89-3

Inductive Refinement of Casual Theories

Claudio Carpineto

MLI 89-2

Inductive Refinement of Causal Theories

Claudio Carpineto
George Mason University *

Abstract

This paper collects ideas from causal analysis, analogical reasoning, empirical learning, and presents an integrated methodology to refining causal and social theories. Its major contribution is that the theory validation method, which consists of an incremental process of generation and pruning of examples and counter-examples, can work in the face of the two following limitations : (a) lack of prior knowledge, (b) lack of specific examples. The approach has been implemented in a program called IR89. We provide an example of IR89 acting as an experimenter of social theories in the domain of Italian Renaissance history. Experimental investigations with "interesting" theories give the work more support.

1. Introduction

The problem of inducing causal and social theories from orderly accounts of events has received much attention in the machine learning field recently (Pazzani et al., 1986; Danyluk, 1987). However, little work has been done on the complementary problem of validating causal and social theories, perhaps acquired by observation.

One notable exception, though restricted to the discovery of laws about the physical world, is the experimentation-based theory revision approach. In an early example of this approach (Langley, 1981) experimentation is used to explain inconsistencies between the system's theory and the real world. More recently (Rajamoney-DeJong, 1987) it has been argued that experiment design may be the key strategy to prune the incorrect explanations produced by an incomplete model.

In this paper we propose a domain-independent methodology based on active observation to refining causal theories. The observation involves identification of examples and counter-examples from a totally ordered set of events, the past examples being used to evaluate the correctness of a theory and suggest refinements. Unlike previous work, we analyze and refine hypothetical causal dependencies in the light of given temporal dependencies. As an interesting byproduct, the comparison between sequences of events of different length reveals subtle effects of ***seeming achievement***, like those described in Machiavelli's The Prince (Machiavelli, 1961) and illustrated here in the example.

In addition to performing theory validation outside of conventional scientific theory discovery, this approach presents two salient features.

The first is that it integrates several existing techniques in causal analysis (Winston, 1986), analogical reasoning (Gentner, 1983), and learning from examples and counter-examples (Dietterich-Michalski, 1983).

* On leave from Fondazione "Ugo Bordoni", via B. Castiglione 59, 00142 Roma, Italy

The second is that it allows for the contemporary absence of the two standard sources of knowledge in learning, namely (a) prior causal theories and (b) specific examples explicitly supplied by the user.

The approach has been implemented in a system called IR89. The rest of the paper is organized as follows. First we will describe the program IR89. Secondly, we give a simplified example of IR89 at work in an historical domain. Thirdly, we compare this work with two approaches that perform a structurally similar learning task by using a different learning method. Finally we provide an empirical evaluation with "interesting" theories.

2. System overview

IR89 confirms, rejects or completes general theories about causal and social processes. IR89 is provided with a memory of temporally ordered events. The theory to be investigated is a causal relationship between the features that describe such events.

After being given the input theory, IR89 employs its operational definition of causality to generate from the event memory a set of positive examples and a set of negative examples for that theory. Then IR89 attempts to reduce the set of negative examples by varying the temporal parameter involved in the definition of causality. At this point, IR89 recursively refines the theory over the two sets, until the set of negative examples reduces to the empty set. The refinement is heuristically driven: analogical heuristics figure out possible refinements and statistical heuristics evaluate their impact on the two example sets. Interestingly, as will become clear later, there is no need to use the event memory to (re)generate the new sets of positive and negative examples at each refinement step .

In the next three sections we illustrate the memory and theory representation in IR89, and the two main mechanisms involved in the algorithm, namely 1) generation of examples and counter-examples and 2) theory refinement.

2.1 Memory and theory representation in IR89

Each event stored in the system is labelled with a progressive ordering number. The events are represented as hierarchical frame descriptions, with two major syntactic types : <name> and <description>. Intuitively, <name> is for naming descriptive features, <description> is for describing them, perhaps in terms of other features. More formally, any event is a pair (<name> <description>), where <name> is an atomic symbol and <description> is either an atomic symbol, or another pair (<name> <description>), or a conjunction of such pairs. This representation reduces to classic attribute-value representation when <description> is atomic.

We assume that each event is self-contained; that is, there is no need to infer missing features from the event memory. Consequently, the system trades clarity of description and better efficiency for redundancy in representation.

It is also worth noting that while equal <names> appearing in different events may have different <descriptions>, we do not adopt an action-state representation. In effect, the only predefined causal assumption is that the occurence of an event is

the trivial cause of any change described in the event. The motivation for this guarded initial assumption is that we consider domains (e.g. history, economics) where 1) there is a great number of potentially active factors, and 2) their interrelations are often unclear.

Any causal theory has the form of implication, of the type: IF A THEN B. Both its antecedent and its consequent have the same syntax as the event descriptions, except that their atomic <descriptions> can be variabilized and treated with selected predicates.

The two parts of a given theory are supposed to partially match some pair of events. The idea is that a partial theory can be seen as a hierarchical relational structure with variable-place predicates and missing arguments. In this way, the task of refining the theory just requires appropriate filling gap.

2.2 Generating examples and counter-examples

We use a simple operational definition of causality:

A causes B if every occurrence of A is strictly followed by an occurrence of B.

This reduces a causal relation to a material implication, whereas a more realistic definition of causality would probably have to be expressed as a combination of necessary and sufficient conditions (Sternberg, 1985). In effect, the normal use of causation seems to be both stronger and weaker than material implication (Shoham, 1988), and therefore a more thorough evaluation would require for instance an analysis of the mutual covariations of the causal antecedent and the causal consequent. However, while these issues may be relevant for inducing causal theories from scratch, they are much less important for a quantitative refinement of given causal theories, as in IR89.

According to our generative account of causality, the memory is scanned, and:

1) each pair of consecutive events, such that the first matches A and the second matches B*, is considered to be a positive example for the theory.

2) each pair of consecutive events, such that the first matches A and the second does *not* match B, is regarded as a negative example.

In practice, as IR89 only diagnoses more restrictive condition for theory applicability, it will only pay attention to the antecedents of the two sets.

Before illustrating the supplementary treatment of the negative examples, it is worth explaining the motivations for introducing a specific component that generates an initial example set. This need is often neglected in machine learning systems. We feel instead that the capability of identifying training examples as a strategic learning component takes the construction of an integrated performance system one step further. One more practical consideration is that, while it may be relatively easy to have a collection of general data, it may be very difficult to obtain specific examples; our observations are usually more complicated and

* Matching requires simple unification of the theory antecedent, and instantiation of the theory consequent's variables. Since the subframes occurring in a theory are intended to be existentially quantified, any pair of events may match the theory in multiple ways.

contain more information than necessary to test or induce one specific theory. Because the relevance of any obervational feature is a function of the causal process being investigated, the approach taken in IR89 is to project any input theory onto the whole set of observations, trying to enlight the portions involved.

Recomputing the example set with a wider ordering step

The generation of the two example sets is intended to identify a causal search space. However, our definition of positive and negative examples is based only on strict temporal contiguity (statistical frequency and similarity will be used later to order the search space). This requirement turns out to be very severe in identifying positive examples. In particular, it prevents from correctly classifying those cases in which incidental observations temporally separate a causal antecedent from its causal consequent. In order to reduce the number of misclassified positive examples, IR89 tests each negative example with a wider ordering step (= 2). This corresponds to allow for *degrees of consistency* of each theory, depending on different partitions of the observation set. In fact, we started with an ordered list of observations (obs_1, ..., obs_n), and then defined for each theory a procedure (f) mapping ordered pairs (obs_t, obs_{t+1}) into the three-valued set {pos-example, neg-example, no-example}. Now we are saying that there are no counter-examples (i.e. the theory is consistent with the data) not only if

$\neg\exists t\ f(obs_t, obs_{t+1}) = \text{neg-example}$,

but also if

$\forall t\ (f(obs_t, obs_{t+1}) = \text{neg-example}) \Rightarrow (f(obs_t, obs_{t+2}) = \text{pos-example})$.

As this definition may produce implausible causal relationships, we want to check if there are features being omitted that can be deemed as *direct* causes of the theory consequent. This is done semi-automatically: the user may supply a causal antecedents that matches the skipped event (obs_{t+1}) and is consistent with the causal consequent, and the program reruns, with a lower temporal step (=1) focusing the search on strict temporal contiguity. This technique for ruling out implausible causal inferences is similar, in spirit, to two other approaches to constraining a procedure for IF-THEN rules acquisition, namely Winston's censors technique (Winston, 1986) and Blum's search for latent variables (Blum, 1982). Here, however, the focus is on the effects produced by a temporal dilution of the causal consequent, whereas Winston and Blum identify as a possible reason of rule misapplication an incomplete description of the causal antecedent.

2.3 Theory refinement

The strategy is to shrink the negative set while minimizing the reduction of the positive set. The tactic to cut off the search space is to focus attention only on a limited number of hypotheses at one time. We seek refinement features connected to the lowest level of the hierarchical structure of the theory antecedent first, and then move on to features connected to the upper levels, interleaving analogical heuristics for candidate generation and statistical heuristics for candidate evaluation.

The inputs of the **refine** routine are: the theory-antecedent (we shall say theory for short), the set of positive examples, the set of negative examples, the innermost level of the input theory description.

It works as follows :

1. A set of candidate refinements is formed just by inspecting the positive set. It contains all the "nodes" of the positive examples which can be added as "brothers" to the current hierarchical level of the theory.

2. The candidate features are evaluated on the basis of how they modify the negative and positive sets. Over-general and over-specific features are discarded. Features that affect both sets in the same way are kept away for later use*. The best refinement among the remaining candidates, if any, is added to the theory.

3. Two new example sets are associated with the refined theory. They contain all of and only the previous examples that (still) match the refined theory. Step 3 is interleaved with step 2.

4. The **refine** routine is recursively applied to the new theory, to the two new sets of positive and negative examples, to the next hierarchical level.

5. The algorithm ends successfully when the negative set becomes empty. In this event, the refined theory is guaranteed to hold in at least two examples and to have no counter-examples. It halts unsuccessfully when either there are no more discriminating features, or the positive set becomes too small (=1), or at the top of the refinement hierarchy.

We use powerful analogical heuristics to reduce the refinement space. While their effectiveness will be empirically discussed later, the general underlying idea is that a feature that belongs to a shared system of encapsulated features is more likely to provide the required theory specialization than isolated common features. This is an adaptation from Gentner's systematicity principle (Gentner, 1983), which states that people prefer structural similarity over literal similarity when generalizing models. It is worth noting, however, that we use this simple technique not to find the best analogies between two fixed relational structures, as Gentner does, but as a means to inductively fill in arguments in a semi-specified structure of variable-space predicates (i.e., the input theory).

Apart from the use of analogical heuristics, there is another major difference between this refinement routine and most learning from examples and counter-examples algorithms (Dietterich-Michalski, 1983). The difference is that the learned concept (i.e., the refined theory) is required to be complete *only* with respect to a subset of the original set of positive examples. Therefore this algorithm is rather a variant, for coping with the cases in which the set of positive examples can be reconfigured, depending on the current concept formulation.

The advantage, in presence of this active feedback between theory reformulation and example identification, is that, at each step, the system can generate itself a customized set of examples from the given sets of events to find a solution.

*The reuse of unrateable features in later refinement steps may greatly reduce the effects of heuristics interaction. In this case the pairing of features allocated in different branches of the theory structure may uncover non-linear causal effects.

3. An example

We have modeled from a history textbook about thirty major events occurring in Italy during the XV and XVI century. Consider as an example what happened at the beginning of the XVI century. The French asked for Spanish alliance to conquer the Kingdom of Naples. The French and the Spanish conquered the Kingdom of Naples and divided it among themselves. Afterwards the Spanish made war on the French and captured the French part. The memory representation of this sequence of events is shown in figure 1.

```
EVENT10
ALLIANCE
  DATE: 1500
  PLACE: GRANATA
  OBJECTIVE
    POWER
      NAME: KINGD-OF-NAP.
      GOVERN: KINGDOM
      LANGUAGE: SPANISH
      MIL-POWER: WEAK
      NUM-OF-TERRIT.: 0
  ALLIES
    POWER
      NAME: FRANCE
      GOVERN: KINGDOM
      LANGUAGE: FRENCH
      MIL-POWER: WEAK
      NUM-OF-TERRIT.: 1
    POWER
      NAME:SPAIN
      GOVERN: KINGDOM
      LANGUAGE: SPANISH
      MIL-POWER: STRONG
      NUM-OF-TERRIT.: 2
    NUM-OF-ALLIES: 2

EVENT11
WAR
  BEGIN-DATE: 1501
  DURATION: 1Y
  BATTLES
    NUM-OF-BATTLES: 1
    BATTLE
      DATE: 1501
      PLACE: NAPLES
  WINNING-SIDE
    POWER
      NAME: FRANCE
      GOVERN: KINGDOM
      LANGUAGE: FRENCH
      MIL-POWER: WEAK
      NUM-OF-TERRIT.: 2
    POWER
      NAME: SPAIN
      GOVERN: KINGDOM
      LANGUAGE: SPANISH
      MIL-POWER: STRONG
      NUM-OF-TERRIT.: 3
  LOSING-SIDE
    POWER
      NAME: KINGD-OF-NAP.
      GOVERN: DESTROYED
      LANGUAGE: SPANISH
      MIL-POWER: DESTROYED
      NUM-OF-TERRIT.: 0

EVENT12
WAR
  BEGIN-DATE: 1502
  DURATION: 2Y
  BATTLES
    NUM-OF-BATTLES: 2
    BATTLE
      DATE: 1503
      PLACE: BARLETTA
    BATTLE
      DATE: 1503
      PLACE: NAPLES
  WINNING-SIDE
    POWER
      NAME: SPAIN
      GOVERN: KINGDOM
      LANGUAGE: SPANISH
      MIL-POWER: STRONG
      NUM-OF-TERRIT.: 4
  LOSING-SIDE
    POWER
      NAME: FRANCE
      GOVERN: KINGDOM
      LANGUAGE: FRENCH
      MIL-POWER:WEAK
      NUM-OF-TERRIT.: 1
```

Figure 1: A three events' sequence from IR89's memory

Suppose to question the sytem about the following theory : *it is not worth making alliance in order to increase one's territories* *. The theory representation is

Does (ALLIANCE(ALLIES(POWER(NAME = X)(NUM-OF-TERRITORIES = N))))
cause (POWER(NAME = X)(NUM-OF-TERRITORIES $\leq$ N)) **?**

*IR89 is incapable of detecting successful refinements if tested with the more intuitive question: *is it worth making alliance in order to increase one's territories*?

Upon being presented with this theory, IR89 searches the set of events and finds positive examples (in which making alliance is not worthwhile), yet not illustrated in this paper, and negative examples (in which making alliance is worthwhile). The sequence in figure 1, for instance, casts two negative examples, one for each power involved. At this point IR89 recomputes the example set. It turns out that some negative examples, included the example of France making alliance with Spain, become positive examples when increasing the ordering step*. Then IR89 applies the Refine routine to the two new example sets. The refinement added to the theory are shown in fig. 2 in bold.

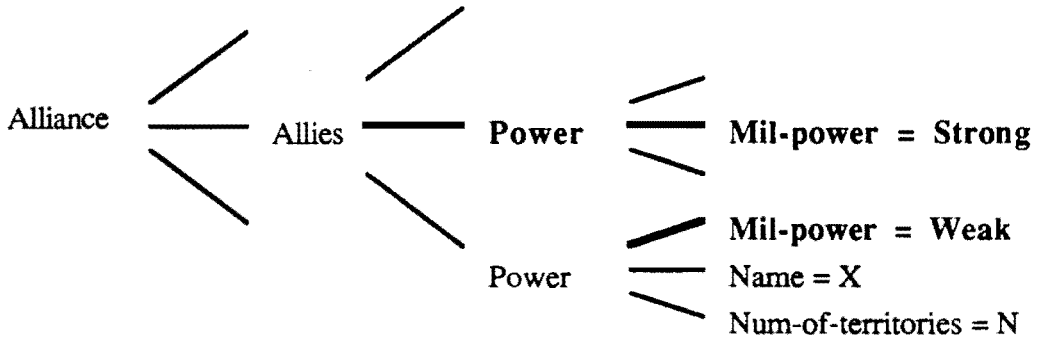

Fig. 2 : Refinement of the causal antecedent

The final theory is very much like one of the principles Machiavelli exposed in his book : "A power ought not to ask a more potent power for alliance". Note that the detection of the subtle causal effect involved in the example of France making alliance with Spain is made possible by the recomputing of the negative set with a wider temporal horizon.

4. Discussion

Given a theory to be refined and a set of events, the method we have discussed is purely empirical. It exploits the implicit analysis embodied in the description of the events, though, and crucially depends on the input theory. In particular, any theory to be refined must contain at least the top hierarchical level (along with one of its sons) of the corresponding complete theory. In order to make the implications of our implicit assumptions more explicit, we can compare IR89 to two distinct but related approaches (Danyluk, 1987; Carpineto, 1988). The learning tasks of the three systems are, in fact, structurally similar :

* Another sequence example is the following: the Venetians allied with the French to conquer Milan. Once they had conquered Milan, the French declared war on the Venetians and threw them out of Milan.

1) These programs are able to acquire *new* knowledge from *complex* descriptions of the same class.
2) The descriptions are represented as hierarchical frames, or as simple sequences of hierarchical frames.
3) The learning involves matching (and mapping) of high-level, partially specified hierarchical frames *of the same type* (e.g., two allied countries, two narrative functions, two terrorist attack locations).

However, they differ in the learning method. Both Danyluk and Carpineto use a combination of explanation-based learning and similarity-based learning; in their approaches, the use of an initial domain theory allows a complete explanation structure to be constructed and mapped onto each description, and this is essential in order to guarantee the feasibility of their subsequent inductive phase. In contrast, IR89 starts out with an incomplete explanation structure (i.e., any of its input theories), and then, in order to select the relevant part of the set of descriptions, integrates it with empirical techniques (in particular, by generating a set of positive and negative examples out of the set of descriptions, and by applying analogical heuristics to the example set). The comparison shows that a combination of empirical techniques can complement certain types of partially specified explanation structures - namely the ones in which all of the <names> are also present in the examples - so as to compensate for the lack of completely specified explanation structures.

Another important difference is in the use of the temporal information associated with the *real* examples. Danyluk does not use time at all; "cause" and "result" are predefined slots of each of her event descriptions, like "location". IVAN (Carpineto, 1988) takes advantage of the temporal ordering of the events told in its story examples, but only for confirming hypothetical causal generalizations suggested by its domain theory. IR89 makes an intensive exploitation of the temporal evidence, by using the ordering of its set of events as a basis for generating the training examples relevant to a given theory. Related to the use of different input theories and to this capability, there are two additional minor feature of IR89 :

- the learning task is not restricted to the acquisition of a single concept,
- the same event can be reused for generating different examples.

5. Experimental evaluation

Because its search is heuristically-driven, IR89 may not find a solution. Also, due to the lack of specific criteria of relevance, it may find a solution of little interest. The effectiveness of the various heuristics used to prune the space of possible refinements can be demonstrated by showing that not only is IR89 able to refine theories, but that it can generate interesting refinements.

For this purpose we took from Machiavelli's The Prince a dozen of theories covered by IR89' event memory and input them to the system in an incomplete form (specifying at least one node for each level of the corresponding complete form). IR89, except for one class of theories, was able to refine most of them. The theories it could not refine typically involved adding refinements with variabilized atomic

descriptions. One example of such class is the following: "In order to keep the possession of a territory, the conquerer power ought to have *the same* language and establish colonies in it (rather than garrison)".

Then we turned our attention to the behavior of the single learning components of IR89. We characterized the *space of interestingness* by means of three dimensions - (a) structuring degree of the causal antecedent, (b) temporal dilution of the causal consequent, (c) reuse of unratable features in the refinement of the causal antecedent - and defined three simple algorithmic measures for these dependent variables. In sum, we found out that :

- The structuring degree of the theories was lower than expected; in fact, the number of unexploited nodes generated at the lower levels of the theory structures was relatively high. As there is a body of psychological evidence (Holland et al., 1986) that in the use of default hierarchies rules based on more specific levels tend to dominate rules based on general levels, one possible explanation is that finding an *interesting* refinement requires shifting up the information specifity level.
- The theories behaved diffomly with respect to the amount of recomputation needed to transform their negative set. This component appeared however to be necessary for any theory involving detection and comparison of macro-sequences of analogical events.
- Many theories, including the example shown above, required simultaneous refinements over different branches of their structure. This effect was probably accentuated by a large presence of analogous features of different powers in the chosen sample.

6. Conclusions

We have presented IR89, a system that refines causal and social theories through active observation of a totally ordered set of events. The problem of learning theory refinements has been cast as a heuristic search through a reconfigurable space of examples and counter-examples. We discussed the advantages of this approach and evaluated how good it is at generating interesting refinements in an historical domain.

The main result of this research is that the combination of several empirical techniques - causal analysis to generate a search space, analogical and statistical heuristics respectively to order and prune it - makes theory refinement without prior knowledge and without specific examples feasible.

Some major weaknesses of this approach, suggesting future research directions, are indicated below.

- IR89 allows for (limited) differences between the causal ordering of the events and their temporal ordering in the memory. Also, it can do pluricausal inferences. However, IR89 cannot deal with *diachronic* causes, such as a temporally disjoined causal antecedent; it can only discover multiple *sinchronic* causes.
- The widening of the temporal horizon may affect the positive set in much the same way as does the negative set. We have ignored this problem because of the difficulty of providing an autonomous definition for negative examples recognition.
- IR89 is unable to induce causal theories by itself; indeed, the refined theories

can only grow around the incomplete structure provided by the user. However, IR89 might be used as a part of a larger system for theory development. Another program, such as OCCAM (Pazzani et al.,1986), might be used to induce *qualitative* rules and then IR89 might *quantitatively* refine these rules*.

Acknowledgements

Most of this research was done at George Mason University. I would like to thank Ryszard Michalski for his support. I would also like to thank Igor Mozetic and Jan Zytkow for many useful comments and criticisms on earlier drafts of this paper. The work was carried out in the framework of the agreement between the Italian PT Administration and the Fondazione Ugo Bordoni.

* True, [Pazzani, 1988] has proposed a mechanism for refining incomplete theories. However this mechanism is not empirical, for it is based on *abstract explanations*, and abstract explanations are theories themselves. This leads us straight back to the problem of refining an incomplete theory (i.e. an abstract explanation).

References

Blum, R.L. (1982). Discovery and Representation of Causal Relationships from a Large Time-Oriented Clinical Database: The RX Project. *Lecture Notes in Medical Informatics*, 19.

Carpineto, C. (1988). An Approach based on Integrated Learning to Generating Stories from Stories. In *Proceedings of the fifth International Conference on Machine Learning* (pp. 298-304). Ann Arbor, Michigan: Morgan Kaufmann.

Danyluk, A.P. (1987). The use of explanations for similarity-based learning. In *Proceedings of the tenth IJCAI* (pp. 274-276). Milan, Italy: Morgan Kaufmann.

Dietterich, T.G., Michalski, R.S. (1983). A Comparative Review of Selected Methods for Learning from Examples. In R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*. Los Altos, California: Tioga.

Gentner, D. (1983). Structure-mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7, pp. 155-170.

Holland, J.H., Holyoak, K.J., Nisbett, R.E., Thagard, P.R. (1986). *Induction Processes of Inference, Learning, and Discovery*. Cambridge, Massachusetts: MIT Press.

Langley, P. (1981). Data-driven Discovery of Physical Laws. *Cognitive Science*, 5, pp. 31-54.

Machiavelli, N. (1961) *The Prince*. Harmondsworth, England: Penguin Books.

Pazzani, M., Dyer, M., Flowers, M. (1986). The Role of Prior Causal Theories in Generalization. In *Proceedings of the fifth AAAI* (pp. 545-550). Philadelphia, Pennsylvania: Morgan Kaufmann.

Pazzani, M. (1988). Integrated Learning with Incorrect and Incomplete Theories. In *Proceedings of the fifth International Conference on Machine Learning* (pp. 291-297). Ann Arbor, Michigan: Morgan Kaufmann.

Rajamoney, S., DeJong, G. (1987). The Classification, Detection, and Handling of Imperfect Theory Problems.In *Proceedings of the tenth IJCAI* (pp. 205-207). Milan, Italy: Morgan Kaufmann.

Shoham, Y. (1988). *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence*. Cambridge, Massachusetts: MIT Press.

Sternberg, R.J. (1985). *Beyond I.Q. A triarchic theory of human intelligence*. Cambridge, Massachusetts: Cambridge University Press.

Winston, P.H. (1986). Learning by Augmenting Rules and Accumulating Censors. In R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), *Machine Learning:An Artificial Intelligence Approach, Vol.2*. Los Altos, California: Morgan Kaufmann.