Using Prior Learning to Facilitate the Learning of New Causal Theories

Michael Pazzani, Michael Dyer, Margot Flowers
UCLA Artificial Intelligence Laboratory
3531 Boelter Hall
Los Angeles, CA 90024

Abstract

We present an approach to learning causal knowledge which lies in between two extremely different approaches to learning:

- **empirical methods** (e.g., [12, 17]) which detect similarities and differences between between examples to reveal regularities.
- **explanation-based methods** (e.g., [13, 4]) which derive a causal explanation for a single event from existing causal knowledge. The event and the causal explanation are generalized to create a new "chunk" of causal knowledge by retaining only those features of the event which were needed to produce the explanation.

In the approach to learning presented in this paper and implemented in a program called OCCAM, prior knowledge indicating what sort of distinctions have proven useful in the past influences the search for causal hypotheses. Our approach to learning shares a goal with explanation-based learning: to allow existing knowledge to facilitate future learning so that fewer examples are required. However, it does not share one shortcoming of explanation-based learning since it can create causal theories which are not implications of existing causal theories.

Introduction

We address the problem of learning causal knowledge by observing examples of actions and state changes. We wish to consider the acquisition of simple causal theories such as those which describe the outcome of common events in the life of a small child (e.g., when a cup made of glass is dropped, it usually breaks and when a cup made of plastic is dropped, it doesn't break).

We take an empirical approach to learning causal theories. A current best hypothesis [12] is formed by noticing similarities and differences among the attributes of an observed event and recalled previous events. We choose to select a current best hypothesis rather than maintain a set of consistent hypotheses (e.g., [12]) for a number of reasons:

- The set of consistent hypotheses can be very large. For example, consider the following situation: Karen (a young girl with blond hair and blue eyes wearing a green sweater) pulls on the refrigerator door but it doesn't open. Mike (an adult male with brown hair and brown eyes wearing a blue sweat shirt) pulls on the refrigerator door and it opens. There are six attributes with different values for Karen and Mike which can generate consistent hypotheses. (e.g., when a person with brown eyes pulls on the refrigerator door, it opens.) In addition, these attributes may be combined conjunctively or disjunctively to form a large set of consistent hypotheses. Psychological evidence (e.g., [1, 11]) indicates that only one or a small number of hypotheses are considered at one time. Thus generating a causal hypothesis is treated as searching the space of possible hypotheses.
- Before a sufficient number of examples have been encountered to rule out alternative consistent hypotheses, it may be necessary to predict the outcome of a new event. The current best hypothesis can serve as the source of this prediction.

When a new example falsifies the current best hypothesis, a new hypothesis is selected from the set of consistent hypotheses. In Winston's ARCH program [17] and in RULEMOD [2] domain-specific heuristics select the new hypothesis. However, since we are assuming no initial domain knowledge, our approach differs from the ARCH program and RULEMOD in the following ways:

- Initially, the current best hypothesis is selected randomly from the set of consistent hypotheses subject to the constraint that simpler hypotheses are selected first: one attribute discriminations are selected before conjunctive combinations and disjunctive combinations.
- Distinctions which have proven useful in the past influence the order in which causal hypotheses are generated. For example, after a number of examples, assume that the current hypothesis indicates that when adults pull on the refrigerator door, it opens. Later, when presented with examples of an adult with brown hair successfully inflating balloons, and a small child with blond hair unsuccessful at the same task, the age attribute would be preferred to the hair-color attribute. The hypothesis that when an adult blows into a balloon, it will inflate is considered before the hypothesis that when persons with brown hair blow into a balloon, it will inflate. As OCCAM learns about causality, domain-specific heuristics (e.g., adults are strong) are also learned which guide the search for the current best hypothesis on new problems.

In RULEMOD, all previous examples are remembered so that the set of consistent hypotheses is always consistent with previous examples. In the ARCH program, no previous examples are saved so that the set of hypotheses may contain hypotheses which are not consistent with previous examples. We take a compromise between these two extreme positions. In OCCAM, the exact number of previous events recalled from memory is dependent on the *retrievability* of each event as determined by the unique features of the events (see [8, 9, 14].) Typically, at least one positive example and at least one negative example are recalled when selecting a new hypothesis. In addition, the current example and the current incorrect hypothesis constrain the set of consistent hypotheses [1, 10].

Background: OCCAM

OCCAM [14, 15] is a program which maintains a memory of events in several domains. As new events are added to memory, generalizations are created which describe and explain similarities and differences between events. OCCAM integrates explanation-based and similarity-based learning techniques. For example, from a number of examples OCCAM induces a rule which indicates that parents have a goal of preserving the health of their children. This rule explains why a parent pays a ransom in a single example of kidnapping. This explanation is generalized by explanation-based learning techniques to create a kidnapping schema.

In this paper, we focus on using prior learning to facilitate the learning of new causal theories. Two aspects of OCCAM relevant to learning causal knowledge are explained in this section: generalization rules and confirming causal hypotheses.

Generalization Rules

In OCCAM, *generalization rules* postulate causal relationships. A generalization rule suggests a *causal explanation* for a *temporal relationship*. For example, the simplest generalization rule is "If an action on an object always precedes a state change for the object, postulate the action results in the state change". Generalization rules serve a number of purposes:

- Explanation-based learning in OCCAM is initiated when a generalization rule suggests an explanation which can be confirmed and elaborated by existing causal theories. In this case, generalization rules focus the search for an explanation.
- In the absence of existing causal theories, generalization rules suggest a causal explanation which can be confirmed or denied by additional examples. In this case, generalization rules serve to generate a set of plausible hypotheses which obey certain constraints on causality [3] (i.e., *covariation*, effects are always present when causes are present, *temporal order*, causes precede effects, and *mechanism*, a physical mediator which "connects" a cause to its effect. Generalization rules may be viewed as weak heuristics which filter the set of possible hypotheses to create a set of plausible hypotheses.

```
(def-gen-rule
  different-actor-gen-result-part          ;name
  ?state-1 = (state type ?stype            ;effect
                    object ?object)
  after
  ?act-1 = (act type ?atype                ;cause
                object (part of ?object))
  ((?act-1 result ?state-1))
  (:difference ?act-1 actor)
  :state-action-exception-actor            ;class
  )
```

Figure 1: A generalization rule (variables are preceded by "?"): If two similar actions performed on a part of an object have different results, and they are performed by different actors, the differing features of the actor are responsible for the different result.

A generalization rule is illustrated in Figure 1. Each generalization rule contains a pattern for an effect (In Figure 1 the effect is a state of an object), a pattern for the cause (an action performed on a part of the object), a temporal relation (after), a set of causal relationships (the action results in the state), an exceptions note (which indicates that the difference in actor may be responsible for the difference in results). This generalization rule would be responsible for generating the set of plausible hypotheses to account for the earlier example of the refrigerator opening after Mike pulls on the door, but not opening after Karen pulls on the door.

Generalization rules are divided into classes. Generalization rules which belong to the same class as the one in Figure 1 all attribute a difference in the result of an action to a difference in the actor of the action. They differ according to role the object of the state plays in the action. In the rule in Figure 1, the action is performed on a part of the object. In other rules in this class, the action is performed on the object, or the object is the destination of some action. The class of a generalization rule plays a part in facilitating the learning of new causal theories.

Confirming Causal Theories

In [15], we discuss our approach to confirming hypotheses. We quickly review two strategies for confirming hypotheses which will be used in a later example. First, confidence in a hypothesis is increased when it makes a correct prediction [9]. Second, confidence in a hypothesis is increased when each alternative hypothesis is ruled out. Later, we also indicate how prior learning can help confirm causal theories.

Facilitating the Learning of New Causal Theories

How can prior learning facilitate the selection of the current best hypothesis of a set of consistent hypotheses? One simple approach might be to keep track of the attributes which have entered into previous successful hypotheses. For example, consider a child eating pieces of a pineapple. The pieces can be different shapes (square or triangular) or different colors (yellow or white). Eventually, the hypothesis that the yellow pieces of pineapple taste better may be considered and supported by a number of examples. Should color be preferred in future hypotheses? Unfortunately, preferring color indiscriminately would hinder rather than facilitate learning in many situations. Consider the earlier example of the refrigerator opening after Mike pulls on the door, but not opening after Karen pulls on the door. If color were preferred, the best hypothesis might be that when a person with a blue shirt pulls on the door, the refrigerator will open. The problem with this simple approach is that the context in which a preference is made is ignored.

The approach that we take in OCCAM differs from the above simple approach in two ways:

1. Attributes which have entered into previous successful hypotheses are preferred in more restricted situations. These situations are determined by the type of the cause in a the generalization rule and the class of the generalization rule. For example, after inducing that the refrigerator door will open after an adult pulls on it, the preference for age applies only to the actor of this type of the action (propel, an application of a force) and to the same class of generalization rules (i.e., those which attribute a difference in a result to a difference in the actor).
2. **The attributes which have entered into previous successful hypotheses are used to create *dispositional* attributes. These dispositional attributes represent capacities or potentials.** For example, after OCCAM induces that the refrigerator door will open when an adult pulls on it, a dispositional attribute which might be called "strength" is created, where strength is the tendency for an *application of force by a particular actor to result in a state change.*

Dispositional attributes serve a number of purposes:

- Distributional attributes serve as intermediate conclusions [6]. Like Fu and Buchanan's intermediate concepts, these dispositional predicates often correspond to named concepts in our domain (see Figure 2). If further information is found out about a dispositional attribute, it applies to all future and past examples. For example, in OCCAM, age is initially associated with strength. If other attributes are found which are indicators of strength (e.g., size of arms), they enter into future predictions.

Role	Action	Attribute	Disposition
actor	propel	age	strength
actor	move	age	dexterity
actor	mbuild	species	intelligence
actor	mbuild	age	intelligence
object	propel	material	fragility

Figure 2: Dispositional attributes. "propel" is the conceptual dependency act for the application of a physical force, "move" is a movement of a body part, "mbuild" is the making of a decision. Note that the attribute listed for each disposition is only the initial attribute associated with the disposition.

- Distributional attributes can viewed as parent predicates [7]. It is in this manner, that distributional attributes facilitate learning new causal theories. When learning that a refrigerator will open after an adult pulls on the door, two hypotheses are created:

1. Adults are strong enough to open a refrigerator door.
2. Adults are strong.

It is this second more general hypothesis which facilitates learning in new domains. For example, this hypothesis can be specialized to indicate that adults are strong enough to inflate balloons. Note that OCCAM does not start with dispositional attributes such as "strength". Instead, dispositional attributes are created to account for differences in capabilities (for actors) or tendencies (for objects). These dispositional attributes serve as domain-specific knowledge which guide the search for causal hypotheses.

- More support is given to hypotheses which are formed by making use of existing dispositional attributes. It is in this manner that prior learning also facilitates confirming hypotheses.

There are a number of issues which arise when using dispositional attributes to facilitate the search for causal hypotheses:

- When are dispositional attributes created? Dispositional attributes are created to account for a difference in the result of two (or more) actions.
- How do we avoid creating a new dispositional attribute for each new example? The reuse of existing dispositional attributes is preferred to the creation of new ones.

An Example

In this section, we demonstrate how learning dispositional attributes facilitates learning new causal theories. The example we consider is the refrigerator opening after Mike pulls on the door, but not opening after Karen pulls on the door. In this case there are two events in memory. Events are input to OCCAM in conceptual dependency [16]. A simplified representation of Mike opening the refrigerator is illustrated in Figure 3.

The generalization rule in Figure 1 suggest that a difference in the actor *accounts for the different results when Mike or Karen pulls on the door.* Since there are not yet any applicable dispositional attributes, OCCAM randomly selects one attribute of the actor which is different in Karen and Mike: eye-color. OCCAM creates a new dispositional attribute[1] (disp-1)

[1]This tendency doesn't have a name in English, so we'll have to refer to it by OCCAM's name: disp-1.

```
(act type propel
     actor (human name (mike)
                  gender (male)
                  hair-color (brown)
                  eye-color (brown)
                  age (adult))
     object (part type (door)
                  of (frige color (tan)))
     after (state type (open)
                  value (yes)
                  object (frige color (tan))))
```

Figure 3: **Simplified Conceptual Dependency representation of Mike opening the refrigerator**

which represents the tendency for an application of a force by a person with brown eyes to result in a state change. The current best hypothesis is that persons with brown eyes are disp-1 enough to open a refrigerator.

Soon, OCCAM is presented with a counterexample of a small child with brown eyes and blond hair who cannot open the refrigerator. This contradicts a prediction made by the current hypothesis. Since very little confidence had been built up for the current hypothesis and disp-1, they are abandoned, and a new current best hypothesis must be generated. There are at least two possible hypotheses: persons with brown hair can open refrigerators, or adults can open refrigerators. OCCAM randomly selects adults to form a new dispositional attribute[2] (strength) which represents the tendency for an application of a force by an adult to result in a state change. The current best hypothesis is that adults are strong enough to open a refrigerator. Further examples give a great deal of support to this hypothesis and to the dispositional attribute called strength.[3]

Once OCCAM has learned a dispositional attribute, future learning is facilitated. OCCAM is next presented with an example of Mike successfully inflating a round yellow balloon, while Karen cannot inflate a long blue balloon. This time, two generalization rules apply, one which would attribute the difference in the result to a difference in the object (round yellow balloon vs. long blue balloon) and one which would attribute the difference to the actor (Mike vs. Karen). Since there are no dispositional attributes for the object difference, one attribute (color) is randomly selected. For the actor difference, the strength dispositional attribute applies (since the generalization rule class, and the act type are the same as the refrigerator example), and the age attribute is selected over other attributes such as hair color. These two competing hypothesis are compared, and since the strength (and, therefore, the age) of the actor has more support than the color of the object,[4] it is favored. The current best hypothesis is that adults are strong enough to inflate balloons. Further examples add support to this hypothesis.

Conclusions

We have presented one point on what appears to be a continuum between explanation-based learning and empirical learning methods. The technique presented in this paper appears limited to domains that have a reason for their regularity (i.e., dispositional attributes). For example, it would not apply to concept learning of physical objects (i.e., there is no reason that color is significant in distinguishing a horse from a zebra, but not significant in distinguishing an arch from a house).

There are a number of possible extensions to this work. First, OCCAM contains generalization rules which attribute the difference of a result to a difference in the actor or the object. These rules could also be learned as dispositional attributes representing internal and external causes for success or failure. This distinction is quite important in determining the affective response to an outcome [5]. Second, the dispositional attributes in OCCAM are learned at a fixed level of generality (i.e., the type of conceptual dependency action.) This works well for the examples we have encountered, but a more general approach would be to learn the level of generality of a dispositional attribute by keeping track of the instances in which it has applied successfully and unsuccessfully. This would be more in the spirit of Goodman's parent predicates and over-hypotheses. Finally, the cases of a dispositional attribute being abandoned can be recorded so that learning is also facilitated by avoiding the same mistake in future cases.

We have presented an approach to learning causal theories which creates dispositional attributes such as "strength" to facilitate future learning. This technique has been applied successfully to a number of examples of causal theories and an example of a social theory: OCCAM required many examples to induce that parents (as opposed to any adult) have a goal of satisfying the hunger of their children. A dispositional attribute (which might be called "affection") was formed which facilitated learning that parents have a goal of preserving the health of their children. This social knowledge provided an explanation for a parent paying the ransom in kidnapping, which enabled OCCAM to create a kidnapping schema by explanation-based learning techniques.

[2] OCCAM's name for this attribute is disp-2.

[3] Note that an event such as an adult not being able to lift a car would not decrease support for the existence of the strength dispositional attribute. Only a counterexample such as a child lifting a car which an adult could not would remove support from this hypothesis since this counterexample would use the same generalization rule.

[4] Recall that the re-use of a dispositional attributes increases the support for a hypothesis

References

[1] Bower, Gordon and Trabasso, Tom. *Attention in Learning: Theory and Research.* John Wiley and Sons, New York, 1968.

[2] Buchanon, B. Mitchell, T. Model-directed learning of production rules. In Waterman, D. & Hayes-Roth, F. (editor), *Pattern-directed Inference Systems.* Academic Press, New York, 1978.

[3] Bullock, M. *Aspects of the young child's theory of causality.* PhD thesis, University of Pennsylvania, 1979.

[4] DeJong, Gerald and Mooney, Raymond. Explanation-based learning: An alternate view. *Machine Learning* 1(2), 1986.

[5] Dyer, M. *In Depth Understanding.* MIT Press, 1983.

[6] Fu, L. and Buchanon, B. Learning Intermediate Concepts in Constructing a Hierarchical knowledge base. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence.* Morgan-Kaufmann, Los Angeles, CA, 1985.

[7] Goodman, Nelson. *Fact, Fiction and Forecast, fourth edition.* Harvard University Press, Cambridge, Mass, 1983.

[8] Kolodner, J. *Retrieval and organizational strategies in conceptual memory: A computer model.* Lawrence Erlbaum Associates, Hillsdale, NJ., 1984.

[9] Lebowitz, M. *Generalization and memory in an integrated understanding system.* Computer Science Research Report 186, Yale University, 1980.

[10] Levine, M. Hypothesis behavior by humans during discrimination learning. *Journal of Experimental Psychology* 71:331-338, 1966.

[11] Levine, M. The size of the hypothesis set during discrimination learning. *Psychology Review* 74:428-430, 1967.

[12] Mitchell, T. Generalization as search. *Artificial Intelligence* 18(2), 1982.

[13] Mitchell, T., Kedar-Cabelli, S. & Keller, R. Explanation-based learning: A unifying view. *Machine Learning* 1(1), 1986.

[14] Pazzani, M. Explanation and generalization-based memory. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum, Irvine, CA, 1985.

[15] Pazzani, M., Dyer, M. & Flowers, M. The role of prior causal theories in generalization. In *Proceedings of the National Conference on Artificial Intelligence.* American Association for Artificial Intelligence, Morgan-Kaufmann, 1986.

[16] Schank, R.C. & Abelson, R.P. *Scripts, plans, goals, and understanding.* Lawrence Erlbaum Associates, Hillsdale, NJ., 1977.

[17] Winston, P.H. Learning Structural Descriptions from Examples. In Winston, P.H. (editor), *The Psychology of Computer Vision.* McGraw-Hill, New York, NY, 1975.