Reinforcement Learning in Constructive Languages

Josep Hernandez-Orallo

Universitat Politècnica de València, Departament de Sistemes Informàtics i Computació,
Camí de Vera 14, Aptat. 22.012 E-46071, València (L'Horta)
E-mail: jorallo@dsic.upv.es

Abstract

We present different ways of measuring reinforcement for eager learning methods and constructive languages. The problem of propagating reinforcement from the evidence into the theory has been shown especially troublesome in high-level languages, like ILP, but the same problem pervades other representations that allow redescription (e.g. neural networks).

In this work, we present an operative measure of reinforcement for general theories, studying the growth of knowledge, theory revision and abduction in this framework. Our approach performs an apportionment of credit wrt. the 'course' that the evidence makes through the learnt theory. The result is compared with other evaluation criteria, like the MDL principle.

Finally, we study a more common view of reinforcement, where the actions of an intelligent system can be rewarded or penalised, and we discuss whether this should affect the distribution of reinforcement.

The most important result of this paper is that the way we distribute reinforcement into knowledge results in a *rated* ontology. In this way, one of the most difficult dilemmas of inductive learning, the choice of a prior distribution, disappears.

Keywords: Reinforcement Learning, Incremental Learning, Ontology, Apportionment of Credit, Abduction, Induction, MDL principle, Knowledge Acquisition and Revision, ILP.

1 Introduction

The study of reinforcement learning in restricted representations has been especially fruitful in this decade (see [19] for a survey) and it has been recently related with EBL [7]. One of the main problems of reinforcement learning is that it is increasingly more difficult to assign and 'propagate' the reinforcement (or apportionment of credit [18]) depending on two factors (which are as well related): (1) how eager is the inductive strategy (vs. lazy methods like instance-based and case-based reasoning [27]) and (2) how expressible is the language where induction must take place. Explanatory Based Learning (EBL) and Inductive Logic Programming (ILP) are two areas where the propagation of reinforcement faces these issues in a more arduous way.

In this paper we shall address the problem of reinforcement with eager learning methods. Eager learning methods extract all the regularity from the data in order to work with intensional knowledge (instead of the extensional knowledge of lazy methods [1]).

Additionally, we will consider the problem with constructive languages. A constructive language is a language that allows dynamical change of its representational bias (what is sometimes known as the possibility of 'redescription'), i.e., new constructed terms can be created to express more compactly the evidence. This is usually known in ILP as predicate invention.

In decision trees or attribute languages, no invented terms are induced and the reinforcement is distributed among the initial attributes. The main drawback of these approaches is the lack of flexibility: when arrived to a 'saturation' point, the data are not abstracted further and the mean reinforcement arrives to a limit. Consequently, the ontology must be given and not constructed (a model of the 'world' is embedded in the system) and the possible extensions of this world are very restricted.

In the case of learning in highly expressible frameworks, a main problem is presented (apart from efficiency): the ontology of the new constructed concepts is indirect. The usual solution to this problem is the assumption of a prior probability. Once the probabilities are assigned, a bayesian framework can be used to 'propagate' the distribution.

In general, there is not justification at all of which prior distribution to choose. In the absence of any knowledge, the most usual one is the MDL (Minimum Description Length) principle [34][35]. The MDL principle is just a formalisation of Occam's razor. Theoretically, its close relation with PAC-learning [41] has been established by [4]. Some high-level representation inductive methods have adapted these ideas (e.g. U-learnability in ILP [32]).

[1] On-line papers: http://www.dsic.upv.es/~jorallo/escrits/escritsa.htm

All of them are based on the assumption of a prior. However, there are many riddles with the management of probabilities and, in particular, the best choice, the MDL principle, has additional ones.

As we will see, most of these difficulties would disappear if no prior distribution is assumed and the knowledge is constructed by reinforcement, as the data suggest. However, the translation of these ideas to general representational frameworks seems difficult. First, the length of the structures which supposedly are to be reinforced is variable. Second, and more importantly, it seems we can always invent 'fantastic' concepts that can be used in the rest of knowledge. Consequently, these 'fantastic' concepts are highly reinforced, increasing the reinforcement ratio of knowledge in an unfair way.

An immediate way out is the combination of reinforcement learning with some prior, mainly the MDL principle, essayed under the name of 'incremental self-improvement' [36] using syntactic minimality to restrict the appearance of these inventions.

Notwithstanding, our approach also avoids 'fantastic' concepts but it is based exclusively on reinforcement. Consequently, compression turns out to be an 'a posteriori' consequence of a well-established reinforcement, instead of an 'arbitrary' assumption.

The paper is organised as follows. Section 2 presents some prior distributions usually assumed in machine learning, especially the MDL principle. Section 3 introduces our framework for incremental knowledge construction. Section 4 essays a first adaptation of reinforcement to realise the problems of 'fantastic' concepts. Section 5 remakes the approach and introduces the idea of 'course' to measure reinforcement. Section 6 discusses the extension of these ideas to wider notions of reinforcement with the presence of reward and penalties. Section 7 considers the length of the reinforced 'units' or 'rules' showing the relation with the MDL principle in the limit. In the same section it is introduced a balanced reinforcement suitable for EBL. Section 8 presents two methods for computing effectively these measures and deals with their limitations and complexity. Section 9 closes the paper discussing the results and the open questions.

2 Prior Selection in Machine Learning

The aim of Machine Learning is the computational construction of hypothetical inferences from facts, as Michalski have pointed out [28]: "*inductive inference was defined as a process of generating descriptions that imply original facts in the context of background knowledge. Such a general definition includes inductive generalisation and abduction as special cases*".

However, given some evidence E, infinite many hypotheses H can be induced ensuring $H \models E$. Obviously, some selection criteria are needed. Depending on different applications, some criteria have been used (e.g. the most specific hypothesis, the most general one, the shortest one, the most informative one, ...). In general, this choice implies the assumption of a prior distribution which can be used to derive the likeliness of the hypotheses.

The principle of simplicity, represented by Occam's razor, selects the shortest hypothesis as the most plausible one. This principle was rejected by Karl Popper because, in his opinion (and at that moment) there *was* no objective criterion for simplicity. However, Kolmogorov complexity [43], denoted $K(x)$, is an objective criterion for simplicity. This is precisely what R.J.Solomonoff proposed as a 'perfect' theory of induction [26]. Algorithmic Complexity inspired J. Rissanen in 1978 to use it as a general modelling method, giving the popular MDL principle [34], recently revised as a one-part code [35] instead of the initial two-parts code formulation.

It is remarkable (and often forgotten) that Kolmogorov Complexity just gives consistency to this theory of induction; Occam's razor is *assumed* but not proven. Nonetheless, some justifications have been given in the context of physics, reliability and entropy, but, in our opinion, it is the notion of *reinforcement* (or cross validation) which justifies the MDL principle more naturally. At a first sight, it *seems* that the higher the mean compression ratio ($length(E)$ / $length(H)$) the higher the mean reinforcement ratio.

Summing up, the MDL principle says that, in absence of any other knowledge about the hypotheses distribution, we should select the prior $P(h) = 2^{-K(h)}$, prevailing short theories over large ones. However, this prior has many riddles. First of all, (1) it is not computable, so the prior must be approximated (e.g. using the time-weighted variant Kt of Kolmogorov complexity [24]) or must dynamically change as the learner knows that something can be further compressed. Second, (2) it presents problems with perfect data; the MDL principle usually 'underfits' the data, because sometimes it is too conservative. Third, the reliability of the theory is not always increasing with the number of examples which have confirmed the theory (e.g., a string of 10^{10} *a*'s is more compressible than a string of 78450607356 *a*'s !). Moreover, (4) is is difficult to work with different and non-exclusive hypotheses, because if we have T_a and T_b, intuition (and logic) says

that $T = T_a \vee T_b$ should have more probability, but MDL assigns less probability to T because it is larger.

Finally, (5) the MDL principle has shown problems for explanation, because, for the sake of maximum mean compression, some part of the hypothesis can be not compressed at all, resulting in a very compressed part plus some additional extensional cases. This extensional part is not validated, making the whole theory weak. An ontology is difficult to construct from here if they are unrelated (not explained) with the other facts. This is closely related with the differentiation between Enumerative Induction and Best Explanation [13] [14] [8] and the distinction between Induction and Abduction [10].

We intend to handle these difficulties with a dynamical reinforcement. However, our approach has additional advantages: (1) no prior assumption has to be made (apart from how to distribute this reinforcement, which is the topic of this paper), i.e. knowledge is constructed just as the data suggest, and (2) reinforcement can be more flexibly managed than probabilities, and allows further insight on the relation between the evidence and the theory.

3 Preliminaries

With this section we just present the schema of incremental learning and the languages we aim to address in the following sections.

3.1 Incremental Knowledge Construction

The field of knowledge construction gathers many other related subfields and usually makes use of very complex techniques for the organisation and revision of the data. We will tackle exclusively the inductive or learning task in knowledge construction.

Incremental knowledge construction (which includes acquisition and revision) generates a theory from an evidence that is gradually supplied example by example. From the very beginning, with an empty knowledge $T=\varnothing$, when new observations or evidences e are received, we can have three possible situations:

- **Prediction Hit** (*or 'matter of course'*). The observations are covered without more assumptions, i.e., $T \models e$. The theory T is reinforced.
- **Novelty**. The observation is uncovered but consistent with T, i.e. $T \not\models e$ and $T \cup e \not\models \square$. Here, the possible actions are:
  1. *Extension*: T can be extended with a good explanation A, (i.e. $T \cup A \models e$).
  2. *Revision*: revised if a good explanation cannot be found,
  3. *Patch*: quoted as an extensional exception (i.e. $T' = T \cup e$), or
  4. *Rejection*: regarded as noise and ignored.
- **Anomaly**. The observation is inconsistent with the theory T, i.e., $T \not\models e$ and $T \cup e \models \square$. In this case, T cannot be extended and there are three possibilities: *revision*, *patch* or *rejection*.

An eager but still non-explanatory approach to theory formation is Kuhn's theory of changing paradigms [23] which basically matches with the MDL principle: as too many exceptions to the paradigm are found, they are increasingly lengthy to quote (*patch*) and the whole paradigm (or part of it) must be changed.

In the preceding sketch, abduction appears as an extension of current knowledge with some assumption (usually one or more facts) and induction is also an extension or revision which performs some kind of generalisation. Nevertheless, this characterisation is not sufficient for a clear distinction (see [15] for sounder considerations about how to distinguish them). In fact, it is a topic of current discussion (for a state of the art see [10]). In this way, abduction has been commonly seen as belief revision [5], usually combined with induction [2]. In other cases is related with validation, justification or ontology [9] in the way the part of the theory where abduction supports must be reliable. Unavoidabily, this reliability must come from a reinforcement produced by the previous evidence.

The previous schema is general enough to include explanatory and conservative knowledge construction. Explanatory knowledge construction should minimise the exceptions, so patches and revisions should not be allowed. Thus, the revisions are much more frequent. Even more, the goal is anticipating, investing, finding more informative and easily refutable hypotheses [33], in contrast to what many approaches to minimal revisions aim for (see e.g. [30]), supported by the obvious fact that a minimal revision is usually less costly, *in short-term*, than a deep revision.

3.2 Representation Languages

For the study of reinforcement we need to introduce some basics for the representation to which it can be applied. A 'pattern' of languages is defined as a set of *chunks* or rules r which are composed of a head (or consequence) and a body (or set of conditions) in the following way $r \equiv \{ h :- t_1, t_2, .. t_s \}$.

Since no restriction of how h and t_i can be (there may be variables, equations, boolean operators...), our definition could be specialised to propositional languages, Horn theories, full logical theories, functional languages, some kind of grammars, and even higher-

order languages. In the following, we leave unspecified the semantics of the representations and we just say that e is a consequence of P, denoted $P \models e$ (in other words, there is a proof for e in P, or, simply, P covers e).

4 Reinforcement wrt. the Theory Use

Whatever the approach to knowledge construction, the revision of knowledge must come from a partial or total weakness of the theory or, in other words, a loss of *reinforcement* (or apportionment of credit [18]). We present a way to compute the reinforcement map for a given theory, depending on past observations.

DEFINITION 4.1

The pure reinforcement $\rho\rho(r)$ of a rule r from a theory T wrt. some given evidence $E = \{e_1, e_2, \ldots, e_s\}$ is computed as the number of proofs of e_i where r is used. If there are more than one proof for a given e_i, *all* of them are reckoned, but in the same proof, a rule is computed only once.

DEFINITION 4.2

The (normalised) reinforcement $\rho(r) = 1 - 2^{-\rho\rho(r)}$.

Definition 4.2 is motivated by the convenience of maintaining reinforcement between 0 and 1. The mean reinforced ratio $m\rho(T)$ is defined as $\Sigma_{r\in T}\, \rho(r)/m$, being m the number of rules. These definitions show that, in general, the most (*mean*) reinforced theory is not the shortest one as the following example shows:

EXAMPLE 4.1

Given the evidence e_1, e_2, e_3, consider a theory $T_a = \{r_1, r_2, r_3\}$ where $\{r_1\}$ covers $\{e_1\}$, $\{r_2\}$ covers $\{e_2\}$ and $\{r_3\}$ covers $\{e_3\}$ and a theory $T_b = \{r_1, r_2, r_3, r_4\}$ where $\{r_1, r_4\}$ cover $\{e_1\}$, $\{ r_2, r_4 \}$ cover $\{e_2\}$ and $\{r_3, r_4 \}$ cover $\{e_3\}$.

From here, T_a is less reinforced than T_b.

In the first case we have $\rho\rho_{a,1} = \rho\rho_{a,2} = \rho\rho_{a,3} = 1$ and $m\rho(T_a) = 0.5$. For T_b we have $\rho\rho_{b,1} = \rho\rho_{b,2} = \rho\rho_{b,3} = 1$, $\rho\rho_{b,4} = 3$ and $m\rho(T_b) = 0.5938$.

In addition, redundancy does not imply a loss of mean reinforcement ratio (e.g. just add twice the same rule).

However, measuring reinforcement of the *theory* presents problems of *fantastic* (unreal) concepts:

THEOREM 4.3

Consider a program P composed of rules r_i of the form $\{ h :- t_1, t_2, .. t_s \}$, which covers n examples $E = \{ e_1, e_2, ... e_n \}$. If the mean reinforcement ratio $m\rho < 1 - 2^{-m}$ then it can always be increased.

PROOF

A *fantastic* rule r_f can be added to the program by modifying all the rules of the program in the following way $r_i = \{ h :- t_1, t_2, .. t_s , r_f \}$. Obviously, all the other rules maintain the same reinforcement but r_f is now reinforced with $\rho\rho_f = n$. Since $\rho_f > m\rho$ then the new $m\rho'$ must be greater than $m\rho$. □

One can argue that these *fantastic* rules could be checked out and eliminated. However, there are many ways to 'hide' a fantastic rule; in fact, cryptography relies on this fact.

5 Reinforcement wrt. the Evidence

It can be derived from this problem that reinforcement must be combined with a simplicity criterion in order to work (maybe neural networks theory is the field where this avoidance of overfitting, ensured by simplicity, has been more thoroughfully studied in combination with reinforcement).

However, there is solution without explicitly making use of simplicity. The idea is measuring the validation *wrt. the evidence*.

DEFINITION 5.1

The course $\chi_T(f)$ of a given fact f wrt. to a theory is computed as the product of all the reinforcements $\rho(r)$ of all the rules r used in the proof of f. If a rule is used more than once, it is computed once. If f has more than one proof, we select the greatest course.

In this case, we can select the theory with the greatest *mean* of the courses of all the data presented so far, defined as $m\chi(T, E) = \Sigma_{e\in E}\, \chi_T(e)/n$, being n the number of facts (examples) in the evidence. We can use the geometric mean instead, denoted by $\mu\chi$. The following example shows the use of this *new* criterion for knowledge construction:

EXAMPLE 5.1

Using Horn theories as representation (Prolog), suppose we have an incremental learning session as follows:

⌦ Given the background theory $B = \{$ s(a,b), s(b,c), s(c,d) $\}$ we observe the evidence $E = \{ e^+_1$: r(a,b,c), e^+_2: r(b,c,d), e^+_3: r(a,c,d), e^-_1: ¬r(b,a,c), e^-_2: ¬r(c,a,c) $\}$:

The following programs could be induced, with their corresponding reinforcements and courses:

P_1 = {r(X,Y,Z) :- s(Y,Z) : ρ = 0.875}
$\chi(e^+_1) = \chi(e^+_2) = \chi(e^+_3) = 0.875$

P_2 = {r(X,c,Z) : ρ = 0.75
r(a,Y,Z) : ρ = 0.75}

$\chi(e^+_1)= \chi(e^+_2)= \chi(e^+_3)= 0.75$

P_3 = {r(X,Y,Z) :- s(X,Y) : ρ = 0.75
r(X,Y,Z) :- s(Y,Z) : ρ = 0.875}

$\chi(e^+_1)= \chi(e^+_2)= \chi(e^+_3)= 0.875$

P_4 = {r(X,Y,Z) :- t(X,Y), t(Y,Z) : ρ = 0.875
t(X,Y) :- s(X,Y) : ρ = 0.875
t(X,Y) :- s(X,Z), t(Z,Y) : ρ = 0.5}

$\chi(e^+_1)= \chi(e^+_2)= 0.7656$, $\chi(e^+_3)= 0.3828$

P_5 = {r(X,Y,Z) :- t(X,Y) : ρ = 0.875
t(X,Y) :- s(X,Y) : ρ = 0.875
t(X,Y) :- s(X,Z), t(Z,Y) : ρ = 0.5}

$\chi(e^+_1)= \chi(e^+_2)= 0.7656$, $\chi(e^+_3)= 0.3828$

At this moment, P_1 and P_3 are the best options and P_4 and P_5 seem 'risky' theories according to the evidence.

⌦ e^+_4 = r(a,b,d) is observed.

P_1 does not cover e_4^+ and it is patched:

P_{1a}' = {r(X,Y,Z) :- s(Y,Z) : ρ = 0.875
r(a,b,d) : ρ = 0.5}

$\chi(e^+_1)= \chi(e^+_2)= \chi(e^+_3) = 0.875$, $\chi(e^+_4) = 0.5$

mχ = 0.78, μχ = 0.76

P_{1b}' = {r(X,Y,Z) :- s(Y,Z) : ρ = 0.875
r(X,Y,d) : ρ = 0.875 }

$\chi(e^+_1)= \chi(e^+_2)= \chi(e^+_3)= \chi(e^+_4) = 0.875$

P_2' is reinforced = {r(X,c,Z) : ρ = 0.75.
r(a,Y,Z) : ρ = 0.875}

$\chi(e^+_1) = 0.875$, $\chi(e^+_2) = 0.75$, $\chi(e^+_3) = \chi(e^+_4) = 0.875$

P_3' is reinforced = {r(X,Y,Z) :- s(X,Y) : ρ = 0.875.
r(X,Y,Z) :- s(Y,Z) : ρ = 0.875}

$\chi(e^+_1)= \chi(e^+_2)= \chi(e^+_3)= \chi(e^+_4) = 0.875$

P_4' is reinforced =

P_4'= { r(X,Y,Z):-t(X,Y), t(Y,Z): ρ = 0.9375
t(X,Y) :- s(X,Y) : ρ = 0.9375
t(X,Y) :- s(X,Z), t(Z,Y) : ρ = 0.75}

$\chi(e^+_1)= \chi(e^+_2)= 0.8789$, $\chi(e^+_3)= \chi(e^+_4) = 0.6592$

mχ = 0.77, μχ= 0.76

P_5' is slightly reinforced

P_5' = { r(X,Y,Z) :- t(X,Y) : ρ = 0.9375.
t(X,Y) :- s(X,Y) : ρ = 0.9375
t(X,Y) :- s(X,Z), t(Z,Y) : ρ = 0.5}

$\chi(e^+_1)=\chi(e^+_2)=0.8789$, $\chi(e^+_3)=0.4395$, $\chi(e^+_4)=0.8789$

mχ = 0.77, μχ= 0.74

At this moment, P_{1b}' and P_3' are the best options. Now P_4' and P_5' seem more grounded.

⌦ We add e^-_3 = ¬r(a,d,d)

P_{1a}' remains the same and P_{1b}' and P_2' are inconsistent, motivating the following 'patches' for them:

P_{2a}' = {r(X,c,Z) : ρ = 0.75.
r(X,b,Z) : ρ = 0.75}

$\chi(e^+_1)= \chi(e^+_2)= \chi(e^+_3)= \chi(e^+_4) = 0.75$

P_{2b}' = {r(X,Y,Z) :- e(Y) : ρ = 0.9375.
e(b) : ρ = 0.75
e(c) : ρ = 0.75}

$\chi(e^+_1)= \chi(e^+_2)= \chi(e^+_3)= \chi(e^+_4) = 0.7031$

P_3' and P_4' remain the same. P_5' becomes inconsistent.

⌦ We add e^+_5 = r(a,d,e)

P_{1a}', P_{2a}', P_{2b}' can only be patched with e^+_5 as an exception because abduction is not possible.

P_3' has abduction as a better option.

P_3" = {s(d,e) : ρ = 0.5
r(X,Y,Z) :- s(X,Y) : ρ = 0.875
r(X,Y,Z) :- s(Y,Z) : ρ = 0.9375}

$\chi(e^+_1)= \chi(e^+_2)= \chi(e^+_3)= 0.9375$,

$\chi(e^+_4) = 0.875$, $\chi(e^+_5) = 0.46875$

mχ = 0.831, μχ= 0.805

P_4' makes the same abduction

P_4" = { s(d,e) : ρ = 0.5
r(X,Y,Z):-t(X,Y),t(Y,Z): ρ=0.96875
t(X,Y) :- s(X,Y) : ρ = 0.96875
t(X,Y) :- s(X,Z), t(Z,Y): ρ = 0.875}

$\chi(e^+_1)=\chi(e^+_2)=0.939$,$\chi(e^+_3)=\chi(e^+_4)=0.82$,$\chi(e^+_5)=0.41$

mχ = 0.786, μχ= 0.754

At this moment, P_3" and P_4" are the best options.

Further examples would be required to distinguish which is the 'intended' one with more reliability.

The example illustrates that in general, and using this new recknoning of reinforcement, the shortest theories are not the best ones. More importantly, it also shows that as soon as a theory gains some solidity, abduction can be applied.

The way reinforcements are calculated makes that very complex programs are avoided, but redundancy is possible. But now there is not any risk of fantastic concepts. As said before, for any program P composed of rules r_i of the form { h :- t_1, t_2, .. t_s }, which covers m examples E = { e_1, e_2, ... e_n } and their reinforcements ρ_i, a *fantastic* rule r_f could be added to the program and all the rules could be modified in the following way r_i = { h :- t_1, t_2, .. t_s , r_f }. The following theorem shows that now it is not reinforced over the original one:

THEOREM 5.1

The course of any example cannot be increased by the use of *fantastic* concepts.

PROOF

Since the *fantastic* concept r_f now appears in all the proofs of the n examples, the reinforcement of r_f is exactly $1 - 2^{-n}$ and the reinforcements of all the r_i remain the same. Hence, the course of all the m examples is modified to $\chi'(e_j) = \chi(e_j) \cdot r_f = \chi(e_j) - \chi(e_j) \cdot 2^{-n}$. Since n is finite, for all $e_j \in E$, $\chi'(e_j)$ can never be greater than $\chi(e_j)$.□

These ideas are being used by [17] in an incremental learning system using Curry as a representation language (a logic functional programming language based on narrowing with some higher-order constructs). The results demonstrate that the *intended* hypothesis is found sooner than when using the MDL principle.

Another advantage of this approach is that a 'rated' ontology can be derived directly from the theory. In this way, the parts which are sound or weak are easy to detect. Intuitively, if a rule only covers just one example, it suggests that the rule is not very real.

6 Rewarded Reinforcement

In reinforcement learning, it is usually assumed that the learner receives some reward (or penalty) value of its actions. In other words, prediction hits can receive different degrees of reward and prediction errors (including novelties and anomalies) can receive different degrees of penalty (or negative reward).

Usually, this broader view of reinforcement is suitable for frameworks where reasoning about action is necessary. The rewards are assigned depending on the actions that the agent performs for each situation. Temporal languages are used for representation, like event calculus or situation calculus [22]. The important issue here is that our model can be used in these cases, by asking the learning system to predict the following situation s_{n+1} after every possible action it can perform in a certain situation s_n. The task of the system is just selecting the one with the greatest reward. In the case the result of the action matches with the evidence, a positive hit happens with the predicted reward. In the case a prediction error occurs, the action may have no awful consequences (no penalty) or it may be fatal. The question is how ontology and 'hedonism' must be combined. It is commonly accepted in psychology the claim that hedonism motivates ontology, and this is stronger the earlier the stage of development of a cognitive system. In our opinion, this motivation does not imply that they must be mixed. Moreover, rewards should be learned as well because they may change.

From here, the choice of the best action must take into account both the reliability of the prediction (i.e. the reinforcement) weighted with the reward, not the action with the best reward alone (because it may be a very weak guess).

Finally, there can be degrees of reliability in the evidence. This degree may come from different reliabilities of the sensors of the system or from intermediate recognition or sensor preprocessing subsystems. Indeed, this should affect ontology in the following way: every fact of the evidence is assigned a real number as a reliability degree, $-1 \leq d_f \leq 1$. In this framework, the completely reliable positive examples are assigned a value of $d_f = 1$ and the completely reliable negative examples are assigned a value of $d_f = -1$.

DEFINITION 6.1

The *'grounded' course* $\chi'(f)$ of a given fact f wrt. to a theory is computed as the normal course $\chi(f)$ multiplied by the reliability degree of f. More formally, $\chi'(f) = \chi(f) \cdot d_f$.

7 Balanced Reinforcement

With the approaches introduced in section 5 and section 6 there is a tricky way of increasing reinforcement: joining rules. If a high-level representation mechanism allows very expressive rules, larger rules can be made in order to stand for the same that was expressed with separated rules, with the advantage of increasing reinforcement:

EXAMPLE 7.1

For instance, the following extended functional programs are equivalent:

T_a = { r_1 = { f(X,a) → g(b) },
r_2 = { f(X,c) → i(d) } }

T_b = { r = { f(X,Y) → if (Y=a) then g(b)
if (Y=c) else i(d) } }

but T_b would be more reinforced than T_a.

The solution to this problem requires the introduction of a factor inversely related with the syntactical length of a rules. It is important to clarify that this syntactical measure is not a prior and it can be effectively computed, in contrast to the MDL principle.

With length(r) we denote the length of a rule r for the concrete language which would be used. The only restriction for length is that for all r, length(r) ≥ 1. Thus we extend the definitions of section 5:

DEFINITION 7.1

The extended pure reinforcement is defined as:

$$\rho\rho^*(r) = \rho\rho(r) \,/\, \text{length}(r).$$

The extended normalised reinforcement $\rho^*(r)$ and the extended courses $\chi^*(r)$ are defined in the obvious way using $\rho\rho^*(r)$ and $\rho^*(r)$.

It is obvious that if length(r) simply assigns 1 to every rule of the program, these definitions are equivalent to those of section 3.

With this extension, it is easy to show that —in the limit [11]— the MDL principle is an excellent principle for achieving reinforcement:

THEOREM 7.2

If the data E are infinite and a theory T is finite, the mean course $m\chi^*(T, E) = 1$.

PROOF

Given some infinite data as evidence $E = \{ e_1, ..., e_n \}$, without loss of generality, consider that T can be exclusively composed of two rules: r_1, which covers all E except e_i and, *independently*, r_2, which covers e_i. The reinforcements are $\rho^*(r_1) = (1-2^{(1-n)/\text{length}(r_1)})$ and $\rho^*(r_2) = (1-2^{-1/\text{length}(r_2)})$ and the mean course $m\chi^*(T, E) = [(n-1) \cdot (1-2^{(1-n)/\text{length}(r_1)}) + (1-2^{-1/\text{length}(r_2)})] / n$. For infinite data, we have that $\lim_{n\to\infty} m\chi^*(T, E) = 1$. □

This theorem shows that maximum reinforcement matches with maximum compression in the limit (simply because both are saturated). However, when the data are finite we have many cases where they differ. The most blatant case occurs when some exception is covered extensionally (as r_2 which covers d_i in the proof of theorem 7.2) and there is an important loss of reinforcement vs. a slight loss of compression. The following example illustrates this point:

EXAMPLE 7.2

Consider the following evidence e_1–e_{10}:

$E = \{$ e_1: e(4) → true, e_2: e(12) → true,
e_3: e(3) → false, e_4: e(2) → true,
e_5: e(7) → false, e_6: e(7) → false,
e_7: e(20) → true, e_8: e(0) → true,
e_9: o(3) → true, e_{10}: o(2) → false $\}$

and that natural numbers are represented using the functor s as the symbol for succesor, e.g. s(s(s(0))) means 3. The length (denoted l) of a rule is computed as $1+n_f+n_v$, where n_f means the number of functors (including constants as functors with arity 0) and n_v the number of variables.

From here, the following theories are evaluated:

	: l	ρρ	ρρ*	ρ*
T_a= { e(s(s(X)) → e(X)	: 7	7	1	0.5
e(0) → true	: 4	5	1.2	0.5647
e(s(0)) → false	: 5	3	0.6	0.3402
o(s(s(s(0)))) → true	: 7	1	0.1429	0.0943
o(s(s(0))) → false	: 6	1	0.1667	0.1091}

The extended courses are $m\chi^*(e_1, e_2, e_4, e_7, e_8) = 0.5 \cdot 0.5647 = 0.28235$, $m\chi^*(e_3, e_5, e_6) = 0.5 \cdot 0.3402 = 0.1701$, $m\chi^*(e_9) = 0.0943$ and $m\chi^*(e_{10}) = 0.1091$.

The mean extended course $m\chi^*$' is 0.2125.

	: l	ρρ	ρρ*	ρ*
T_b= { e(s(s(X)) → e(X)	: 7	7	1	0.5
e(0) → true	: 4	5	1.2	0.5647
e(s(0)) → false	: 5	3	0.6	0.3402
o(s(s(X)) → o(X)	: 7	2	0.2857	0.1797
o(0) → false	: 4	1	0.25	0.1591
o(s(0)) → true	: 5	1	0.2	0.1294}

The extended courses are $m\chi^*(e_1, e_2, e_4, e_7, e_8) = 0.5 \cdot 0.5647 = 0.28235$, $m\chi^*(e_3, e_5, e_6) = 0.5 \cdot 0.3402 = 0.1701$, $m\chi^*(e_9) = 0.1797 \cdot 0.1294 = 0.02325$ and $m\chi^*(e_{10}) = 0.1797 \cdot 0.1591 = 0.02859$.

The mean extended course $m\chi^*$' is 0.1974.

	: l	ρρ	ρρ*	ρ*
T_c= { e(s(s(X)) → e(X)	: 7	9	1.2857	0.5898
e(0) → true	: 4	6	1.5	0.6464
e(s(0)) → false	: 5	4	0.8	0.4257
o(X) → not(e(X))	: 6	2	0.3333	0.2063
not(true) → false	: 4	1	0.25	0.1591
not(false) → true	: 4	1	0.25	0.1591}

The extended courses are $m\chi^*(e_1, e_2, e_4, e_7, e_8) = 0.5898 \cdot 0.6464 = 0.3813$, $m\chi^*(e_3, e_5, e_6) = 0.5898 \cdot 0.4257 = 0.2511$, $m\chi^*(e_9) = 0.2063 \cdot 0.5898 \cdot 0.4257 \cdot 0.1591 = 0.00824$ and $m\chi^*(e_{10}) = 0.2063 \cdot 0.5898 \cdot 0.6464 \cdot 0.1591 = 0.0125$.

The mean extended course $m\chi^*$' is 0.2681.

Note that the lengths ($l(T_a)$=29, $l(T_b)$=32, $l(T_c)$ = 30) would not give many hints about which theory to select.

The example also shows the advantages of this approach for explanation-based learning. Since all the data must be explained, if a part is left in an extensional way (or unrelated with the rest), it is penalised. On the other hand, we have seen in the preceding sections that *fantastic* concepts are also avoided, so it results to be a *balanced* criterion for the 'intensionality' of theories, without falling into fantasy.

Regarding T_c of example 7.2, our measure can be adapted to situations where a more compensated theory is required, using a *geometric mean* instead of an *arithmetic mean*. In addition, and concerning T_a, if we do not want exceptions (extensional parts) at all, we can discard theories where a fact has a course value less than the mean divided by a constant. Moreover, this case should trigger theory revision in an incremental framework in order to integrate (or reconcile) the example with the theory.

Finally, another straightforward extension to our approach is considering the length of the examples, too. However, it could also be included in the reliability value which was discussed in section 6.

8 Computing Reinforcement

First of all, it should be stated clear that our theory of reinforcement is not an inductive learning method. We have not dealt about how the theory could be constructed from the evidence. On the contrary, this paper presents a set of measures that allow a detailed study of the relation between the theory and the evidence, assisting the evaluation, the selection, and the revision of theories.

A general method of computing reinforcement is as it has been used in the examples:

GENERAL METHOD:

Consider the theory T, with m rules $r_1..r_m$, and the evidence E, with n examples $e_1..e_n$, such that $T \models E$. First we must *prove* all the examples and compute $\rho\rho^*$ and ρ^* for each rule. In a second stage, we *prove* again the n examples, computing χ^* from the ρ^* obtained in the first stage.

The complexity of the previous method *seems* to be, in the worst case, in $O(m \cdot n)$. However it is not, because we have not stated any restriction about the computational cost of the theory, and each proof has its cost.

However, it would be more realistic to consider the computing of reinforcement in an incremental setting:

INCREMENTAL METHOD:

We will use four arrays: $l_{1..m}$, $\rho\rho^*_{1..m}$, $\rho^*_{1..m}$, $\chi^*_{1..n}$ for the lengths, the pure and normalised reinforcements and the courses, respectively. An additional boolean bidimensional array $U_{1..m,1..n}$ assigns *true* to $U_{j,i}$ iff e_i uses r_m in its proof and *false* otherwise.

For each new example e_{n+1} which is received we have different possibilities:

1. If it is a *hit*, we remake $\rho\rho^*_{1..m}$, $\rho^*_{1..m}$, according to the proof of e_{n+1}, U is extended with $U_{\cdot,n+1}$ and $\chi^*_{1..n+1}$ is updated using U.
2. If it is a *novelty* and no revision is made to T, only an extension $T' = T \cup \{r_{m+1}, ..., r_{m+k}\}$, the steps are very similiar to the previous case, except that the arrays must be extended to $m+k$.
3. Finally, if it is a *novelty* or an *anomaly* and the theory is revised in some rules $\{r_1, ..., r_p\}$ and extended in others $\{r_{m+1}, ..., r_{m+k}\}$, only the $U_{\cdot,j}$ which does not use any rule from $\{r_1, ..., r_p\}$ can be preserved. The rest must be remade.

The previous method ignores two exceptional cases: that a *hit* could trigger a revision of the theory to readjust reinforcements and that case 2. could produce alternative proofs for previous examples.

Further optimisation could come from a deeper study of the static dependencies (i.e. some rule always depends on others) and the topology of dependencies that the theory generates. On the other hand, an appropriate approximation could be used. Even more, some of the past evidence can be 'forgotten' if it is covered by very reinforced rules, so minimising the cost.

However, in the case that an inductive learning method uses reinforcement for evaluating the theories it is constructing, the complexity of these methods would surely be very modest compared to the usual huge costs of machine learning algorithms.

Moreover, reinforcement measures are a very adequate tool to guide a learning algorithm. For instance, in [17], the examples and rules with low reinforcement were mixed in order to 'conciliate' them and to obtain more compact and reinforced theories.

9 Conclusions

We have presented a framework to distribute or propagate reinforcement into a theory depending on the observation (or evidence). The advantage of this approach is that it makes no assumptions about the prior distribution. Also in this framework, knowledge can have alternative descriptions, without reducing the evidence's courses. Moreover, "deduction in the knowledge" can affect positively to reinforcement, something that the MDL principle or other syntactic priors avoid because the theory cannot change its syntax without changing its a posteriori probability.

Reinforcement allows a more detailed treatment of exceptions and provides different ratings for different parts of a theory, not the single probability value given by the priors which is assigned to the whole theory. Moreover, different predictions or assumptions are provided with different reliability values.

We have seen it working in the context of knowledge construction, showing that abduction is feasible as long as the theory gets reinforced. We think that the role of reinforcement in induction and abduction in knowledge acquisition is portable even from expert systems and diagnostic systems to neural networks (training=induction, recognition=abduction). It is more obvious the relation of this work with the distribution of reinforcement in neural networks, and the problems of overfitting and underfitting in the learning of linear functions. It even resembles some popular algorithms, like back-propagation. However, a symbolical framework seems an extremely adequate tool to advance and combine different areas and applications: ILP, EBL, Analogical Reasoning, Reinforcement Learning and some kinds of non-monotonic reasoning.

As future work, the measures could be extended to consider time-complexity and/or negative cases in the courses. In addition, a deeper study of how deduction affects reinforcement could be of capital interest in knowledge-based systems which use inductive and deductive reasoning techniques. Finally, we plan to apply our ideas in domains with actions, probably using situation or event calculus [22][37], and treating rewards in a more direct way (connecting with the work of [7]), in order to re-associate our notion of reinforcement with more classical notions of reinforcement learning.

Acknowledgements

Voldria agrair als revisors de la CCIA'98 per llurs comentaris, especialment per suggerir el millorament de la secció 7 i la introducció de la secció 8.

References

[1] David W. Aha, "Lazy Learning. Editorial" Special Issue about "Lazy Learning" *AI Review*, v.11, Nos. 1-5, Feb. 1997.

[2] Aliseda, A. "A Unified Framework for Abductive and Inductive Reasoning in Philosophy and AI" in M. Denecker, L. De Raedt, P. Flach and T. Kakas (eds) ECAI'96 Workshop on *Abductive and Inductive Reasoning*, pp. 7-9, 1996.

[3] Barker, S.F. *Induction and Hypothesis* Ithaca, 1957.

[4] Blumer, A.; Ehrenfeucht, A.; Haussler, D.; Warmuth, M.K. "Occam's razor" *Inf. Proc. Letters*, 24, pp. 377-380, 1987.

[5] Botilier, C.; Becher, V. "Abduction as belief revision" *Artificial Intelligence* 77, 43-94, 1995.

[6] Bylander, T.; Allemang, M.C.; Tanner, M.C.; Josephson, J.R. "The computational complexity of abduction" *Artificial Intelligence*, 49:25-60, 1991

[7] Dietterich, T.G.; Flann, N.S. "Explanation-Based Learning and Reinforcement Learning: A Unified View" *Machine Learning*, 28, 169-210, 1997.

[8] Ernis, R. "Enumerative Induction and Best Explanation" *J. of Philosophy*, LXV (18), 523-529, 1968.

[9] Flach, P. "Abduction and Induction: Syllogistic and Inferential Perspectives" in M. Denecker, L. De Raedt, P. Flach and T. Kakas (eds) *Working Notes of the ECAI'96 Workshop on Abductive and Inductive Reasoning*, pp. 7-9, 1996.

[10] Flach, P.; Kakas, A. (eds) *Abduction and Induction. Essays on their relation and integration*, in press, Kluwer 1998.

[11] Gold, E.M. "Language Identification in the Limit" *Inform. and Control*., 10, pp. 447-474, 1967.

[12] Grünwald, P. "The Minimum Description Length Principle and Non-Deductive Inference" in Peter Flach and Antonis Kakas (eds), Proceedings of the IJCAI'97 Workshop on *Abduction and Induction in AI*, Nagoya, Japan 1997.

[13] Harman, G. "The inference to the best explanation" *Philosophical Review*, 74, 88-95, 1965.

[14] Hempel, C.G. *Aspects of Scientific Explanation*, The Free Press, New York, N.Y. 1965.

[15] Hernández-Orallo, J.; García-Varea, I. "Distinguishing Abduction and Induction under Intensional Complexity" in Flach, P.; Kakas, A. *Proc. of the ECAI'98 Ws. on Abduction and Induction in AI*, to appear 1998.

[16] Hernández-Orallo, J.; García-Varea, I. "On Autistic Interpretations of Occam's Razor", http://www.dsic.upv.es/~jorallo/escrits/autistic21.ps.gz, 1998.

[17] Hernández-Orallo, J.; Ramírez-Quintana, M.J. "Inductive Inference of Functional Logic Programs by Inverse Narrowing" J. Lloyd (ed) *Proc. JICSLP'98 CompulogNet Meeting on Comp. Logic and Machine Learning*, pp. 49-55, 1998.

[18] Holland, J.H.; Holyoak, K.J.; Nisbett, R.E.; Thagard, P.R., *Induction, Processes of Inference, Learning and Discovery*, The MIT Press, 1986.

[19] Kaelbling, L.; Littman, M.; Moore, A. "Reinforcement Learning: A survey" *J. of AI Research*, 4: 237-285, 1996.

[20] Karmiloff-Smith, A., *Beyond Modularity: A Developmental Prespective on Cognitive Science*, The MIT Press 1992.

[21] Kolmogorov, A.N. "Three Approaches to the Quantitative Definition of Information" *Problems Inform. Transmission*, 1(1): 1-7, 1965.

[22] Kowalsi, R.; Sachi, F. "Reconciling the Event Calculus with the Situation Calculus" *J. Logic Prog.* 31 (1-3): 39-58, 1997.

[23] Kuhn, T.S., *The Structure of Scientific Revolution*, University of Chigago 1970.

[24] Levin, L.A. "Universal search problems" *Problems Inform. Transmission*, 9, pp. 265-266, 1973.

[25] Levinson, R. "General game-playing and reinforcement learning" *Computational Intelligence*, 12(1): 155-176, 1996.

[26] Li, M.; Vitányi, P., *An Introduction to Kolmogorov Complexity and its Applications*, 2nd Ed. Springer-Verlag 1997.

[27] López de Mántaras, R.; Armengol, E. "Machine Learning from examples: Inductive and Lazy Methods" *Data & Knowledge Engineering* 25, 99-123, 1998.

[28] Michalski, R.S. "Concept Learning" in S.C. Shapiro (ed). *Encyclopedia of AI*, 185-194, John Wiley , 1987.

[29] Mitchell, T.M. *Machine Learning*, McGraw-Hill Series in Computer Science, 1997.

[30] Mooney, R.J. "Integrating Abduction and Induction in Machine Learning" in Peter Flach and Antonis Kakas (eds), *Proceedings of the IJCAI'97 Workshop on Abduction and Induction in AI*, Nagoya, Japan 1997.

[31] Muggleton, S.; De Raedt L. "Inductive Logic Programming — theory and methods" *J.Logic Prog.*, 19-20:629-679, 1994.

[32] Muggleton, S.; Page, C.D. "A Learnability Model for Universal Representations" Unpublished Manuscript, Oxford University Computing Laboratory, 1995.

[33] Popper, K.R., *Conjectures and Refutations: The Growth of Scientific Knowledge*, Basic Books, 1962.

[34] Rissanen, J. "Modelling by the shortest data description" *Automatica-J.IFAC*, 14:465-471, 1978.

[35] Rissanen, J. "Fisher Information and Stochastic Complexity" *IEEE Trans. Inf. Theory*, 1(42): 40-47, 1996.

[36] Schmidhuber, J.; Zhao, J.; Wiering, M. "Shifting Inductive Bias with Success-Story Algorithm, Adaptive Levin Search, and Incremental Self-Improvement" *Machine Learning*, 28, 105-132, 1997.

[37] Shanahan, M. "Explanation in the Situation Calculus" Proceedings of IJCAI'93, pp. 160-165.

[38] Shapiro, E. "Inductive Inference of Theories form Facts" Research Report 192, Dep. of Computer Science, Yale Univ., 1981, also in Lassez, J.; Plotking, G. (eds.) *Computational Logic,* The MIT Press 1991.

[39] Solomonoff, R.J. "A formal theory of inductive inference" *Inf. Control* v.7, 1-22,Mar., 224-254,June 1964.

[40] Sutton, R.S. "Special issue on reinforcement learning" *Machine Learning*, 1991.

[41] Valiant, L. "A theory of the learnable" *Communication of the ACM*, 27 (11), pp. 1134-1142, 1984.

[42] van den Bosh, *Simplicity and Prediction*, Master Thesis, Dep. of Science, Logic & Epistemology of the Fac. of Philosophy at the Univ. of Groningen, 1994.

[43] Vitányi, P.; Li, M. "On Prediction by Data Compression", *Proc. 9th European Conference on Machine Learning*, Lecture Notes in AI, Vol. 1224, Springer-Verlag, 14-30, 1997.