Generalization and Explanation Revisited

Luca Console Lorenza Saitta

Università di Torino, Dip. di Informatica
Corso Svizzera 185, 10149 TORINO (Italy)
Phone: (+39) 11- 7429.111 Fax: (+39) 11- 751.603
{lconsole,saitta}@di.unito.it

Introduction

In the last years the interest toward explanation grew considerably in AI researches. Explanations, in fact, aim at making apparent the motivations and the mechanisms underlying observed phenomena, helping reducing systems brittleness and improving systems performance (in particular, in the paper we shall limit tom a deductive notion of explanation). The paper addresses two related topics, namely, a discussion of the role of non truth-preserving reasoning mechanisms (such as induction and abduction) in explanation-seeking tasks and a critical analysis of the very notions of *generality* and *explanation.* The aim is twofold: on the one hand, a logical framework, in which abduction and induction can be precisely defined by introducing a generalized notion of *explanation,* is proposed. On the other hand, the several definitions, put forward in the machine learning literature for the *more-specific-than* relation in First Order Logic, will be critically revised and compared, pointing out some underlying implicit misunderstandings and inconsistencies. As a side-effect, the paper shows how different tasks, such as abduction related ones (e.g. diagnosis) and learning, can be unified into a unique framework, in which they are nothing else than two aspects of the same process, i.e., providing an explanation for a set of observations.

2. The "More-Specific-Than" Relation Revisited

Let us consider a function-free First Order Logic language $\mathbf{L}$. Let $\mathbf{P}$ be the set of basic predicates and Ω the set of individuals of the universe. The formulas of L can be partitioned into two subsets: the *open* formulas (called *concepts* following Frege) , with some occurrence of free variables, and the *closed* one, with no free variables (sentences).A concept does not have a truth value associated with it; rather, it partitions Ω into the concept *extension* and its complement. The concept extension is the set of individuals or tuples of individuals which satisfy the concept definition. On the other hand, predicates which are true of a given n-tuple $<a_1, a_2, \dots, a_n> \in \Omega^n$ are said to belong to the "intension" of that n-tuple. Intension and extension are dual properties of concept hierarchies. A certain confusion between these two aspects has influenced some of the current definitions of the *more-specific-than* (or of its inverse *more-general-than*) relation. The basic idea behind the definition of the *more-specific-than* relation, that we propose, derives from the simple observation that specificity (or generality) is an essentially *extensional* property and, hence, it only pertains to concepts, i.e. open formulas, that have an associated extension. Closed formulas (i.e. sentences) are statements about the generality of the associated concepts and, hence, they convey information about their degree of generality. Concepts become sentences by binding their variables. Sentences tells us something about the extension of a concept. Thus, while concepts can be compared on the basis of their specificity, sentences can be compared on the basis of their informativeness.

Definition 2.1 – Given two concepts $f(x_1,x_2,...,x_n)$ and $g(x_1,x_2,...,x_n)$ belonging to the language L, the concept $f(x_1,x_2,...,x_n)$ will be said *more specific than* the concept $g(x_1,x_2,...,x_n)$ (denoted by $f \mid< g$ [Michalski, 1983]) iff:

$$\models \forall x_1,x_2,...,x_n \, [f(x_1,x_2,...,x_n) \rightarrow g(x_1,x_2,...,x_n)]$$

where $\rightarrow$ denotes material implication.

As an example, $\models \forall x \, [\text{square}(x) \wedge \text{big}(x) \rightarrow \text{square}(x)]$. Definition 2.1 is consistent with the universally accepted extensional interpretation of the *more-specific-than* relation, according to which:

$$f \mid< g \quad \text{implies} \quad EXT(f) \subseteq EXT(g)$$

However, the inverse is not true: co-extensionality does not mean equivalence in generality. As Plotkin [1970] and Buntine [1988] point out, it would be much more useful, in practice, to have a generality ordering that takes into account a domain theory T:

Definition 2.2 – Given a theory T and two concepts $f(x_1,x_2,...,x_n)$ and $g(x_1,x_2,...,x_n)$ belonging to L, the concept $f(x_1,x_2,...,x_n)$ will be said *more specific than* the concept $g(x_1,x_2,...,x_n)$ *with respect to* T (denoted by $f \mid<_T g$) iff:

$$T \models \forall x_1,x_2,...,x_n \, [f(x_1,x_2,...,x_n) \rightarrow g(x_1,x_2,...,x_n)]$$

where $\rightarrow$ denotes material implication.

As mentioned above, generality is an extensional property and, hence, it pertains only to concepts (open formulas). In order to compare sentences, we will introduce a relation taking into account the amount of information that a true sentence conveys about the generality of the associated concept. Given a universe of discourse Ω (containing all the individuals of the domain), a (possibly empty) domain theory T and a set of predicates P, a *possible world* is a set of assignments, consistent with T, of truth values to every basic n-ary predicate for each n-tuple of objects in Ω^n ($n \geq 1$). Let W be the set of possible worlds. Let us suppose, for the sake of simplicity, that both Ω and P are finite; then, W is also finite. Let, moreover, $|\Omega| = N \neq 0$ and $|W| = M$. The amount of confusion we have about the true world, given a certain knowledge K, can be quantitatively measured by the entropy value [Shannon & Weaver, 1949]:

$$H(\mathcal{W} \mid \mathcal{K}) = -\sum_{w \in \mathcal{W}} p(w \mid \mathcal{K}) \cdot \lg p(w \mid \mathcal{K})$$

where $p(w \mid K)$ is the probability that w is the actual world, given K, and K is a subset of sentences of L known to be true in the actual world. If nothing is known about the actual world (i.e., $K = \varnothing$), every element of W is a candidate to be the actual world; if every world has the same probability, then the entropy reaches its maximum value, $H_0 = \lg M$. The minimum value, $H_{min} = 0$, is reached when all the worlds, but one, have a null probability (complete knowledge). If the information K consists of a domain theory T, the maximum value H_0 is reduced to a value $H(W \mid T) \leq H_0$, because the only possible worlds are the models of T. Obviously, the lower $H(W \mid T)$, the more informative (constraining) the theory is.

<u>Definition 2.3</u> – Given two sentences φ and ψ of L (and, possibly a theory T), φ will be said *more informative than* ψ (denoted by $\varphi >_T \psi$) iff:

$$T \models (\varphi \rightarrow \psi)$$

<u>Property 2.1</u> – Given a theory T and two sentences φ and ψ, if $\varphi >_T \psi$, then $W_T(\varphi) \subseteq W_T(\psi)$ and $H(W \mid T, \varphi) \leq H(W \mid T, \psi)$.

As in the case of generality, equivalence in informativeness implies equality in information content, but not vice-versa.

3. Towards a General Definition of Explanation

Given a theory T and some observation OBS, an explanation is a set of formulae E such that OBS follows from T and E. In the traditional definition of abduction, OBS and E are constrained to be atomic formulae (usually ground). However, the same pattern can provide also a declarative definition of induction (see, e.g., [Genesereth & Nilsson 87]). We can therefore provide a general declarative definition of explanation; we shall then distinguish between abductive and inductive explanations.

<u>Definition 3.1</u> – Given a domain theory T and a set of sentences α_{obs}, describing observations performed on the world, a set E of sentences will be said an *Explanation* for α_{obs} (with respect to T) iff the following conditions are satisfied:

(a) $T \cup E \models \alpha_{obs}$ (b) $T \cup E$ is consistent (c) $T \not\models \alpha_{obs}$

(d) E has some additional properties that make it an "interesting" explanation.

Condition (d) allows to single out a set of preferred explanations to be selected among a set of possible alternatives. There are, in particular, many criteria that have been defined to such a purpose, for example: avoiding redundant explanation, avoiding trivial explanations, imposing a language for explanation.

In the following we will try to formalize our intuitive understanding of the inductive and abductive nature of explanations: an inductive hypothesis "explains" a phenomenon by asserting that the same phenomenon has been already observed several times in the past, under the same circumstances. An abductive hypothesis "explains" a phenomenon by specifying enabling conditions (in particular, "causes") for it. As a consequence, abduction needs a domain theory, induction does not. We believe that it is very important to establish the nature of an explanatory hypothesis, because this nature evidentiates the grounds on which the hypothesis relies and, hence, supplies relevant information for assessing its credibility. If we want to explain, for instance, that the light appears in a bulb when we turn a switch on, an inductive explanation would say that this happens because it happened hundreds of times before, whereas an abductive one could supply an explanation in terms of the electric current flowing in the bulb filament. If, some time, turning the switch on does not produce the phenomenon, the inductive explanation just fails, whereas the abductive one can supply hints for repairing the failure.

Before formally characterizing this difference, let us consider an example.

Example 1 – Suppose, first, that we are given the following $T \equiv \forall x\ [p(x) \rightarrow q(x)]$ and the observation $\alpha_{obs} \equiv q(a)$. Three alternative explanations for α_{obs} are:

$E_1 = p(a)$ $\qquad\qquad E_2 = \{\forall x\ q(x)\}$ $\qquad\qquad E_3 = \{\forall x\ p(x)\}$

We would like to say that E_1 has an abductive nature; $E_{2\ n}$as an inductive nature and E_3 as a hybrid abductive/inductive nature.

This labelling of hypotheses as abductive or inductive or abductive/inductive tries to capture the intuitive feeling that the only support of an inductive explanation is a supposed "similarity" between unobserved individuals and observed ones. In other words an inductive hypothesis allows the validity of properties, observed on a set of individuals, to be extended to other individuals not in the observations (and for which the property was unknown), whereas an abductive one allows unobserved properties to be applied to observed individuals. Hence, the generation of an inductive hypothesis does not need a theory relating properties to each other, whereas the generation of an abductive hypothesis does. Notice that the distinction between inductive and abductive hypotheses strictly parallels the dichotomy "extension vs. intention". In fact, inductive hypotheses are related to (concept) extensions (i.e., they extend the extension of observed properties, or "they generalize observed properties to other individuals), whereas abductive hypotheses are related to (individual) intensions (i.e., they extend the intension of individuals in the observation - that is the set of properties that are true for such individuals). the explanations that are abductive/inductive are in some sense hybrid since the extend observed properties to other individuals not in the observation but they need the domain theory in order to do that. Notice that we do not want to say that this is the right classification; however we believe that it has interesting advantages.

In the following we provide the formal definition in the simpler case

__**Definition 3.2**__ – Given a domain theory T, a universe of individuals Ω, an observation $\alpha_{obs} \equiv p(X) \equiv_{Def} p(a_1) \wedge \ldots \wedge p(a_n)$, where $X = \{a_1, \ldots, a_n\}$ is the set of individuals for which "p" has been observed, an explanation E for α_{obs}, satisfying Definition 3.2, will be said:

- *(Purely) Inductive* iff:

  E ˆ p(Y) ˆ p(X) (with $X \subset Y$)
- *Abductive/Inductive* iff:

  $T \cup E$ ˆ p(Y) ˆ p(X) (with $X \subset Y$) and $E \odot p(X)$

- *(Purely) Abductive* iff:

  $T \cup E$ ˆ $p(X)$ and $E \odot p(X)$ (unless $E=p(X)$) and $T \cup E \odot p(Y)$ with $X \subset Y$

Definition 3.3 can be generalized to the case where more than one predicate symbol occurs in the observation, and to the case where the observation is not ground and contains quantified formulae (the reader interested in such formal definition is referred to the longer version of this paper).

4. Examples

We analysed many different approaches to explanation in the A.I. literature analysing their abductive/inductive nature according to our classification. Some of the results are as expected and can be summarized as follows:

- the explanation produced by the classical definition of abduction as used in model-based diagnosis (or planning) have an abductive nature
- learning from examples produces explanation which have an inductive nature
- the operators of inverse resolution used in inductive logic programming have inductive nature (truncation) or abductive-inductive nature (absorpion and identification), see also [Console et al. 91].

5. Related Works

There are at least three aspects to be considered in comparing our work to related ones (the reader interested in a detailed discussion is referred to the longer version of the paper available from the authors).

The *More-Specific-Than* Relation. Several alternative definitions of the *more-specific-than* relation have been proposed to compare First Order Logic formulas. However, all these definitions fail to capture, in our opinion, the very nature of generality and this misunderstanding produces several undesirable consequences. Most of the definitions in the literature suffer of some problems (see e.g., the definition of *θ-subsumption* between clauses in Plotkin [1970] or the definitions in [Vere, 1978; Buntine, 1988; Dietterich & Michalski, 1983; Kodratoff & Ganascia, 1986; Niblett, 1988, Muggleton 91] just to cite some examples). The first among these is a confusion between two different dichotomies, i.e., "general *vs.* particular" and "generic *vs.* specific", generated by the use of the same term "specific" for both. In fact, the first dichotomy relates to extensional properties of concepts, whereas the second one to intensional properties. By using the two separated dimensions proposed here, the problems can be avoided. The second negative consequence, related to the effort of attributing generality to closed formulas, is an internal contradiction. In fact, almost all those cited above ([Niblett, 1988] is an exception) would agree in defining the formula $p(a) \land p(b)$ more specific that $p(a)$ and the formula $\forall x\, p(x)$ more general than $p(a)$. However, this is an internal contradiction.

General frameworks and inductive logic programming. Interesting general framework to induction have recently proposed by Helft [89] and Muggleton [91]. The former argued that defining induction in terms of explanation has some drawbacks; he then proposes a new definition. He provides a definition (based on non-monotonics logic) which is a special case of our framework and, moreover, fails to capture some important aspects concerning the different nature of the theory and the observations. Muggleton introduces operators for inductive logic programming; we showed that such operators provide explanations with inductive or abductive/inductive nature according to our framework.

Abduction and Induction. Some attempts to compare and or unify abduction and induction can be found in [Michalski 1991,Kodratoff, 1991] who, however simply identify the two mechanisms without recognizing their different nature. Josephson [1991] suggests that inductive generalization is a special form of abduction but he fails to provide a definition for distinguishing between the two.

References

W. Buntine (1988): "Generalized Subsumption and Its Applications to Induction and Redundancy", *Artificial Intelligence, 36,* 149-176.

L. Console, A. Giordana, L. Saitta (1991): "Investigating the Relationships between Abduction and Inverse Resolution in Propositional Calculus", LNCS 542, Springer Verlag, pp. 316-325.

T. Dietterich, R.S. Michalski (1983): "A Comparative Review of Selected Methods for Learning from Examples". In R. Michalski, J. Carbonell & T. Mitchell (Eds.), *Machine Learning: An A.I. Approach, Vol. I,* Morgan Kaufmann , pp. 41-81.

M. Genesereth, N. Nilsson (1987): "Logical Foundations of Artificial Intelligence", Morgan Kaufmann.

N. Helft (1989): "Induction as non-monotonic inference", Proc. *KR 89* , pp. 149-156.

J.R. Josephson (1991): "Abduction: Conceptual Analysis of a Fundamental Pattern of Inference", Technical Report, Ohio State University, LAIR.

Y. Kodratoff, J. G. Ganascia (1986): "Improving the Generalization Step in Learning". In R. S. Michalski, J. G. Carbonell & T. Mitchell (Eds.) *Machine Learning: An A.I. Approach, Vol. II,* Morgan Kaufmann, pp. 215-244.

Y. Kodratoff (1991): "Induction and the Organization of Knowledge", Proc. *Multi-Strategy Learning Workshop* (Harpers Ferry, WV), pp. 34-48.

R. Michalski (1983): "A Theory and Methodology of Inductive Learning", In R. S. Michalski, J. G. Carbonell & T. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach, Vol. I,* Morgan Kaufmann (Los Altos, CA), pp. 83-134.

R. Michalski (1991): "Inferential Learning Theory as a Basis for Multi-Strategy Task-Adaptive Learning", Proc. *Multi-Strategy Learning Workshop* , pp. 3-18.

S. Muggleton (1991): "Inductive Logic Programming", *New Generation Computing, 8,* 295-318.

T. Niblett (1988): "A Study of Generalisation in Logic Programs", Proc. *EWSL-88* (Glasgow, UK), pp. 131-138.

G. Plotkin (1970): "A Note on Inductive Generalization". *Machine Intelligence, 5,* 153-163.

C. Shannon, W. Weaver (1949): *The Mathematical Theory of Communication,* University of Illinois Press (Urbana, IL).

S. Vere (1978): "Inductive Learning of Relational Productions", in *Pattern-Directed Inference Systems,* Academic Press, pp. 281-295.