbody>
</table>

4 Metagol\(_{AI} \)

Metagol\(_{AI} \) extends Metagol\(^1 \), an existing MIL implementation, to support Abstractions and Invention by learning with interpreted BK. Figure 5 shows the implementation of Metagol\(_{AI} \) as a generalised meta-interpreter [Muggleton et al., 2015], similar in form to a standard Prolog meta-interpreter.

**Background knowledge** The key difference between Metagol\(_{AI} \) and Metagol is the introduction of the second prove \(_{aux} \) clause in the meta-interpreter, denoted in boldface.

\(^1\)https://github.com/metagol/metagol
null

choice is incorrect for a random subset of $k$ drinks. The robot can perform the following fluents and actions (details omitted for brevity) defined as compiled BK: $\text{at\_end}/1$, $\text{wants\_tea}/1$, $\text{wants\_coffee}/1$, $\text{move\_left}/2$, $\text{move\_right}/2$, $\text{turn\_cup\_over}/2$, $\text{pour\_tea}/2$, and $\text{pour\_coffee}/2$.

Results Figure 7a shows that Metagol$_{AI}$ learns more accurate programs than Metagol, refuting null hypothesis 1. Figure 7b shows that Metagol$_{AI}$ learns programs quicker than Metagol, refuting null hypothesis 2. Figure 1 shows example programs learned by Metagol (c) and Metagol$_{AI}$ (d). Although both programs are general and can handle any number of guests and any assignment of drink preferences, program (b) is smaller because it uses the higher-order abstractions until and ifthenelse. This compactness affects predicate accuracies because whereas Metagol$_{AI}$ can find solutions in the allocated time, Metagol struggles because the solutions are too long.

![Predictive accuracies](image1.png) ![Learning times](image2.png)

(a) Predictive accuracies  (b) Learning times

Figure 7: Robot waiter experiment results

5.2 Chess strategy

Programming robust chess playing strategies is an exceptionally difficult task for human programmers [Bratko and Michie, 1980]. Consider the concept of maintaining a wall of pawns to support promotion [Harris, 1988]. In this case, we might start by trying to inductively program the simple situation in which a black pawn wall advances without interference from white. Having constructed such a program one might consider using negative examples involving interposition of white pieces to deal with exceptional behaviour. Figure 8 shows such an example, where in the initial state pawns are at different ranks, and in the final state all the pawns have advanced to rank 8, but the other pieces have remained in the initial positions. In this experiment, we try to learn such strategies.

![Initial state](image3.png) ![Final state](image4.png)

(a) Initial state  (b) Final state

Figure 8: Chess initial/final state example

Materials The state is a list of pieces, where a piece is denoted as a triple of the form $(\text{Type}, \text{Id}, X/Y)$, where $\text{Type}$ is the type ($\text{king}=k$, $\text{pawn}=p$, etc.), $\text{Id}$ is a unique identifier, and $X/Y$ is the position. We generate positive examples as follows. For the initial state, we select a random subset of $n$ pieces from the interval $[2, 16]$ and randomly place them on the board. For the final state, we update the initial state so that each pawn finishes at rank 8. To generate negative examples, we repeat the aforementioned procedure but we randomise the final state positions, whilst ensuring that the input/output pair is not a positive example. We use the compiled BK shown in Figure 9.

```plaintext
at\_rank8((\_\_\_/8)).
\text{is\_pawn}/(p,\_\_\_).
\text{not\_pawn}/X:\:-\text{is\_pawn}/X).
\text{empty}/().
move\_forward/(\text{Type}, \text{Id}, X/Y1), (\text{Type}, \text{Id}, X/Y2)):\:-Y1 < 8,Y2 is Y1+1.
move\_forward/(A,B,\text{Id}):\:-append(Prefix,[(\text{Type}, \text{Id}, X/Y1)\text{Suffix}]),A),
Y1 < 8,Y2 is Y1+1.
append(Prefix,[(\text{Type}, \text{Id}, X/Y2)\text{Suffix}],B).
```

Figure 9: Compiled BK used in the chess experiment

Results Figure 10a shows that Metagol$_{AI}$ learns programs approaching 100% accuracy after two examples. By contrast, Metagol learns programs with around default accuracy. This result refutes null hypothesis 1. The semi-log plot in Figure 10b shows that Metagol$_{AI}$ learns programs quicker than Metagol, refuting null hypothesis 2. We can explain these results by looking at the sample programs learned in Figure 11. Metagol$_{AI}$ (b) learns a small higher-order program using the abstractions map and until. In this program, the map operation decomposes the problem into smaller sub-problems of finding how to move a single piece to rank 8. These sub-goals are solved by the chess1 predicate. By contrast, Metagol (a) learns a larger recursive and more specific first-order program.

![Predictive accuracies](image5.png) ![Learning times](image6.png)

(a) Predictive accuracies  (b) Learning times

Figure 10: Chess experiment results

5.3 Drop lasts

In this experiment, the goal is to learn a program droplasts which drops the last element from each sublist of a given list, a problem frequently used to evaluate IP systems [Kitzelmann, 2007]. Figure 12 shows input/output examples for this problem.
The experiments in this paper were largely related to the use of functional constructs, such as `map` and `reduceback`, within logic programs. However, we would like to investigate the use of relational constructs. For instance, consider the following higher-order definition of a closure.

\[
\text{closure}(P,X,Y) \leftarrow P(X,Y).
\]

This definition could be used to learn compact abstractions of relations such as the following.

\[
\text{ancestor}(X,Y) \leftarrow \text{closure}(\text{parent},X,Y).
\]

Moreover, the issue of how metarules might themselves be learned could be treated in a similar fashion using higher-order programs such as the following.

\[
\text{chain}(P,Q,R,X,Y) \leftarrow Q(X,Z), R(Z,Y).
\]

In summary we believe that the use of abstractions in machine learning provides an important new approach to the use of powerful programming constructs within IP. We believe that such approaches could have wide application in AI domains such as planning, vision, and natural language processing.