In: S. Muggleton (ed.), Proc of the 3rd Int. Workshop on Inductive Logic Programming, ILP-93, Bled, Slovenia.

A Comparative Study Of Structural Most Specific Generalizations Used In Machine Learning

Jörg-Uwe Kietz
German National Research Center for Computer Science, I3-KI,
P.O.Box 1316, D-5205 St.Augustin, Germany

October 3, 1997

Abstract

In this paper we compare the two main lines of research in learning most specific generalizations (MSG's) in a unifying framework. By reducing them to each other we show that even in some simple subset of first-order logic, the MSG grows exponentially in the number of examples. We then review two polynomial approaches, learning most specific Hornclauses without existential variables and learning most specific ij-determinate Hornclauses within this framework. We also show that ij-determinate Hornclauses are a maximal subset of Hornlogic, which is polynomial learnable, as the relaxation from ij-determinate Hornclauses to determinate Hornclauses lead to exponentially longer MSGs. Keywords: Inductive Logic Programming, Most Specific Generalization, Least General Generalization.

1 Introduction

Recently the limited expressiveness of attribute-based descriptions has lead to an increased interest in learning from logical descriptions which, however, leads to an increased complexity of the learning problem. In this paper, we study this trade-off between expressiveness of the knowledge representation formalism and the computational complexity of the resulting learning situation. We focus on the description and comparison of inductive learning of most specific generalizations (MSGs, also called least general generalizations, LGGs). MSGs are of special interest in machine learning, because they have some nice properties. They build the lower bound of Mitchell's Version Space [Mitchell 1982], so some of Mitchell's results hold for them. If there is no MSG of all positive examples, which do not cover also some negative examples, there is no consistent generalization of all examples at all. The importance of MSGs in inductive learning is reflected by the fact that there are several machine learning approaches which use MSGs or approximations of them [18, 26, 22, 6, 12, 8, 15, 5, 14, 10, 1, 16]. In this paper, we identify two main lines of this previous research and compare them in a unifying framework. Whereas one line of research regard learning from a logical point of view, the other line see it from a graph matching point of view. In this paper we show, that the main difference between these two points of view is at the implementation level! Whereas the logical view matches the predicates and retains as many variable bindings (object identities) as compatible with the predicate match, the graph matching view matches the objects and retains as many descriptors (predicates) as compatible with the object match. We will unify the approaches in our framework. As learning in Hornlogic is intractable in general [9], we will investigate the remainder of the paper in the identification of tractable subsets of Hornlogic.

2 A Framework for Induction

For inductive learning systems, the input typically consists of examples E given in a language LE and possibly background knowledge B in the language LB. The task of an inductive system then consists of constructing in a language LH a hypothesis that "explains" the examples E. As there is usually more than one explanation, we need a preference order over the hypothesis language to select the "right" hypothesis (table 1). In the following, we will review the learning problems defined by Plotkin [18] and Helft [8] as prototypes for the logical approaches in this framework. We then switch our attention to the learning problem as defined by Haussler [5] as a prototype for the graph matching approaches. These three general approaches differ in LE and LH, but have in common that they do not use background knowledge. Whereas Plotkin and Helft are able to show, that there always exist a unique (up to equivalence) MSG, Haussler proves, that in his framework there are in the worst-case exponentially many different MSGs and searching a consistent concept in this space in an NP-hard problem. We will prove, that the unique MSG of Helft is not only more specific than every MSG of Haussler, but is in fact more specific than the union of all Haussler MSGs. This means, the MSG construction of Helft and Plotkin is a kind of breath-first search in the exponential concept space of Haussler. This insight enables us to give an example learning situation, where even the reduced MSG of Helft grows exponentially in the number of examples.

One way to avoid this intractability is to search only for heuristic approximations of the MSGs. Bisson's [1] KBG is an example of such a system.

The other way to avoid this exponential explosion is to restrict the hypothesis language. As this explosion is mainly due to problems with the existentially variables[1] one possibility is to use Hornclauses without existential variables - the so called constraint atoms of Page and Frisch [16]. They prove, that the MSG

[1] Existentially variables are the variables, which only occur in the body of the Hornclause

Given:

- background knowledge B in a Language LB
- positive and negative examples E = $E^+ \cup E^-$ in a language LE consistent with B (B,E $\not\models \Box$) and not a consequence of B (B $\not\models$ E).
- a hypothesis language LH.
- a preference partial-order $\leq$ on LH.

Find H $\subseteq$ LH, such that $\forall$h $\in$ H:

(I) (B,h,E $\not\models \Box$), h is consistent with B and E.

(II) (B,h $\models E^+$), h and B explain E^+.

(III) (B, h $\not\models E^-$), h and B doesn't explain E^-.

(IV) $\forall$h' $\in$ H: (h' $\leq$ h $\Rightarrow$ h $\leq$ h') h is a most preferred hypothesis

If $\leq$ is a generality order (h $\leq$ h' $\Rightarrow$ h' $\models$ h), then each h$\in$H is called a most specific generalization (MSG) of E with respect to LH, B and $\leq$.

Table 1: Logical Framework for Induction.

in this restricted hypothesis language grows only polynomially in the number of examples.

This result was subsumed by the polynomial MSG learning system GOLEM of Muggleton and Feng [14]. They use so called ij-determinate Hornclauses as hypothesis language, i.e. Hornclauses with existential variables, which value can be uniquely determined given the examples and the background knowledge, and which need not more that i relations to link them deterministically to a head variable.

Let us first recall some results concerning background knowledge and its incorporation into the learning process through the generality order $\leq$ or through saturation of the examples.

2.1 The effects of LB and $\leq$

In the general case, where LB is first order logic (FOL) and $\leq$ is implication, the learning problem is undecidable because consistency (B,h,E $\not\models \Box$) and the generality order itself (h $\leq$ h') are undecidable. This also holds if LH is FOL and B is empty [20]. This means that we have to restrict ourselves to less than FOL and implication if we want to come to a solution. For LB and/or LH, possible restrictions are function free logic, Hornclauses, generative clauses, linked clauses[2] or a combination of them. For the generality relation, possible restrictions are θ-subsumption and relative (or generalized) subsumption[3].

Plotkin [19] and Niblett [15] gives some results for Hornclause logic (LH as well as LB) and implication. If we have background knowledge, it may happen that no most specific generalization exists (H = {}): for each generalization h' there exists a more special one h. If we do not have background knowledge, h $\leq$ h' is decidable for Hornclauses [15], but there need not exist a unique MSG ($|$H$|\geq$1) of two Hornclauses.

The situation is much better for clausal logic and θ-subsumption. The learning problem is decidable and there exists a unique MSG ($\mid H \mid = 1$) reduced under equivalence. C_1 is equivalent to C_2 written $C_1 \equiv_\theta C_2$ iff $C_1 \leq_\theta C_2$ and $C_2 \leq_\theta C_1$ [18]. θ-subsumption is a weaker form of implication, i.e. if $C_1 \leq_\theta C_2$, then $C_1 \leftarrow C_2$, but not vice versa. In the general case, θ-subsumption between two clauses is NP-complete (Garey, Johnson, 1979), but Helft [8] states that he has an n^3 approximation, for reduction under $\equiv_\theta$ of function-free Hornclauses.

To incorporate background knowledge, Plotkin [19] and Buntine [2] define relative subsumption $\leq_B$ as subsumption relative to background knowledge B. They show that θ-subsumption and relative subsumption are equivalent if B is empty. But relative subsumption is also equivalent to θ-subsumption between saturated clauses[4]. Hence, it is equivalent to use background knowledge and generalized subsumption, and to use saturated examples and θ-subsumption without background knowledge, as long as we can assure that the saturation process will terminate.

As the choice of implication as generality order has several drawbacks, we will try θ-subsumption. θ-subsumption in general is NP-complete, but the result of Helft offers hope that we can learn MSGs without having to decide θ-subsumption in general.

2.2 The Logical Approach of Learning MSGs in clausal and Hornlogic

One of the first approaches of learning MSGs is from Plotkin [18]. This approach uses clausal logic as LH and LE, no background knowledge and θ-subsumption as preference order (table 2)[5].

[2] Generative clauses are clauses where all variables in the head also occur in the body (see [11, 14]). A clause is linked if all its literals are. A literal is linked if at least one of its arguments is. An argument is linked if either the literal is in the head of the clause or another argument in the same literal is linked [8].

[3] A clause C1 is θ-subsumed by a clause C2 (written C1 $\leq_\theta$C2), iff $C_1 \supseteq C_2\theta$ for some substitution θ. Plotkin [19] and Buntine [2] defines relative (generalized) subsumption $\leq_B$ as subsumption relative to some background knowledge.

[4] This follows immediately from Theorem 4.2 of Buntine [2] and the properties of saturation as described by Rouveirol [21]. If saturation is not finite, generalized subsumption is also undecidable.

[5] These results are generalized in [19] to learning with background knowledge expressed in clausal logic. He also studies the effects of several other preference orders, not based on

LB = {},   LH = LE = {C | C is a FOL clause} $C_1 \leq C_2 :\Leftrightarrow C_1 \supseteq C_2\theta$, for a substitution θ.

Table 2: Plotkin (1970) in our Logical Framework.

Plotkin's main results are:

- There always exists a MSG (|H|≥1). It is the empty clause, only if E^+ has no selection. A selection of a set of clauses $S = \{C_i \mid 1 \leq i \leq n\}$ is a set of literals $K = \{L_i = \phi(t_{1i}, \ldots, t_{mi}) \mid 1 \leq i \leq n, L_i \in C_i\}$, i.e. a selection of a set of clauses consists of one literal from each clause, all having the same predicate symbol and sign ϕ.
- The MSGs are unique up to equivalence ($\mid H_\equiv \mid \leq 1$), whereas $\equiv$ is the equivalence by $\leq$: $C_1 \equiv C_2 \Leftrightarrow C_1 \leq C_2 \wedge C_2 \leq C_1$ and $H_\equiv$ is the set of equivalence classes of H under $\equiv$.
- Each equivalence class in LH under $\equiv$ has a unique (up to the naming of variables) reduced member C. C is reduced if it is not equivalent to a proper subset of itself: D ⊂ C, D ≡ C ⇒ D = C.
- Plotkin further proves, that the $msg(e_1, e_2) = msg(e_2, e_1)$, and that $msg(e_1, e_2, \ldots, e_n) = msg(msg(\ldots(msg(msg(e_1, e_2), e_3)\ldots), e_n)$, i.e. the MSGs are order independent and incremental learning produces the same result as one-shot learning.

He also gives an algorithm that computes a member of H from E^+, i.e. a MSG. The algorithm as well as the length of the computed MSG is linear to the number of selections of E^+. It is easy to see that the number of selections is bounded by m^n, where m is the length of the longest clause of E^+ and n is the number of clauses of E^+. Plotkin says nothing about the length of the reduced MSG. But, we will show later on, that even the reduced MSG grows exponential in the number of examples.

Here is a simple example:

$$
\begin{aligned}
E^+ = \{ & \{+(e_1), \neg part(e_1, a), \neg small(a), \neg red(a), \\
& \quad \neg part(e_1, b), \neg big(b), \neg blue(b)\}, \\
& \{+(e_2), \neg part(e_2, c), \neg small(c), \neg blue(c), \\
& \quad \neg part(e_2, d), \neg big(d), \neg red(d)\}\}
\end{aligned}
$$

$$
\begin{aligned}
\text{msg}(E^+) = \{ & +(E), \neg part(E, X_1), \neg small(X_1), \\
& \neg part(E, X_2), \neg red(X_2), \\
& \neg part(E, X_3), \neg big(X_3), \\
& \neg part(E, X_4), \neg blue(X_4)\}
\end{aligned}
$$

generalization.

$B = \{\{t_1 \neq t_2\} \mid t_1, t_2$ are different terms in LE$\} \cup \{X \neq Y \leftarrow Y \neq X\}$,
LH = {C | C is a linked, function-free Hornclause},
LE = { C | C is a ground member of LH},
$C_1 \leq C_2 :\Leftrightarrow C_1 \leq_B C_2$

Table 3: Helft (1987) in our Logical Framework.

After Plotkin, there are several other approaches using MSGs in the logical view for learning, e.g. Kodratoff's and Ganascia's structural matching [12] and Vrain's OGUST [25]. Helft [8] is another more theoretical paper, which uses Plotkin's framework and restricts LH to linked, function-free Hornclauses and LE to ground clauses of LH. In addition to Plotkin, the difference of objects present in the examples is stated in form of inequality literals between terms in the MSG. At the level of our framework, this is equivalent to having an (in)equality theory in the background knowledge and using this background knowledge in the preference order through relative subsumption.

Let us look at the above example in Helfts framework:

$$\begin{aligned} B = \{&\{a \neq b\}, \{a \neq c\}, \{a \neq d\}, \{b \neq c\},\\ &\{b \neq d\}, \{c \neq d\}, \{e_1 \neq a\}, \{e_1 \neq b\},\\ &\{e_1 \neq c\}, \{e_1 \neq d\}, \{e_2 \neq a\}, \{e_2 \neq b\},\\ &\{e_2 \neq c\}, \{e_2 \neq d\}, \{e_1 \neq e_2\},\\ &\{X \neq Y, \neg Y \neq X\}\}\end{aligned}$$

E^+ as above.

$$\begin{aligned} \text{msg}(E^+) = \{&+(E), \neg X_1 \neq X_3, \neg X_2 \neq X_4\\ &\neg part(E, X_1), \neg small(X_1), \neg E \neq X_1,\\ &\neg part(E, X_2), \neg red(X_2), \neg E \neq X_2,\\ &\neg part(E, X_3), \neg big(X_3), \neg E \neq X_3,\\ &\neg part(E, X_4), \neg blue(X_4), \neg E \neq X_4\}\end{aligned}$$

As seen in section 3.2, instead of using the background knowledge in the generalization process it is equivalent at the knowledge level to saturate the examples and use θ-subsumption as preference order. This means that all of Plotkin's results also hold for Helft's learning problem as Plotkin's learning problem is a more general one.

Turning it the other way around, if the saturation process is not only finite, but a polynomial transformation, then a lower bound complexity result for learning Helft's MSGs also holds for learning Plotkin's MSGs, as Plotkin's MSGs contain Helft's as a special case. Clearly, the saturation process adds at most n*(n-1) linked literals to a clause, where n is the number of terms in the clause. So indeed it is a polynomial transformation.

2.3 The Graph Matching Approach of Learning MSGs

Starting with Winston [26], the other main line of computing MSGs was established. Other approaches in this line of research are Hayes-Roth's and McDermott's system SPROUTER [6, 7] and the work of Vere [22, 23, 24]. A formal description and a proof of the non PAC-learnability of these MSG approaches is done by Haussler [5]. This framework of learning MSGs forms the other main line of learning MSGs. Whereas Plotkin and Helft regard learning from a logical point of view, Haussler analyzes it from a graph matching point of view. Whereas the logical view matches the predicates and retains as many variable bindings (object identities) as compatible with the predicate match, the graph matching view matches the objects and retains as many descriptors (predicates) as compatible with the object match. We will unify the approaches in our framework. But, first let us look more formally at Haussler's framework.

The hypothesis language in Haussler's framework consists of so called existential conjunctive expressions over a fixed set of descriptors of unary attributes and binary relations (see also Michalski's [13] VL_1):

$C = \exists^* x_1, x_2, \ldots, x_r : f_1 \& f_2 \& \ldots \& f_s$

where $\exists^* x_1, x_2, \ldots, x_r$ means, there exists at least r different objects and the f_i $1 \leq i \leq s$ represent required attributes of an object or required relations between two different objects.

Such an expression is satisfied by all instances that consist of at least r different objects and there is a 1 to 1 mapping φ between the x_k $1 \leq k \leq r$ and r different objects of the instance, such that for all f_i $1 \leq i \leq s$ of C, if $f_i = f_i(x_k)$, then $\varphi(x_k)$ satisfies f_i, if $f_i = f_i(x_k, x_j)$, then $(\varphi(x_k), \varphi(x_j))$ satisfies f_i. The set of all instances that satisfy C is called the concept represented by C. An expression C_1 is more specific than an expression C_2 iff the concept represented by C_1 is a subset of the concept represented by C_2. The MSG of a set of instances E^+ is the set of all most specific expressions C_i, where E^+ is a subset of the concept of C_i. An expression C is consistent with $E = E^+ \cup E^-$, if the intersection of the concept of C and E^- is empty and E^+ is a subset of the concept of C.

Lets look at our example in this framework:

e_1: size(a)=small & color(a)=red &
size(b)=big & color(b)=blue

e_2: size(c)=small & color(c)=blue &
size(d)=big & color(d)=red

$msg(\{e_1, e_2\}) = \{$
$\exists^* x_1, x_2$: size(x_1)=small & size(x_2)=big,
$\exists^* x_1, x_2$: color(x_1)=blue & color(x_2)=red $\}$

As we have seen in this example, the MSG in this framework is not unique, but consists of several expressions with very different concepts.

This leads to the main result of Haussler: The problem of determining if there is an existential conjunctive concept consistent with a sequence of m examples over an instance space defined by n attributes (where m and n are variables) is NP-complete, even when there are no binary relations defined, each attribute is Boolean valued and each example contains exactly two objects. [5, pp. 17]

The description given so far seems not to fit very well into our framework, but we transform it into our logical framework. This allows us to relate the NP-completeness result of Haussler to the problem of learning MSGs in the logical framework.

In general, a n-place attribute with value $f(x_1, \ldots, x_n) = v$ can be transformed into a n-place predicate $v_f(x_1, \ldots, x_n)$ stating that the value is true for that relation, e.g. $distance(t_1, t_2) = far$ is transformed into $far_{distance}(t_1, t_2)$ or simply into $far(t_1, t_2)$ if this introduces no ambiguities. Now, we can define a transformation Γ for attributes, instances and existential conjunctive expressions into linked, function-free Hornclauses. $\Gamma(f(t_1) = v) = v_f(t_1)$ $\Gamma(f(t_1, t_2) = v) = v_f(t_1, t_2)$ An instance e consisting of m objects, described by n attributes over these objects:

$$e : f_1 \& \ldots \& f_n$$

is transformed by Γ into a ground, linked, function-free Hornclauses:

$$\Gamma(e) = \{+(e), \neg p(e, o_1), \ldots, \neg p(e, o_m), \neg\Gamma(f_1), \ldots, \neg\Gamma(f_n)\}$$

An existential conjunctive expressions consisting of m exist different quantified variables, i.e variables, which must be instantiated with different objects, described by n attributes over these variables:

$$C = \exists^* x_1, \ldots, x_m : f_1 \& \ldots \& f_n$$

is transformed by Γ into a linked, function-free Hornclauses:

$$\begin{aligned}\Gamma(C) = \{&+(E), \neg X_1 \neq X_2, \ldots, \neg X_{m-1} \neq X_m, \neg X_1 \neq E, \ldots, \neg X_m \neq E, \\ &\neg p(E, X_1), \ldots, \neg p(E, X_m), \neg\Gamma(f_1), \ldots, \neg\Gamma(f_n)\}\end{aligned}$$

The relation p is introduced in Γ to achieve linked Hornclauses.

Theorem 1: The transformation Γ has the following property: e is a member of the concept represented by C, if and only if
$\Gamma(e) \leq_B \Gamma(C)$, with
$B = \{\{t_1 \neq t_2\} \mid t_1, t_2 \text{ are different terms in } \Gamma(e)\} \cup \{\{X \neq Y, \neg Y \neq X\}\}$.

Proof: If e is a member of the concept represented by C, then there exists a 1 to 1 mapping φ between the m different variables x_k of C and a subset of the more than m different objects o_k of e, say $\varphi(x_k) = o_k, 1 \leq k \leq m$. This mapping establishes a substitution $\sigma_1 = \{X_k/o_k, 1 \leq k \leq m\}$ between the variables in $\Gamma(C)$ and the constants in $\Gamma(e)$. Together with the substitution $\sigma_2 = \{E/e\}$ the substitution $\sigma = \sigma_1\sigma_2$ allows to θ-match all literals
$\Gamma(C) = \{+(E), \neg p(E, X_1), \ldots, \neg p(E, X_m), \neg\Gamma(f_1), \ldots, \neg\Gamma(f_n)\}$
of $\Gamma(C)$ to the clause

$B = \{\{t_1 \neq t_2\} \mid t_1, t_2$ are different terms in LE$\} \quad \cup\{X \neq Y, \neg Y \neq X\}$,
$LH = \{\Gamma(C) \mid$ C is an existential conjunctive concept over two variables, described along n Boolean attributes$\}$,
$LE = \{\Gamma(e)\} = \{C \mid C$ is a ground member of LH$\}$
$C_1 \leq C_2 :\Leftrightarrow C_1 \leq_B C_2$

Table 4: Subproblem of Haussler (1989) in our Logical Framework.

$\Gamma(e) = \{+(e), \neg p(e, o_1), \ldots, \neg p(e, o_m), \neg\Gamma(f_1), \ldots, \neg\Gamma(f_n)\}$.
The remaining literals grounded by σ $\{\neg X_1 \neq X_2, \ldots, \neg X_{m-1} \neq X_m, \neg X_1 \neq E, \ldots, \neg X_m \neq E\}\sigma$ of $\Gamma(C)$ can be θ-matched to clauses of B, since σ substitutes different terms for different variables, and for different terms there is a clause $\{t_1 \neq t_2\}$ in B. This proves the "if" direction.
The "only if" direction follows immediately. Because of the inequality literals in $\Gamma(C)$, whenever the literals $\{\neg X_1 \neq X_2, \ldots, \neg X_{m-1} \neq X_m, \neg X_1 \neq E, \ldots, \neg X_m \neq E\}$ of $\Gamma(C)$ are θ-matched to clauses of B, the substitution σ must be a 1 to 1 mapping between variables and terms. q.e.d.
Corollary 2: The problem of deciding if there exists a Hornclause h $\in$ LH consistent with E, $|\mathrm{E}| = m$ where the examples consist of two objects described by n Boolean attributes (where m and n are variables) is NP-hard.
Proof: This follows immediately from Haussler's Theorem, Lemma 1 and the fact that Γ is a polynomial transformation. q.e.d.

2.4 A comparison of the two approaches

Now, after we have described both approaches in our framework, let us compare both approaches. Within our framework it is easy to see, that the hypothesis language of Haussler's approach $LH_{Haussler}$ is a real subset of Helft's hypothesis language LH_{Helft}. Therefore, the unique $h \in H_{Helft}$ is possibly not in $LH_{Haussler}$. But, all MSGs of Haussler are in the hypothesis space of Helft. As we have seen in the example, the $h \in H_{Haussler}$ are possibly not unique. But as $LH_{Haussler} \subset LH_{Helft}$ and $C_1 \leq_{Haussler} C_2 \Leftrightarrow C_1 \leq_{Helft} C_2$, the following lemma holds:
Lemma 3: If $E_{Helft} = E_{Haussler}$ and both are expressions of $LE_{Haussler}$, then $\forall h_1 \in H_{Helft}, h_2 \in H_{Haussler} : h_1 \leq h_2$.

We can improve the result in lemma 1. The unique h_{Helft} is in fact more specific than the conjunction of all clauses in $H_{Haussler}$. And, if the $E_{Haussler}$ does not contain any relation (except our linking relation p and the $\neq$ relation), then the unique h_{Helft} is equivalent to the conjunction of all clauses in $H_{Haussler}$.
Theorem 4: If $E_{Helft} = E_{Haussler}$ and both are expressions of $LE_{Haussler}$, then there exist a substitution σ, such that:
$\forall h \in H_{Helft} : h \leq \mu\sigma$, with $\mu = \bigcup_{h' \in H_{Haussler}} h'$

If furthermore $LE_{Haussler}$ does not contain any relation except p and $\neq$, then there exist a substitution σ, such that:
$\forall h \in H_{Helft} : h \equiv \mu\sigma$, with $\mu = \bigcup_{h' \in H_{Haussler}} h'$
Proof: As all $h \in H_{Helft}$ are equivalent, if one $h \in H_{Helft}$ has the desired properties, all have them.
Show: $h \leq \mu\sigma$:
Let σ be the $\sigma = mgu(\{p \mid p \text{ is a head of a clause of } H_{Haussler}\})$. σ exists, as all heads have the form $+(E)$. Then $\mu\sigma$ is a Hornclause as all heads are unified. As all $h' \in H_{Haussler}$ are linked, $\mu\sigma$ is also linked, because σ is also applied to the linking relation p. Clearly $\mu\sigma$ is function-free. So $\mu\sigma$ is a member of LH_{Helft}. As each $h' \in H_{Haussler}$ is a generalization of all positive examples and σ does nothing but rename the head variable, $\mu\sigma$ is a generalization of all positive examples. The $h \in H_{Helft}$ are the most specific generalizations of all positive examples in LH_{Helft}.
Show: $\mu\sigma \leq h$, if no additional relations involved:
Let h the unique (up to renaming of variables) reduced member of H_{Helft} and σ be the mgu as above. Let h' and and $\mu\sigma'$ the saturated linked Hornclauses corresponding to h and $\mu\sigma$. As stated in section 2.1, $\mu\sigma \leq h$ if and only if $\mu\sigma' \leq_\theta h'$ if and only if $\mu\sigma \supseteq h'\theta$. θ cannot do anything else than renaming of variables. If it would do anything else, it would produce a specialization of h', which would be a more special generalization of all examples than h', which contradicts that h' is a MSG. We already know that $h'\theta \supseteq \mu\sigma'$. If $\mu\sigma \supseteq h'\theta$ does not hold, then
$R := h'\theta - \mu\sigma' \neq \{\}$
If we show that R $= \{\}$, we have shown that $\mu\sigma \supseteq h'\theta$. Let us add $+(E)$ to R
$R' = R \cup \{+(E)\}$
and for each variable X_i in R which does not occur in any $\neq$ literal in R let us add $\{\neg p(E, X), \neg X \neq X_i\}$, where E is the head variable in R and X is a variable not occurring in R
$R'' = R' \cup \{\neg p(E, X), \neg X \neq X_i \mid X_i \text{ in R which does not occur in any} \neq \text{literal}\}$.
It is clear that R'' covers an example of $LE_{Haussler}$ if and only if R' covers the example. As R' is a subset of $h'\theta$, R' and R'' are both generalizations of all positive examples. Let
$r_i = \{+(E)\neg p(E, X_j), \neg p(E, X_k), \neg X_j \neq X_k\} \cup$
$\{\neg\Gamma(f) \mid \Gamma(f) \text{ contains only the variables } X_j \text{ and/or } X_k\} \subseteq R''$.
It is clear that R'' can be completely divided into a set of clauses r_i, with $1 \leq i \leq$ number of $\neq$ literals in R'', if there are no relations other that p and $\neq$ in R''. Each r_i is an element of $LH_{Haussler}$ and a generalization of all examples, i.e. $R'' \subseteq \mu\sigma'$ and therefore empty. q.e.d.

The complete division of R'' into the R_i, in the prove of $\mu\sigma \leq h$, requires that there is no relation between hypotheses in $H_{Haussler}$. But this is not true in the general case. where there are other relation, than p and $\neq$, as the following example shows.

$$
\begin{aligned}
E^+ = \{&\{+(e_1),\ \neg part(e_1,a1), \neg part(e_1,b1), \neg part(e_1,c1),\\
&\qquad \neg rel(a1,b1), \neg rel(a1,c1)\},\\
&\{+(e_2),\ \neg part(e_2,a2), \neg part(e_2,b2), \neg part(e_2,c2),\\
&\qquad \neg rel(a2,b2), \neg rel(b2,c2)\}\}\\
\mathrm{msg}_{Haussler}(E^+) &= \{+(E), \neg part(E,X_1), \neg part(E,X_2), \neg part(E,X_3),\\
&\quad \neg X_1 \neq X_2, \neg X_2 \neq X_3, \neg X_1 \neq X_3,\\
&\quad \neg rel(X_1,X_2)\}\\
\mathrm{msg}_{Helft}(E^+) &= \{+(E), \neg part(E,X_1), \neg part(E,X_2), \neg part(E,X_3), \neg part(E,X_4),\\
&\quad \neg X_1 \neq X_2, \neg X_1 \neq X_4, \neg X_2 \neq X_4, \neg X_3 \neq X_4,\\
&\quad \neg rel(X_1,X_2), \neg rel(X_3,X_4)\}
\end{aligned}
$$

We could use Helft's MSG also to compute all Haussler MSGs using the function
$select(C) := \{h \in LH_{Haussler} \mid C \supseteq h \wedge \forall h' \in LH_{Haussler} : (C \supseteq h' \wedge h' \supseteq h \Rightarrow h \supseteq h')\}$.

Lemma 5: If $E_{Helft} = E_{Haussler}$ and both are expressions of $LE_{Haussler}$ and $h \in H_{Helft}$, then $select(h) = H_{Haussler}$, and computing select(C) is polynomial in the length of the clause C and 2^d, where each example contains at most d objects.

Proof: $select(h) = H_{Haussler}$:
By definition select(h) produces all most specific subsets of h, which are elements of $LH_{Haussler}$. As $h \leq \mu\sigma$, all elements of select(h) are members of $H_{Haussler}$ and each $h' \in H_{Haussler}$ is a subset of h.
select is polynomial computable in $\mid C \mid$ and 2^d:
Each member of $LH_{Haussler}$ consist of mostly d p relations connecting the head to mostly d different variables with a different literal between every pair of them There are mostly $\mid C \mid *2^d$ different possibilities[6]. Searching all descriptors in C containing only these variables is linear to $\mid C \mid *d$.

Corollary 6: Computing a $h \in H_{Helft}$ from m examples is NP-hard, even if the examples contain only one relation (p) linking only two different objects two the head of the clause and each object is described by at most n 1-place predicates (where m and n are variable).

Proof: As select computes the solution of the NP-complete problem of finding $H_{Haussler}$ in polynomial time given a $h \in H_{Helft}$ (lemma 4 and 5), computing this h is NP-hard. q.e.d.

In fact this proves, that in the worst case even the reduced MSG $h \in H_{Helft}$ is exponentially longer than the examples. Let us give an set of n positive examples, where the MSG is exponential longer, than the examples:

$$
\begin{aligned}
E^+ = \{\ &\{+(e) \leftarrow p(x,x_1), a_1(x_1), \ldots, a_{i-1}(x_1), a_{i+1}(x_1), \ldots, a_{2n}(x_1)\\
&\qquad p(x,x_2), a_1(x_2), \ldots, a_{n+i-1}(x_2), a_{n+i+1}(x_2), \ldots, a_{2n}(x_2)\}\\
&\mid i : 1 \leq i \leq n\}
\end{aligned}
$$

[6] This corresponds to the NP-complete CLIQUE problem [4], which is polynomial for any fixed number d of objects/nodes

$B_G = \{\{b\} \mid \text{b is a ground atom}\}$
$LH_G = \{C \mid C \text{ is a Hornclause, where every variable also occurs in the head}\}$
$LE_G = \{\{+(t_1, \ldots, t_n)\} \mid +(t_1, \ldots, t_n) \text{ is ground}\}$
$C_1 \leq C_2 :\Leftrightarrow C_1 \leq_B C_2$

Table 5: Constrained Atoms in our Logical Framework.

The MSG of these n examples looks as following:

$$\begin{aligned} msg(E^+) = \{\ & +(e) \leftarrow p(x, X_i), D_i(X_i)C_i(X_i) \mid i : 1 \leq i \leq 2^n : \\ & D_i(X_i) \text{ is the } i \text{ element of the powerset of } \{a_1(x_i), \ldots, a_n(x_i)\} \\ & C_i(X_i) = \{a_{n+i}(X_i) \mid a_i(X_i) \notin D_i(X_i)\} \end{aligned}$$

The length of E^+ is $n*(1+2*(2n-1)) = O(n^2)$, whereas the length of $msg(E^+)$ is $1 + 2^n * (n+1) = O(2^n)$. And the msg is no further reducible under $\equiv_\theta$, as no $p(x, X_i), D_i(X_i)C_i(X_i)$ θ-subsumes a $p(x, X_j), D_j(X_j)C_j(X_j)$, for all $i \neq j$ and $1 \leq i, j \leq 2^n$.

2.5 Two polynomial solutions

This exponential explosion of the MSG of n examples is due two the 2^n different possibilities to map the 2 objects of the n examples. Each such mapping gives a new existential variable, and if sufficient descriptors are available, none of these variables can be reduced under $\equiv_\theta$. On way two avoid this explosion, is to forbid existential variables in the generalization. This has been done by Page and Frisch [17, 16] with their so called constrained atoms. They prove that constrained atoms (c.f. Table 5) are PAC-learnable in polynomial time. The MSG has a size of at most $Ct^r + t + 1$ Literals, where C is the number of predicates in B, r is the greatest arity of any predicate in B and t is the number of subterms in the $msg_\theta(E^+)$, where $msg_\theta(E^+)$ is the MSG of the example atoms without background knowledge. Another nice property is that the MSGs of constrained atoms shrinks with the number of examples, i.e.

$$\mid msg(\{e_1, \ldots, e_n\}) \mid \leq \mid msg(\{e_1, \ldots, e_n, e_{n+1}\}) \mid,$$

as the the number of terms (variables) cannot increase and if a relation between some terms of the head is not present in the MSG of the first n examples, it will not be present in any further MSG.

The positive result of constrained atoms is subsumed by the positive result of GOLEM [14]. In principle, GOLEM learns ij-determinate Hornclauses from examples, which are ground facts and background knowledge, which is a finite set of ground facts. ij-determinate Hornclauses can be defined in this framework as follows: Let $\{A, \neg B_1, \ldots, \neg B_m, \neg B_{m+1}, \ldots, \neg B_n\}$ be an ordered Hornclause and t be a term found in $\neg B_{m+1}$ if and only if for every substitution θ such that $A\theta \in E_G^+$ and $\{\{B_1\}, \ldots, \{B_m\}\}\theta \subseteq B_G$ there is a unique clause $\{B_{m+1}\}\theta\delta \in$

$B_G = \{\{b\} \mid$ b is a ground atom$\}$
$LH_G = \{C \mid C$ is an ij-determinate linked Hornclause$\}$
$LE_G = \{\{+(t_1, \ldots, t_n)\} \mid +(t_1, \ldots, t_n)$ is ground$\}$
$C_1 \leq C_2 :\Leftrightarrow C_1 \leq_B C_2$

Table 6: GOLEM in our Logical Framework.

B_G, i.e. for any such θ there is a unique valid ground substitution δ whose domain is the variables in t. The degree of B_{m+1} with respect to t is the number of variables in B_{m+1} which must be instantiated to determine t. Every unit clause is 0j-determinate. An ordered clause $\{A, \neg B_1, \ldots, \neg B_m, \neg B_{m+1}, \ldots, B_n\}$ is ij-determinate if and only if $\{A, \neg B_1, \ldots, \neg B_m\}$ is (i-1)j-determinate, and every literal $\neg B_k$ in $\{\neg B_{m+1}, \ldots, B_n\}$ contains only determinate terms and has degree at most j.

It is easily seen, that an constrained atom with maximum arity r of B is a 0r-determinate Hornclause.

Note, that all literals in a ij-determinate Hornclause must be either linked literals or literals unifying with only one literal of a clause of B_g.

This means GOLEM is not able to learn the MSGs defined in Haussler's framework. The only ij-determinate Hornclause (for any i and j) which is the generalization of any set $E^+ \subseteq LH_{Haussler}$ is $\{+(E)\}$. This is so, because the objects $o_1, \ldots, o_n$ in the examples are not determinate in the literals $p(e, o_i)$ $1 \leq i \leq n$ and these are the only literals linking e to $o_1, \ldots, o_n$. In the example above GOLEM will also produce $\{+(E)\}$ as the MSG. This means, GOLEMs restriction of LH is in fact also an restriction of LE, as GOLEM uses only the determinate parts of the examples.

Muggleton & Feng prove that all h $\in$ H are unique up to $\equiv$, and in the worst case h has $O((Crt)^{i^j})$ literals in the body, where C is the number of predicates as above, r is the greatest arity of any predicate in B, and t is the number of subterms in the $msg_\theta(E^+)$ as above and i and j are fixed for LH. This means, GOLEM complexity is polynomial in the examples and the background knowledge for any fixed i and j. The fact that i and j must be fixed is shown by the next theorem, which is an adoption of a theorem of Cohen, Borgida and Hirsh [3] for the same-as constructor of term-subsumption languages.

Theorem 7: If we use determinate linked Hornclauses as LH and ground determinate linked Hornclauses as LE, and no background knowledge, the MSG of n examples is only bounded by $\Omega(\frac{2^n}{p(n)})$ for some polynomial p.

Proof: Consider $E_1, E_2, \ldots, E_n$ with E_i defined as
$E_i = \{a(a_{1i}), \neg r_i(a_{1i}, a_{2i}), \neg r_i(a_{2i}, a_{1i}),$
$\neg r_j(a_{1i}, a_{1i}), \neg r_j(a_{2i}, a_{2i})\}$ for all $j \neq i$,
where the r_j are different relations. Clearly, all r_j are determinate, but the reduced MSG of these n examples contains 2^n variables and at least $4 * 2^n$ relations.

3 Conclusion

We have shown that learning of most specific generalizations in Hornlogic - even without function symbols and background knowledge - leads to exponential longer MSGs, if m of the examples contain at least two not determinate terms t_1 and t_2, and these terms are described by n predicates, where m and n are variable. We have obtained this result by comparing learning Hornclauses to learning existential conjunctive expressions, which together cover almost all approaches of most-specific (bottom-up) learning in structural domains. We have related the intractability result to the polynomial result obtained for constrained atoms, and a superset of them, the ij-determinate MSGs as produced by GOLEM. We also have shown, that dropping the ij restriction, i.e. using determinate examples could produce exponentially longer determinate MSGs.

From the example at the end of section 2.4 it is seen, that we need an unbounded number of descriptors for each existential variable to produce the exponential grow of the MSG. In further research we will investigate, if a bound on the number of descriptors will give a polynomial growing MSG.

Acknowledgments: I would like to thank Katharina Morik for the support and suggestions she has provided to this work. I also like to thank Stefan Wrobel and the members of the MLT and ILP project for interesting discussions about topics presented in this paper. Parts of this work are partially supported by the European Community ESPRIT program under contract number 6020 "Inductive Logic Programming".

References

[1] Gilles Bisson. Learning of rule systems in a first order representation. Rapport de Recherche 628, LRI, University de Paris-Sud, 1991.

[2] Wray Buntine. Generalized subsumption and its applications to induction and redundancy. *Artificial Intelligence*, 36:149 – 176, 1988.

[3] W. Cohen, A. Borgida, and H. Hirsh. Computing least common subsumers in description logic. In *Proc. of the 10th National Conference on Artificial Intelligence*, San Jose, California, 1992. MIT-Press.

[4] Michael R. Garey and David S. Johnson. *Computers and Intractability - A Guide to the Theory of NP-Completeness.* Freeman, San Francisco, Cal., 1979.

[5] David Haussler. Learning conjunctive concepts in structural domains. *Machine Learning*, 4(1):7–40, 1989.

[6] Frederick Hayes-Roth and John McDermott. Knowledge acquisition from structural descriptions. In *Proc. of the 5th International Joint Conference on Artificial Intelligence*, pages 356–362, 1977.

[7] Frederick Hayes-Roth and John McDermott. An interference matching technique for inducing abstractions. *CACM*, 21:401–410, 1978.

[8] Nicolas Helft. Inductive generalisation: A logical framework. In *Proceedings of the 2nd European Working Session on Learning*, 1987.

[9] Jörg-Uwe Kietz. Some lower bounds for the computational complexity of inductive logic programming. In *Proc. of European Conference on Machine Learning*, Vienna, Austria, to apper 1993. Springer.

[10] Jörg-Uwe Kietz and Katharina Morik. Constructive induction: Learning concepts for learning. Arbeitspapiere der GMD, GMD (German Natl. Research Center for Computer Science), P.O.Box 1240, W-5205 St. Augustin 1, Germany, 1991. Presented at the Workshop on Concept Formation in Man and Machine, May 1991. To appear.

[11] Jörg-Uwe Kietz and Stefan Wrobel. Controlling the complexity of learning in logic through syntactic and task-oriented models. In Stephen Muggleton, editor, *Proc. of Int. Workshop on Inductive Logic Programming*, pages 107 – 126, Viana de Castelo, Portugal, 1991. Also available as Arbeitspapiere der GMD No. 503.

[12] Yves Kodratoff and Jean-Gabriel Ganascia. Improving the generalization step in learning. In Ryzard S. Michalski, Jaime G. Carbonell, and Tom M. Mitchell, editors, *Machine Learning - An Artificial Intelligence Approach*, volume II, chapter 9, pages 215–244. Morgan Kaufman, 1986.

[13] Ryszard S. Michalski. A theory and methodology of inductive learning. In *Machine Learning — An Artificial Intelligence Approach*, volume I, pages 83 – 134. Morgan Kaufman, Los Altos, CA, 1983.

[14] Stephan Muggleton and Cao Feng. Efficient induction of logic programs. In *Proceedings of the 1th conference on Algorithmic Learning Theorie*, 1990.

[15] Tim Niblett. A study of generalisation in logic programs. In *Proceedings of the 3th European Working Session on Learning*, 1988.

[16] C. D. Page and A. M. Frisch. Generalisation and learnability: a study in constained atoms. In S. Muggleton, editor, *Inductive Logic Programming*. Academic Press, 1992.

[17] C. David Page, Jr. and Alan M. Frisch. Generalizing atoms in constraint logic. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, 1991.

[18] Gordon D. Plotkin. A note on inductive generalization. In B. Meltzer and D. Michie, editors, *Machine Intelligence*, volume 5, chapter 8, pages 153–163. American Elsevier, 1970.

[19] Gordon D. Plotkin. *Automatic methods of inductive inference*. PhD thesis, Univ. of Edinburgh, 1971.

[20] Gordon D. Plotkin. A further note on inductive generalization. In B. Meltzer and D. Michie, editors, *Machine Intelligence*, volume 6, chapter 8, pages 101–124. American Elsevier, 1971.

[21] Celine Rouveirol. Semantic model for induction of first order theories. In *Proc. 12th International Joint Conference on Artificial Intelligence*, 1991.

[22] Steven A. Vere. Induction of concepts in the predicate calculus. In *Proc. of the 4th International Joint Conference on Artificial Intelligence*, pages 281–287, 1975.

[23] Steven A. Vere. Induction of relational productions in the presence of background information. In *Proc. of the 5th International Joint Conference on Artificial Intelligence*, pages 349–355, 1977.

[24] Steven A. Vere. Multilevel counterfactuals for generalizations of relational concepts and productions. *Artificial Intelligence*, 14:139–164, 1980.

[25] Christel Vrain. Ogust: A system that learns using domain properties expressed as theorems. In Yves Kodrtoff and Ryszard Michalski, editors, *Machine Learning - An Artificial Intelligence Approach*, volume III, chapter 13, pages 360–382. Morgan Kaufmann, 1990.

[26] P. H. Winston. Learning structural descriptions from examples. In P.H. Winston, editor, *The psychology of computer vision*. McGraw-Hill, 1975.