Similarity, Conservatism, and Rationality

Jon Doyle

Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

February 26, 1988

Abstract: We examine some formalizations of the notion of similarity as it appears in learning, and of the notion of conservatism as it appears in reasoning. We show that the underlying formalizations are closely related to each other and to a central notion of the formal theory of rationality. These connections indicate how the structures of similarity judgments and conservatism can arise naturally in agents which rationally govern their own representation and reasoning.

This research was supported by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 4976, Amendment 19, monitored by the Air Force Avionics Laboratory under Contract F33615-87-C-1499. The views and conclusions contained in this document are those of the author, and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Government of the United States of America.

1 Introduction

The subjects of learning and reasoning are often fairly distinct in artificial intelligence, with learning focusing on acquiring new knowledge and reasoning focusing on using old knowledge. Of course, learning may involve reasoning (consider explanation-based generalization [Mitchell et al. 1986]), and reasoning may involve learning (consider SOAR's chunking [Laird et al. 1987]). But by and large, the different aims of learning and reasoning in the eyes of most researchers have led to separate theories.

In this paper we attempt to connect the concept of similarity prominent in many discussions of learning with the concept of conservatism prominent in some theories of reasoning. The best examples within artificial intelligence of the notion of similarity concern learning by analogy and clustering of data. In making analogies, one considers two situations and identifies as similar some of the objects and relationships present in one situation with objects and relationships present in the other. There are always many such identifications possible, and it is common to compare analogies, saying that one analogy is better than another. For example, Winston [1975] described a distance measure between analogies, with better analogies involving smaller distances, and Carbonell [1983] employed such a measure to guide heuristic search for analogies. In clustering data, one seeks to classify the differences between items as either significant or insignificant, in order to group together all similar (insignificantly differing) items as instances of the same event or concept. For example, [Michalski and Stepp 1983] cluster data by using neighborhood-dependent distance functions on points. On the other hand, in conservative reasoning, by which we mean conserving as much mental state as possible over time, not political opinions or aversion to risk, each mental state of the reasoner is similar to its predecessors, or more similar than alternative states. Similarity and conservatism have previously been intimately connected in the derivational analogy method of learning proposed by [Carbonell 1986], in which the arguments for a new concept are chosen to be as similar as possible to the arguments for an earlier concept. Here we view similarity and conservatism formally, and indicate how they arise naturally from the more fundamental notion of rationality.

2 Similarity

Similarity has been studied in its own right more thoroughly in psychology than in artificial intelligence. Tversky [1977], for example, summarizes a large literature and introduces a formal theory of similarity. His idea is that judgments of similarity are based on the differences in the sets of features of the items under consideration. Thus if a and b are objects characterized respectively by feature-sets A and B, the degree of similarity of a and b, written $s(a, b)$, is a function of the sets of shared and unshared features of a and b, namely $A \cap B$, $A - B$, and $B - A$. With such a function, one can compare degrees of similarity, so that $s(a, b) \geq s(c, d)$ means that a is more similar to b than c is to d, a comparison very familiar from Evan's and Winston's treatments of analogies. This view of similarity recognizes the asymmetry of some judgments, and also how common elements can sometimes outweigh differences, or vice versa. However, Tversky's axioms place strong requirements on similarity judgments, requirements sufficient to guarantee the existence of essentially unique continuous real-valued functions s for representing degrees of similarity. In the following, we focus on a somewhat more general conception, concentrating on the central notion of comparison of similarities, whose qualitative theory is simpler than Tversky's strong numerical theory.

Studies of similarity also appear in philosophy in the analysis of counterfactual statements such as "if it had not rained, the crops would have failed." For example, Lewis [1973] employs the notion of similarity to say that a counterfactual is true if its conclusion is true in those possible worlds most similar to the actual world, thus viewing counterfactuals as making analogies between the actual world and "possible worlds" in which the hypothesis is true instead of false. Of course, if the counterfactual is not an ordinary entailment, there are always possible worlds in which the hypothesis is true but the conclusion is false: hence the requirement that the alternative world be as similar as possible to the actual one, in order to make the interpretation nontrivial.

Lewis formalized the notion of similarity in a *comparative similarity relation* over possible worlds. In essence, this is a ternary relation between worlds A, B, and C, written $B \underset{A}{\preceq} C$, which states that B is more similar to A than C is to A. (The corresponding notion in Tversky's theory is expressed by $s(A, B) \geq s(A, C)$.) Formally, if U is the set of all possible worlds (including the actual world), then for all $A, B, C, D \in U$,

1. $A \underset{A}{\preceq} B$,

2. $B \underset{A}{\preceq} B$, and

3. If $B \underset{A}{\preceq} D$ and $D \underset{A}{\preceq} C$, then $B \underset{A}{\preceq} C$.

The first of these conditions says that the most similar world to a world is itself. The second and third conditions say that for each world A, the relation $\underset{A}{\preceq}$ is a reflexive and transitive relation, that is, a quasi-order over U. Lewis's notion is actually more involved than just these three axioms, but his extra conditions are not important here.

To use comparative similarity relations for describing analogies, let M be the class of all sets of items and relations, that is, of all model-theoretic structures over some universe. Let U be the class of all analogies between elements of M, that is, the class of all homomorphic mappings of structures in M. Then comparison of analogies can be captured as a comparative similarity relation over U. We pretend that $M \subseteq U$ by identifying each $A \in M$ with its identity mapping $id_A \in U$. We then write $B \underset{A}{\preceq} C$ just in case $B : A \rightarrow A'$ and $C : A \rightarrow A''$ are two analogies starting from A and B is a better analogy than C.

Note that these comparisons need not be complete, in that neither $B \underset{A}{\preceq} C$ nor $C \underset{A}{\preceq} B$ need hold. This means that this formulation is more general than using a single numerical function to measure degree of similarity. In that case, we would define $B \underset{A}{\preceq} C$ iff $s(A, B) \leq s(A, C)$. We say that a measure function s is compatible with $\preceq$ iff $s(A, B) \leq s(A, C)$ whenever $B \underset{A}{\preceq} C$. The possible incompleteness of $\preceq$ means that in general, many different measures s will be compatible with the comparative similarity relation. Incomplete orders arise when distances can be measured along several distinct dimensions, so that one must compare vectors of dimensions rather than a single number. Most practical systems, however, have been required to offer complete judgments. Thus in Carbonell's [1983] multidimensional comparison of analogies, and in Michalski and Stepp's [1983] clustering of points in a multidimensional space, the need to make overall comparisons leads to combining all the different dimensions into a single function, as in Tversky's treatment of similarity.

3 Conservatism

[Doyle 1983] developed a formal notion of conservatism identical to the formalization of similarity considered above to describe the programs for reason maintenance (RMS) and dependency-directed backtracking (DDB) described (under the name TMS) in [Doyle 1979], using a comparative similarity relation to compare states of the reasoner. In RMS, changing assumptions requires updating their consequences. Though RMS recorded and traced reasons to perform the update, its intent is better described as seeking to revise as few conclusions as possible. In DDB, contradictions are resolved by tracing reasons as well, but again the intent is better expressed as removing contradictions with as few consequences as possible.

In formalizing RMS, we view states as sets of nodes (representing propositions or reasons) drawn from a universal set $\mathcal{D}$. We write $\mathcal{I}$ to mean the set of all states of the agent, so that $S \subseteq \mathcal{D}$ for each $S \in \mathcal{I}$. The operation of reason maintenance can then be viewed as starting with a state S and a "kernel" set K of new assumptions and then moving to a new state S' which contains the new assumptions, that is, $K \subseteq S'$. Define $\mathcal{I}(K) = \{X \in \mathcal{I} \mid K \subseteq X\}$, and define $S' \underset{S}{\preceq} S''$ to hold iff $S \bigtriangleup S' \subseteq S \bigtriangleup S''$, where $X \bigtriangleup Y$ is the symmetric difference of X and Y, that is $(X - Y) \cup (Y - X)$. Since the aim of reason maintenance is to make as small a change as possible, we pick $S' \in \mathcal{I}(K)$ so that $S' \underset{S}{\preceq} S''$ for every $S'' \in \mathcal{I}(K)$, thus minimizing the set of changed elements. For reasons that are not relevant here, RMS failed to reliably achieve this ideal, but it is nevertheless a natural way to view its operation.

The specific sort of state-similarity appearing in RMS's conservatism has close connections with the basic notions of Tversky's theory of similarity. As described earlier, in that theory similarity judgments are mediated through comparisons of sets of features, just as conservatism compares states of mind by viewing them as sets of mental elements. But more interestingly, the "monotonicity" condition Tversky places on similarity measures is reflected in RMS's conservatism as well. Specifically, Tversky requires that $s(a, b) \geq s(a, c)$ whenever

1. $A \cap B \supseteq A \cap C$,
2. $A - B \subseteq A - C$, and
3. $B - A \subseteq C - A$.

But since $A \triangle B = (A - B) \cup (B - A)$, we see that Tversky's three conditions imply that $A \triangle B \subseteq A \triangle C$, so that $s(a,b) \geq s(a,c)$ implies $B \underset{A}{\preceq} C$ under RMS's definition of conservatism.

Though DDB actually removed contradictions by tracing reasons and removing "maximal" assumptions, we can use comparative similarity relations to capture its intent as well in the following way. If we write $\mathcal{I}^+$ to mean the set of contradiction-free states in $\mathcal{I}$, DDB seeks to conservatively remove contradictions from a contradictory state S by picking a state $S' \in \mathcal{I}^+$ such that $S' \underset{S}{\preceq} S''$ for every $S'' \in \mathcal{I}^+$.

4 Rationality

We have seen that the notion of comparative similarity relation appears naturally in similarity and conservatism comparisons in learning and reasoning. The abstractness of the notion of comparative similarity relation seems justified because in some cases there are many plausible candidates for such relations. For example, RMS might be just as interesting if it minimized the number of changed elements than the set of changed elements. But it seems natural to ask if some sorts of comparisons have better motivation than others. To answer that question, we must know more about where these comparisons come from. We propose to view comparative similarity relations as stemming from the notion of rationality. For the notion of rationality we employ the standard notion from decision theory, according to which an action is said to be rational for an agent at some instant if it is of maximal expected utility according to the agent's beliefs and preferences about current and future events, where the agent's preferences may be a function of its goals and plans. (See, for example, [Jeffrey 1983]).

The central notion of decision theory is that the agent may order its alternatives according to their expected utility. Expected utility is usually thought of as a numerical assessment of states, but this need not be so. One may also develop the theory in a "behavioristic" fashion from the choices a rational agent would make among different alternatives. Developed this way, the theory begins with a set of preferences among alternatives rather than probability and utility functions. If we let U stand for the set of all alternatives at an instant, then the agent's preferences are comparisons of the form $a < b$ for $a, b \in U$, such that the combined preference relation $<$ satisfies certain axioms of rationality, for example that preference is transitive. If we write $a \sim b$ to mean that the agent does not prefer either alternative to the other, that is, that $a \not< b$ and $b \not< a$, then the equivalence classes induced in U by $\sim$ can be identified as degrees of expected utility, and can be represented with a numerical measure of expected utility.

We propose to understand the quasi-orders appearing in similarity and conservatism in terms of the preference and indifference relations $<$ and $\sim$. In essence, we view $s(a,b) \geq s(a,c)$ as meaning that choosing b is preferred or equivalent to choosing c when the decision is one of selecting the most similar alternative to a, that is, that $c < b$ or $b \sim c$ when these choices refer to a. Correspondingly, we view $S' \underset{S}{\preceq} S''$ as meaning that in state S choosing to move to state S' is preferred or equivalent to choosing to move to state S'', that is, that $S'' < S'$ or $S' \sim S''$ in state S.

The nominal rationality of the comparative similarity relation is most apparent in the case of conservatively updating the agent's state. Simply minimizing the set of changes made, as RMS does, is not always reasonable, for this criterion does not distinguish between guesses and gospel. Thus RMS can in some cases discard fundamental laws and keep trivial hypotheses. To avoid such senseless revisions, conservatism needs to be rational conservatism, rationally selecting the update of maximal expected utility with, for example, frequently employed fundamental beliefs counting much more than ephemeral assumptions. Russell and Grosof [1987] propose to use this sort of rational revision for the case of shift of bias in learning, but it is an idea applying to all updates, not just shift of bias. Similarly, DDB's choice of a "maximal" assumption to remove or replace in order to restore consistency is made with the expectation that this choice will lead RMS to find a consistent conservative revision. A better way of proceeding is to choose the assumption rationally so that the expected payoff is highest among all possible choices of assumptions to remove. Thus RMS and DDB achieve only weak approximations to rational conservatism and rational backtracking.

Viewing similarity as stemming from rationality can be tricky. While it is easy to admit to having preferences about what to believe, it is less natural to think of observed similarities between objects as having anything to do with rationality. To view similarity as arising from rationality, we must step back and recognize that the rational agent thinks rationally as well as acts rationally. That is, the rational agent economizes on its mental resources. Every representation of information entails costs in memory and time: the memory needed to store the information, and the time needed to use it. There are well known tradeoffs between the succinctness of axiomatizations and lengths of proofs, and between expressiveness of languages and difficulty of finding proofs. These economic tradeoffs motivate different organizations for representation and reasoning, among them organizations of concepts according to similarity. In fact, Lakoff [1987] argues that many of the associations among concepts exhibited in human languages do not result from any fundamental connection of meaning as much as from the conceptual convenience these associations offer, and it is natural to interpret "conceptual convenience" as another phrase for rational choice of representation and clustering.

Consider, for example, the case of prototypes in taxonomic hierarchies. Tversky defines good prototypes to be those which (roughly speaking) maximize average similarity to the class of instances. But viewed more generally, this is just a special case of (or approximation to) maximization of expected storage costs. Hierarchies, of course, compress lots of data into simple descriptions and so offer dramatic (often exponential) economies of storage space. Some sorts of hierarchies are purely logical, with subconcepts entailing superconcepts, but in general the use of prototypes stem from economical rather than logical concerns. For example, say a prototypical concept is defined by a set or conjunction of n properties or aspects, and that objects satisfying any $n-1$ of these properties are counted as instances of the concept, albeit exceptional ones. Suppose further that we wish to describe $n+1$ individuals, one of which satisfies the concept perfectly, and n exceptional instances representing every possible exception. To describe these individuals in a system of prototypes and exceptions requires only $3n+1$ statements: n to describe the prototype, a typing or IS-A statement for each instance ($n+1$ all together) and a statement of the exceptional property for each of the exceptional instances (n of these). But to make the same descriptions using the ordinary logical connectives and implication requires n^2+n+1 statements: n for the prototype, one implication for the perfect instance, and n statements, variations of the prototype's definition, to describe each of the exceptional cases, n^2 in all. Thus the nonlogical system of prototypes can be much more efficient than a strictly logical representation. Of course this example considers only storage costs and does not take into account either limits on storage resources or expectations about what sorts of questions will be asked during retrieval, or with what frequency. These may greatly change the form of rational representation, but that is a topic we cannot pursue here.

5 Conclusion

We have seen how similarity, an important notion in the theory of learning, is closely related at the formal level to conservatism, an important notion in the theory of reasoning. Each of these topics deserves attention in its own right, but both may be naturally interpreted as arising directly from the agent's rational choices about how to represent information and how to reason with it. Recognizing this sort of rational self-government, moreover, helps to understand and connect numerous other aspects of thinking which, through the global character of rationality, can influence judgments of similarity and the practice of conservatism. (See [Doyle 1988a, 1988b] for further discussions.)

Acknowledgments

I thank Jaime Carbonell, Joseph Schatz, and Richmond Thomason for helpful discussions.

References

Carbonell, J. G., 1983. Learning by analogy: formulating and generalizing plans from past experience, *Machine Learning: An Artificial Intelligence Approach* (R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, eds.), Palo Alto: Tioga, 137-161.

Carbonell, J. G., 1986. Derivational analogy: a theory of reconstructive problem solving and expertise acquisition, *Machine Learning: An Artificial Intelligence Approach, Volume 2* (R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, eds.), Los Altos: Morgan-Kaufmann, 371-392.

Doyle, J., 1979. A truth maintenance system, *Artificial Intelligence 12(3)*, 231-272.

Doyle, J., 1983. Some theories of reasoned assumptions: an essay in rational psychology, Pittsburgh: Carnegie-Mellon University, Department of Computer Science, report 83-125.

Doyle, J., 1988a. Artificial intelligence and rational self-government, Pittsburgh: Carnegie Mellon University, Computer Science Department.

Doyle, J., 1988b. On rationality and learning, submitted to *Seventh Natl. Conf. on Artificial Intelligence.*

Jeffrey, R. C., 1983. *The Logic of Decision*, second edition, Chicago: University of Chicago Press.

Lakoff, G., 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, Chicago: University of Chicago Press.

Laird, J. E., Newell, A., and Rosenbloom, P. S., 1987. SOAR: an architeture for general intelligence, *Artificial Intelligence*, V. 33.

Lewis, D., 1973. *Counterfactuals*, Cambridge: Harvard University Press.

Michalski, R. S., and Stepp, R. E., 1983. Learning from observation: conceptual clustering, *Machine Learning: An Artificial Intelligence Approach* (R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, eds.), Palo Alto: Tioga, 331-363.

Mitchell, T. M., Keller, R. M., and Kedar-Cabelli, S. T., 1986. Explanation-based generalization: a unifying view, *Machine Learning* Vol. 1, No. 1, 47-80.

Russell, S. J., and Grosof, B. N., 1987. A declarative approach to bias in concept learning, *Proc. Sixth Nat. Conf. on Artificial Intelligence*, 505-510.

Tversky, A., 1977. Features of similarity, *Psych. Rev.* Vol. 84, No 4 (July 1977), 327-352.

Winston, P. H., 1975. Learning structural descriptions from examples, *The Psychology of Computer Vision* (P. H. Winston, ed.), New York: McGraw-Hill, 157-209.