Ames

IN-65-TM

146701

[redacted]

150

Induction of Models Under Uncertainty

Peter Cheeseman

NASA Ames Research Center
Mail Stop 244-7
Moffett Field, CA 94035

(NASA-TM-101173) INDUCTION OF MODELS UNDER UNCERTAINTY (NASA) 15 p CSCL 12A

N88-24230

Unclas
G3/65 0146701

Abstract

This paper outlines a procedure for performing induction under uncertainty. This procedure uses a probabilitic representation and uses Bayes' theorem to decide between alternative hypotheses (theories). This procedure is illustrated by a robot with no prior world experience performing induction on data it has gathered about the world. The particular inductive problem is the formation class descriptions both for the tutored and untutored cases. The resulting class definitions are inherenty probabilistic and so do not have any sharply defined membership criterion. This robot example raises some fundamental problems about induction—particularly it is shown that inductively formed theories are *not* the best way of making predictions. Another difficulty is the need to provide prior probabilities for the set of possible thoeries. The main criterion for such priors is a proagmatic one aimed at keeping the theory structure as simple as possible, while still reflecting any structure discovered in the data.

1 Introduction

Historically, the scientific (hypothetical-deductive) method is the most successful known method for inducing models under uncertainty, and AI can profit by emulating this approach. This method consists of the following cycle:

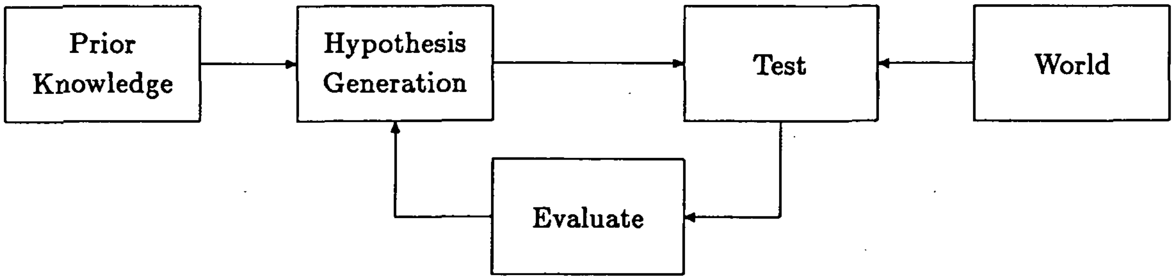

Figure 1: The Hypothetical-Deductive Cycle

For example, the cause of AIDS (Aquired Immune Deficiency Syndrome) took many such cycles to identify. Initially, there were many hypotheses that fitted the available evidence (mostly related to the lifestyles of the known risk groups), however all but one of these hypotheses failed to fit more detailed information. The blood transfusion victims in particular made the theory that there was some sort of infectious agent carried by blood the most likely. Futher iterations of this cycle finally made the specific theory that a particular virus was responsible the most likely theory.

Adapting this style of reasoning to AI poses a number of questions, noted below:

1. How can evidence be used to chose among competing theories?
2. How are theories (hypotheses, models etc.) refined with more evidence?
3. How to represent possible theories (including families of theories)?
4. How are theories (hypotheses, models etc.) generated?
5. How to utilize prior knowledge (general background information) in the derivation of theories from evidence and in evaluation of the resulting theories.

The AI literature on induction (e.g. [12]) has concentrated on "logical" induction—i.e. induction with noise free examples. Such procedures generally find the simplest theory that satisfies the given examples. Since a single example can eliminate many possible theories, noise (errors) in the data can cause the correct theory to be eliminated [9]. Actual logic based induction procedures use ad hoc methods to provide a degree of robustness with noisy data. In fact, some induction procedures, such as ID3 [14], are implicitly based on probability because they internally use entropy criterion and significance tests, even though the end result is catagorical.

Surprisingly, the basic theory for induction with noisy data has been available for many years (e.g. [15]), but is largely unknown in the AI literature (see [10] for an exception). This paper outlines this theory, with the help of a robot example. The remaining problems have been the mainstay of philosophy (particularly philosophy of science), without a consensus emerging. AI is well-suited to making progress in this area because the need to write working programs cuts through the verbal maze created by the philosophers. Another aim of this paper is to present an initial attempt at answering the remaining questions.

2 A Robot Example

To avoid the problems associated with induction in a rich semantic context, the example developed in this paper concerns a (hypothetical) robot that is switched on for the first time in a new environment. This robot is assumed to have no prior knowledge of the environment, and so the problem of how to relate new information to an existing knowledge framework is avoided. To keep the example concrete, imagine that the new environment is a normal office building with rooms, corridors, doors etc. Also, assume that the robot is equiped with high-level sensors that report attributes, such as, the existence, type, color and location of objects, and relations between objects. Objects include such items as chairs, tables, doors, etc. Such sensors do not yet exist, but nothing in the following argument depends on the semantic level of reporting by the sensors—the argument could be developed (less convieniently) even for pixel-level information.

The robot is programmed so that when it is switched on it begins actively exploring to improve its knowledge of the new environment. It uses it sensors to discover "primative" facts about its new world, such as properties of observed objects and relations among them, and stores this information in its data base. Note that "primative" here means directly observable by the sensors. For example, it might discover and store such predicates as "in(room27, table3)", "on(book43, table3)" and "open(door4)". At this stage, the robot is only capable of answering questions about what it has (and has not) observed. This information is the raw material for performing induction, as described below.

3 Untutored Learning

We may ask our robot (or it may ask itself) "Is there a terminal in room27 ?", and the necessary information is not in the robot's data base. Rather than just answering "I don't know", we would like our robot to answer "the probability is .8", based on information it has gathered itself, such as: "room 27 is an office, and the probability of an office having a terminal is .8". Unfortunately, there is no *logical* basis for making such statements! This is partly because logic is only capable of answering true or false, and partly because no finite amount of evidence can lead to logical implications, such as "If office(x) then in(x,terminal(x))". Probability provides the necessary generalization of the concept of truth (a measure of belief), by assigning a number (between 0 and 1) to any logical expression, given the relevent evidence (see [3] for an overview of the use of probability in AI). The standard "axioms" of probability are logical consequences of very basic properties that we would require of any calculus of belief using real numbers (see [7] for a proof of this statement).

The only way a robot can make useful probability statements about its world is if it finds "patterns" (models, associations, theories etc.) that significantly differ from chance expectations. These patterns can be used to make probabilistic predictions in particular cases. In the robot example, the only type of pattern considered is (static) classification. That is, the robot looks for evidence in the data that a set of objects are sufficiently similar to each other (i.e. they have sufficient attribute values in common) that they should be considered as a class. The resulting class (a set of similar objects) can be used to assign a probabilty that a new object is a member of the class, based on the similarity of its attribute values to other examples of the class. This probabilistic definition means that there is no sharp membership criterion—there is only a best match criterion. The probabilistic interrelationships among attributes within a class can be used to make (conditional) probabilistic predictions about unobserved attributes, given the known attribute values. This probabilistic behavior is very different from classical set theory. Many other types of "patterns" could be utilized—for example, the robot could search for temporal patterns (e.g. room occupancy) or joint probabilities (e.g. filing cabinets and desks tend to occur together) or spatial patterns, etc. The inductive methods illustrated using the classification example can be developed for other such theoretical frameworks.

4 Induction Procedure

The robot's class induction procedure requires a criterion for deciding whether a particular partition of the set of objects into classes is better than another. This criterion is used in combination with a search procedure to find the best classification. The result could be a single class (which is likely to occur if the procedure is given a random set of objects) or many well separated classes. A brute-force search procedure would generate all possible class hypotheses and test them against the best current hypothesis. Such a procedure is too computationally expensive in nontrivial cases. A much better procedure would only generate good class hypotheses, based on observed similarities among the objects. Unless the good candidate generation procedure is guaranteed to include the best, it is possible that the more efficient procedure will select a sub-optimal class hypothesis. This section defines the decision criterion for chosing one class hypotheses over another.

The first step is to chose the set of objects for which a classifaction is sought. In this particular example, the robot considers the set of all the rooms it knows. The next step is to build an ordered set of attribute values for each room, such as its size, how many doors, contains a desk (or not), etc., for all the room attributes known to the robot. Note that relations, such as "connects(door3, room2, room5) and connects(door4, room2, room9)" are translated into attributes such as "door-count(room2, 2)". There is a practical problem in chosing a "small" relevent subset of attributes when there are a very large number of possible attributes, however, the robot initially has no information on which to base relevency decisions. When the robot has gained a much better understanding of is environment, it can use information it has discovered, such as associations between different attributes, to restrict its inductive procedures to combinations of attributes that are likely to be significant. In this example, we assume that the robot uses all the primative attributes known to it.

The next step is for the robot to test whether these room descriptions provide significant evidence for assigning the rooms to different classes. The result of such an analysis would be to produce class descriptions that roughly correspond to our concepts of "office", "conference room", "bathroom" etc. Note that these classes are produced by the robot discovering characteristic patterns in the attributes associated with individual rooms themselves, and not by the robot being told the classification of each room in advance (i.e. the robot is performing untutored learning).

The fundumental classification question is "Are these individual examples (rooms) "clustered" in attribute space?". That is, do some rooms have more attribute values in common with each other than the rest. The pattern recognition literature contains many criteria for clustering [8], all based on some sort of "similarity" measure. However, this richness is an embarassment—how should the robot chose from this set of *ad hoc* measures? Even worse is the problem of deciding whether the observed clustering (under a particular similarity measure) represents a real effect or could be attributed to chance. Most systems contain a parameter that sets this threshold, but it is not clear how this parameter value should be chosen for our robot.

Fortunately, there is a class induction criterion that can be derived from first principles. This criterion uses Bayes' theorem to assign a relative probability between two class hypotheses H_i and H_j. A class hypothesis H_i is a particular partition of the known objects into specific sets (classes)—each class is implicitly defined by the set of objects (examples) assigned to it.

The *relative* probability of these hypotheses given the data D, by Bayes' theorem is:

$$\frac{P(H_i/D)}{P(H_j/D)} = \frac{P(H_i)}{P(H_j)} \frac{P(D/H_i)}{P(D/H_j)} \tag{1}$$

If the robot has no prior knowledge implying that one class hypothesis with the same number of classes is more likely than another, it assigns them equal probability. This is just an application of the principle of indifference. In this case, in Eqn. 1, this means that $P(H_i) = P(H_j)$, so that the term involving the relative prior probabilities can be ignored. However, evaluation of the relative prior probability of hyptheses with different numbers of classes is more complex. A person may have prior knowledge that favors a particular number of classes, but our robot is not as privileged. In the absence of prior information, the robot should pragmatically favor hypotheses with fewer classes to reduce the number of classes it has to remember and process. That is, it should prefer hypotheses with fewer classes unless the data provides sufficient evidence to suggest further class divisions are justified. Here, Occam's razor is being used—not because the world is simple, but because the robot wants it to be simple (for computational reasons) and only accepts greater complexity when it gives significantly improved predictive power. A relative prior probability with this property is:

$$\frac{P(H_i)}{P(H_j)} = \frac{N_j}{N_i} \tag{2}$$

where N_i is the number of classes associated with H_i, etc. This particular choice biases against a large number of classes, and is discussed further in section 6. This leaves the problem of determining $P(D/H_k)$, for a class partition denoted by H_k. In Eqn. 1, D is the entire set of examples, which can be further partioned into subsets (D_l) corresponding to the particular class partition H_k. That is:

$$P(D/H_k) = P(D_1/C_1) \cdots P(D_l/C_l) \cdots P(D_n/C_n) \tag{3}$$

i.e., a product of component probabilities, where C_l is the lth class under the class hypothesis H_k, and D_l is the set of examples that define the class C_l. That is, Eqn. 3 shows the explicit partition of the examples into independent classes under the hypothesis H_k. This expansion assumes that information about the members of one class is non-informative about member of another class (i.e. the classes are independent). This assumption is not true, for example, for hierachical classes.

Each example e_i from the set $D_l = \{e_1 \cdots e_m\}$, is described by an ordered set of attribute values, where each attribute value is drawn from a fixed set of possible values associated with each attribute. The probability of an example is dependent on the other examples of the same class, i.e.

$$\begin{aligned} P(D_l/C_l) &= P(\{e_1, \cdots, e_m\}/C_l) \\ &= P(e_1/C_l)P(e_2/e_1, C_l)P(e_3/e_1, e_2, C_l) \cdots P(e_m/e_1, \cdots, e_{m-1}, C_l) \end{aligned} \qquad (4)$$

This equation is just the multiplication theorem of standard probability theory corresponding to a particular expansion order. The expansion order is just the particular order in which the examples are presented, so that the expected probability of the next example can depend on all the previously presented examples of the same class. An important property of probability theory is that the joint probability is the same regardless of expansion order. The problem is now to calculate terms such as:

$$P(e_p/e_1, \cdots, e_{p-1}, C_l) \qquad (5)$$

Each e_p consists of an ordered set of attribute values, so the desired probability can be expanded as:

$$P(< a_1, \cdots, a_n >_p \, / < a_1, \cdots, a_n >_1, \cdots, < a_1, \cdots, a_n >_{p-1}, C_l) \qquad (6)$$

That is, the probability of the ordered set of attribute values of a particular example is conditioned by all the previously given examples of the same class. If the previous examples are strongly predictive, the conditional probabilities in Eqn. 6 will be different from the probability that would result from assigning all examples to the same class. As in Eqn. 4, the set of attributes can be expanded to give:

$$\begin{aligned} P(< a_1, \cdots, a_n >_p /e_1, \cdots, e_{p-1}, C_l) &= P(a_1/e_1, \cdots, e_{p-1}, C_l)P(a_2/e_1, \cdots, e_{p-1}, < a_1 >, C_l) \cdots \\ &\quad \cdots P(a_n/e_1, \cdots, e_{p-1}, < a_1, \cdots, a_{n-1} >, C_l) \end{aligned} \qquad (7)$$

Now, the remaining task is to show how to evaluate terms in the above equation, such as: $P(a_q/e_1, \cdots, e_{p-1}, < a_1, \cdots, a_{q-1} >, C_l)$ where a_q is a particular attribute that can assume one of an ordered set of possible values; i.e. $a_q =< v_1, \cdots, v_r, \cdots, v_Q >_q$. The above sequence of decompositions can be summarized as follows:

$$\begin{array}{lcll} D_l & = & \{e_1 \cdots e_p \cdots e_m\}_l & \text{—the set of examples in the } l\text{th class} \\ & & \overbrace{< a_1 \cdots a_q \cdots a_n >_p} & \text{—an ordered set of attributes for the } p\text{th example} \\ & & \overbrace{< v_1 \cdots v_r \cdots v_Q >_q} & \text{—an ordered set of attribute values for the } q\text{th attribute} \\ & & v_{rqpl} & \text{—a particular value} \end{array}$$

That is, our final goal is to be able to calculate:

$$P(v_{rqpl}/e_1 \cdots e_{p-1}, < a_1 \cdots a_{q-1} >) \tag{8}$$

where v_{rqpl} is the rth value of the qth attribute of the pth example in the lth class, given all the previous $(p-1)$ examples and all earlier attribute values of the current example.

4.1 Independent Attributes Case

The simplest approach is to assume that each attribute value is independent of other attributes of the same class. In this case Eqn. 8 is just:

$$P(v_{rqpl}/v_{rq1l}, v_{rq2l}, \cdots, v_{rq(p-1)l}) \tag{9}$$

That is, in this case the desired probabilities are only dependent on the total frequency of occurence of the particular attribute value in the examples examined so far. The required probability is given by:

$$P(v_{r,q,p,l}/v_{r,q,1,l} \cdots v_{r,q,(p-1),l}) = \frac{(n_{r,q,p-1,l}) + \frac{1}{2}}{(N_{q,p-1,l}) + \frac{Q}{2}} \tag{10}$$

where $n_{r,q,p-1,l}$ is the number of occurences of the rth value of the qth attribute in all the $p-1$ examples examined so far in the lth class, and $N_{q,p-1,l}$ is the total number of occurences of the qth attribute in the examples so far, and Q is the total number of possible values for the qth attribute. Note that this formula gives a meaningful value (i.e. 1/Q) even when N is zero. This formula is the mean probability derived from inverting the Multinomial Distribution using Bayes' theorem. The derivation uses a locally non-informative prior probability as described in Box and Tiao [1].

Using Eqn. 10 it is possible to compute the required probability ratio in Eqn. 1 by unwinding the sequence of decompositions given above. In essence the method is just serially computing the probability of each observed attribute value for each example of a particular class, in order of presentation. This probability is just a frequency ratio based on the already known examples. Once the probability of the new example is computed, the observed attribute value increments both n and N in Eqn. 10, and so changes the expected probability of the next example, and so on. Because of the fundamental properties of probabilities, the order of evaluation is not significant.

4.2 Dependent Interactions

In section 4.1 we assumed that the probability of individual attribute values for examples of a particular class were independent of each other. That is, knowing the value of an attribute (e.g. size(room27, large)) is non-informative about any other attribute value of the same example (e.g. door-count(room27, 3)). Clearly, this is not true in general. Fortunately, the theory presented above can be extended to accomodate any interactions between attribute values (not attributes) that the robot might discover.

The fundumental assumption that must be made in order to make useful predictions is that attributes are independent of each other *unless there is significant evidence to the contrary*, refered to as causal closure. In principle, every attribute could be dependent on every other attribute, but any system (human or robot) that uses this approach will never make useful predictions in any realistic situation! The reason for this impass is that if a required probability is conditioned on all situations with exactly the same combination of attribute values as the case of interest, then it is unlikely that any such exact match will occur. That is, situations almost never exactly recur, so there usually no prior cases on which to base a conditional probability calculation. Because of this situation, we (and the robot) make the opposite assumption— that every attribute is independent of every other unless there is significant evidence to the contrary. As a result, the robot will often make decisions that would have been different if it had known about other relevent dependencies. When decisions must be made with incomplete information, use of the independence assumption seems inescapable.

The procedure for searching for significant combinations of attribute values is conceptually simple, but computationally expensive. Essentially, it requires testing each possible combination of attribute values to see if the observed frequency is significantly different from chance expectations. If so, the corresponding observed joint probability becomes a constraint on the underlying total joint probability space, as described in [5]. However, a particular conditional probability is difficult to calculate if the known joint probability constraints overlap in complex ways. If the constraints form a tree structure then the probabilities are conditionally independent, and are easy to calculate [6],[13]. For non-conditionally independent cases, a generalized notion of independence is necessary to give a particular probability value, subject to the already discovered constraints. Generalized independence is achieved if the underlying total joint probability distribution is chosen that maximizes the total joint probability distribution entropy, subject to the known probability constraints. The entropy of the total probability distribution is defined to be:

$$H = -\Sigma_{ijkl\ldots} P_{ijkl\ldots} \log P_{ijkl\ldots} \qquad (11)$$

where $P_{ijkl\ldots}$ is the probability of the ith value of the first attribute etc.—i.e. it is an element of the total joint probability space. If there are a large number of attributes, the cost of finding the maximum entropy distribution using Eqn. 11 directly is prohibitive. However, by using Lagrange multipliers and carefully factoring the resulting form, the maximum entropy distribution subject to the known constraints can be evaluated with low computational cost [4]. This means that it is possible to find the conditional probability of any attribute value given any particular combination of attribute values and the known probability constraints. These conditional probabilty calculations include all the probability constraints between attribute values, as well as the specific information about particular attribute values of the current example.

The remaining problem is: "How does the robot discover what are the significant constraints?". We assume that the only form of constraints that the robot searches for are joint probabilities of combinations of attribute values (from different attributes), because these are symmetric and all other constraints can be defined in terms of them. Given the robot's initial lack of knowledge, there is no alternative to a brute-force search through the set of possible combinations of attributes. Again, this search is too expensive in general, and so the robot should start by searching for significant *pairs* of attribute values and only look for higher order combinations if the cost of the search is justified by the expected improvement in the classification that results. Prior knowledge of independence of attributes could dramatically reduce the robots search. For example, knowledge about how wall colors are selected for particular rooms could be used to exclude color as a useful room attribute. How to utilize prior knowledge to select only the relevent combinations of attributes (and thereby drastically reduce the search required) is largely a mystery.

It might seem that the probabilistic class criterion, Eqn. 1, would favor the creation of a large number of classes, each containing only a few members that are particularly "close" to each other. However, Eqn. 1 using the prior probabilities of Eqn. 2 favors hypotheses with a smaller number of classes. Also, Eqn. 10 is a combination of prior information (the Q term) and the observed frequency information, so it requires many examples before the expected probabilities significantly differ from prior expectations. The result of these two effects is that unless the members of a class are strongly predictive about the attributes of each other, it requires many examples to justify a particular class. In particular, for a random set of examples, it is very unlikely that the examples will be sufficiently similar to each other for Eqn. 1 to justify multiclass hypotheses.

5 Tutored Learning

The method described in the previous section can be easily adapted to deal with the tutored learning case. This case would arise for our robot if the user wishes to communicate his understanding of the world to the robot—that is to make sure that the classes the robot induces correspond to those familiar to the user. Because the robot has different sensing capabilities and goals, the robot's untutored classes may be very different from the user's. This situation is similar to the Eskimos seeing (or at least naming) forty three different kinds of "snow" that are imperceptible to those who have no need to distinguish.

The method is for the user to supply the class labels for the individual examples, and require the robot to try to find the corresponding class definition. This definition is necessarily probabilistic because of Eqn. 3. There is no guarantee that the class description the robot generates will correspond exactly with that intended by the user (who probably doesn't have an exact idea anyway).

Fortunately, the method for performing tutored learning is even simpler than the method for untutored learning. The method is to divide the examples into classes as specified by the user, then use Eqn. 3 to compute $P(D_l/C_l)$. Here, D_l are the examples (i.e. ordered sets of attribute values) and C_l is the particular class whose definition is being generated. In this case the "definition" is just the set of probabilities of attribute values (including any interactions) found by evaluating all members of the class, as explained in the previous section. With the probabilistic definitions generated for each class, it is now possible to compute the (relative) probability that a new case is a member each class. If the new case *must* be assigned to one of the known classes, then the one with the maximum probability should be selected. However, if the (relative) probabilities are normalized, the result may be that the new case does not fit any known class very well, and so indicates that the set of classes may need to be extended.

For example, the question posed earlier: "Is there a terminal in room27", can now be answered. The first step is to determine which class room27 belongs to. This requires taking all the attributes about room27 known to the robot and deciding which of the room classes it is most likely to belong to. If the answer is that room27 is most likely to be in class "office", the robot can then compute the probability of there being a terminal in an office with the known attributes of room27. This is not an optimal procedure—the robot should compute the probability of room27 being a member of each of the possible room classes then form the weighted average of the probability of there being a terminal in the room for each class. If the probability of room27 being an office is overwhelming, there is essentially no difference between the answers given by the two procedures.

Note that the probabilistic class definitions discussed above do not have a sharp membership criterion, so there does not have to be any particular set of characterisics that distinguish members from non-members. Instead, an overall probabilistic match of each new case with the expected probabilities associated with each class is what defines the "nearness" of the match. A membership decision can only be created by imposing a threshold on the probabilistic nearness function. This "fuzzy" set membership approach is close to Wittgenstein's concept of "family resemblence" [17], which he introduced to avoid the philosophical difficulties associated with the classical "sharp" concept of set membership. Note that probabilistic class membership, as introduced in this paper, is close to concept of "Fuzzy Sets" defined by Zadeh [18]. Fuzzy sets are based on the concepts of "degree of membership". The similarity of these concepts is not coincidental—it has been shown in [2] that all the typical reasoning of fuzzy logic (or equivalently with fuzzy sets) is either identical to probabilistic reasoning or counter-intuitive when they significantly differ.

6 Theory of Induction

The robot example presented in the previous sections illustrates the basic probabilistic (Bayesian) approach to induction. The idea is to compute the most likely hypothesis (theory, explanation, model, etc.) out of a set of possible hypotheses, based on the prior probability of the theories and their likelihood (i.e. the probability of the observed data given the hypothesis). The most contoversial part of this theory is the assignment of prior probabilities over the set of possible theories. This is because in most inductive situations we only have a poor idea of the necessary priors over theories, yet without them it is impossible to decide which theory should be prefered over another. Note that the set of possible hypotheses is not necessarily finite (as in the classification example) provided the sum of the probabilities over all the hypotheses is one.

The above equations involve a product of probabilities, so if we take $-log$ of these equations, we turn the products into sums (of positive quantities). Information theory interprets $(-\log P)$ to be the minimum message length required to encode the corresponding event—the unlikely events requiring a longer message. Consequently, the above induction procedure can be recast as a search for the hypothesis with the shortest message length (when minimally encoded). Thinking about an inductive problem as a problem of minimal encoding is a heuristic aid in choosing prior probabilities for the possible hypotheses, because a "natural" encoding scheme for communicating possible hypotheses implies a particular prior probability. This association arises because there is a one-to-one correspondence between a minimal message length (M) and a probability (P) via the formula $M = -\log P$. For example, the particular choice of prior probability implied by Eqn. 2 corresponds to a message length to encode the number of classes (N) given by $M = \log N$—a natural encoding proportional to the number of digits in the number N.

A fundamental difficulty with any method of induction is how to decide which general class of theories to use as a basis. In our robot example, the type of theory used was classification of particular objects based on similarity of their attribute descriptions. Many other types of theories are possible, particularly in dynamic situations. Humans seem predisposed (by evolution ?) to search for causal theories, even in such apparently static situations as our robot room clustering. Here, we understand the cause of the similarity among the different rooms because we understand their function. The robot is in the same situation we would be in if we noticed similarity among different rooms in an alien civilization, but had no idea of the cause. However, we, like the robot, believe there *is* an underlying cause for the observed similarity, even if we do not know what it is.

An important feature of the inductive procedure described in this paper is that it can be applied hierachically to the results of earlier inductions. For example, the results of class induction is a set of classes implicitly defined by the set of examples belonging to the class. These classes can now be regarded as examples at a higher level and so the robot can form classes of classes, and so on. However, the number of examples at any level drops exponentially, so there is a built in limit to this process. That is, due to the limited lowest level sample size, there is an upper bound on the levels of theory that can be built. In science, this heirachical theory building is clear. For example, many observations of the behavior of gases produced specific theories such as Boyle's law. Later, such theories were combined into a more comprehensive theory–the kinetic theory of gases. This theory was further extended with the arrival of quantum mechanics, and it is unlikely that the last word in the theory of gases has been said.

7 Discussion

The class induction procedure presented in this paper removes much of the *ad hoc* flavor of existing AI induction methods, but it also raises some very fundumental questions. The most important is why would anyone want to induce (untutored) classes at all! If the goal is to predict properties of the world, such as the probability of finding a terminal in a room, then using class properties to make the prediction is *not* the optimal method. The full Bayesian method is to compute the required probability conditioned on *all* the known information about rooms. That is, instead of trying to decide to what class of rooms the room in question belongs, then computing the required probability conditioned on that class; the correct procedure is to compute the required probability conditioned on each class, then sum the probabilities weighted by the probability of the particular room belonging to each class.

In practice, if one class is much more likely than all the others combined, then the answers obtained by the different methods will be virtually identical. However, if the particular room does not seem to belong to any particular class, then chosing the most likely and making prediction on that identification can be highly misleading. In view of this sub-optimality, it appears that computers (because of their superior computational power) should use a strategy of basing predictions on a single class identification if the probabilistic match is strong enough, but using the more expensive weighted sum when no clear identifcation can be made. This strategy requires a kind of "truth maintenance" system so that predictions based on particular identifications can be revised if new evidence casts doubt on the identification.

Another fundamental question is: "Are the classes discovered in untutored learning real?". Instead of a deep philosophical discussion on the nature of reality, the robot example illustrates the following relevent points. If the attribute values associated with the examples (e.g. rooms) are generated at random, then it is highly likely that the induction procedure will only find evidence for a single (all inclusive) class. That is, the robot is unable to find evidence for making class distinctions. It is always possible that a randomly generated set could produce examples that are highly clustered in probability space, leading to the induction of classes. This is because the class critierion, Eqn. 1, is inherently probabilistic and so there is no way in which an observed pattern can be "proved" absolutely to be "real" rather than a chance occurance. However, if the relative probabilities calculated using Eqn. 1 are strongly biased, then it is highly unlikely that the corresponding classes are due to chance.

Another major consideration is the limitations produced by restricting the set of theories to existance of "classes". If the robot only looks for classes then it will miss patterns such as the serial/spatial ordering of the room numbers, or spatial groupings of rooms into, say, sets of three. Humans use an extremely rich theory language, and typically switch to a new language (paradigm) when there is insufficient progress in the current one. An intelligent robot must also have a rich theoretical language, or at least the fundumental concepts to build more complex theories. Humans seem to be endowed by evolution to seek causal explanations (theories) of the world, and goal directed explanations seem to also be important. These seem to be the building blocks from which more elaborate theories are constructed. It is interesting to speculate what theories of the world a robot might construct, (including human), if it started from a fundamentally different theoretical framework. Would its predictions strongly differ from ours, or are the "real" patterns in the world strong enough to overcome any a priori theory framework biases.

The above method for class induction appears to suffer from a logical flaw following the arguments presented by Watanabe [16]. He points out that the individual attribute values, such as used by the robot, can be combined into more complex logical expressions that appear to have equal claim to be considered as do the original attributes. For example, the robot could form the compound (non-primative) attribute "in(x, terminal) OR in(x, typewriter)". This new "attribute" could be added to the list of attributes and included in any class definition. However, if all such possible logical compounds are added in this way, Watanabe has shown ("Theorem of the Ugly Duckling") that there is no logical basis for induction! If all possible compounds of attributes are treated the same as the primative attributes, the same difficulty applies to the probabilistic case as well. The method proposed by Watanabe to avoid this problem is to assign unequal weights to different attributes (including the compound attributes), but the user is expected to supply these weights. The robot however is intended to form classes without help, and so it seems as if it would be overwhelmed by the explosion of possible attributes.

The apparent difficulty noted above disappears when we realize that the primative attributes (i.e. those that are directly sensed) have a prefered status compared to possible compound attributes that can be formed from them. In Watanabe's terms, this means the primative attributes are given considerable weight. This is because a sensed attribute value is a real property in the world and not a possible property in some hyperthetical world, and so should have a prefered status. For example, if the robot can detect typewriters and terminals, it makes the default assumption that they occur independently, unless it has information to the contrary (see section 4.2). If it observes that the joint occurance of these objects in a room is *not* independent (e.g. one or the other (but not both) tends to occur) then it will revise its prior independence assumption. The independence assumption in this case, amounts to assuming equal probabilities for every *conjunction* of the given attribute values (e.g. $A_1 \wedge B_3$). Consequently, this gives a *non*uniform prior probability assignment to other logical combinations (e.g. $A_1 \vee B_3$). This nonuniform prior probability assignment does not prevent such logical combinations from being judged significant—it just makes them more unlikely. That is, it requires more evidence before a compound logical expression is accepted than a conjunction of the same primative properties. This nonuniform assignment is just a consequence of the basic independence assumption, which is defined in terms of conjunctions (e.g. $Pr(P \wedge Q) = Pr(P).Pr(Q)$).

In summary, induction of theories under uncertainty is a complex problem that raises most of the basic epistemological issues in philosophy. In AI, these issues are not just interesting theoretical issues, but arise from the need to build agents that can make reasonable decisions under uncertainty, within available computing resources. The basic theory for performing induction is illustrated using a particular robot example. This theory is shown to be a consequence of Bayes' theorem—no new principles are required. The main difficulty with this approach is the need to provide prior probabilities over the space of possible theories. In the absence of known priors, Occam's razor suggests we should prefer the simplest theory supported by the data. The reason for this choice is essentially a pragmatic one. Also, the use of classifications as a way of making predictions about unobserved attribute values is shown to be sub-optimal, particularly when the class identification of the object in question is not strong. These class indentifica

References

[1]Box, G. P. and Tiao, G. C., "Bayesian Inference in Statistical Analysis", Addison-Wesley, 1973.

[2]Cheeseman, P. C., "Probabilistic verses Fuzzy Reasoning", to appear "Uncertainty in AI", Nth. Holland, Eds. Kanal and Lemmer, 1986.

[3]Cheeseman, P. C., "In Defense of Probability", Proc. Ninth International Conference on Artificial Intelligence, Los Angeles, Aug. 1985, pp 1002-1009.

[4]Cheeseman, P. C., "A Method of Computing Generalized Bayesian Probability Values for Expert Systems, Proc. Eight International Conference on Artificial Intelligence, Karlruhe, Aug. 1983, pp 198-202.

[5]Cheeseman, P. C., "Learning Expert Systems from Data, Proc. Workshop on Principles of Knowledge-Based Systems, Denver, pp 115-122, Dec. 1984.

[6]Chow, C. K. and Liu, C. N., "Approximating Discrete Probability Distributions with Dependence Trees", IEEE Trans. on Information Theory, Vol. IT-14, No. 3, pp462-467, 1968.

[7]Cox, R. T., "Of Inference and Inquiry—An Essay in Inductive Logic, In The Maximum Entropy Formalism, Ed. Levine and Tribus, M.I.T. Press, 1979.

[8]Duda, R. O. and Hart, P. E., "Pattern Recognition and Scene Analysis", Wiley-Interscience, 1973.

[9]Gaines, B. R., "Behaviour/structure Transformations under Uncertainty", Int. J. Man/Machine Studies, 8, pp337-365, 1976.

[10]Georgeff, M., "A General Selection Criterion for Inductive Inference", Proc. 6th. European Conf. on AI, Pisa, Italy, Sept., 1984.

[11]Jaynes, E.T., "Where do we stand on Maximum Entropy, in "The Maximum Entropy Formalism, Levine and Tribus Eds. M.I.T Press 1979.

[12]Michalski, R. S., Carbonell, J. G. and Mitchell, T. M. (Eds), "Machine Learning: An Artificial Intelligence Approach", Tioga Press, Palo Alto, 1983.

[13]Pearl, J., and Kim, J. H., "A Computational Model for Causal and Diagnostic Reasoning in Inference Systems, Proc. 8th. International Conf. Artificial Intelligence, Karlsruhe, pp 190-193, Aug., 1983.

[14]Quinlan, J. R., "Learning from Noisy Data", Proc. International Machine Learning Workshop, June 1983, pp58-64.

[15]Wallace, C.S. and Boulton, D.M. "An Information Measure for Classification, Computer Journal, 11, 2, pp 185-194, 1968.

[16]Watanabe, S., "Pattern Recognition: Human and Mechanical", Wiley-Interscience, 1985.

[17]Wittgenstein, L., "Philosophical Investigations", Macmillan, 1958.

[18]Zadeh, L. A., "Possibility Theory and Soft Data Analysis, In Mathematical Frontiers of the Social and Policy Sciences, Ed. L. Cobb and R. M. Thrall, pp 69-129.