Learning from Indifferent Agents

LISA DENT LISA.DENT@CS.CMU.EDU
JEFFREY C. SCHLIMMER JEFF.SCHLIMMER@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

(Received: March 14, 1990)

Abstract. In many situations, learners have the opportunity to observe other agents solving problems similar to their own. While not as favorable as learning from fully explained solutions, this has advantages over solving problems in isolation. In this paper we describe the general situation of *learning from indifferent agents* and outline a preliminary theory of how it may afford improved performance. Because one of our long-term goals is to improve educational methods, we identify a domain that allows us to observe humans learning from indifferent agents, and we summarize verbal protocol evidence indicating when and how humans learn.

Keywords: multi-agent domains, learning by doing, learning from examples, protocol analysis

1. Introduction

While getting out of my car, I saw another driver lock a bar-like object across his parked car's steering wheel, and I asked myself, "Why did he do that?" (As it turns out, it is a theft deterrence device designed to inhibit steering the car.) People often find themselves asking such questions, indicating that they build predictive models, apply them to other people, and attempt to explain surprising behavior. One benefit of this curiousity-driven behavior is that it can reveal novel strategies used by other agents; strategies that might otherwise be found only after extensive experimentation.

In more detail, let us consider three possible learning situations. In each, an agent has operators, a task to perform, and an indication of when the task is completed successfully; the agent must learn when to apply operators to complete the task. In the first learning situation, an agent learns in isolation, perhaps by trying operators and seeing what they do. This is called *learning by doing* (cf. Simon & Anzai, 1989), and the agent only sees outcomes that result from his/her actions. In the second situation, a teacher demonstrates successful and unsuccessful sequences of operators. This is called *learning from examples* (Carbonell, Michalski, & Mitchell, 1983), and the outcomes an agent sees are not biased by the current state of his/her learning. In the third situation, an agent learns with other agents that have similar operators and tasks but without a teacher. We call this *learning from indifferent agents* to denote our interest in situations where the agents are neither cooperative nor antagonistic. Here the agent has two sources of outcome observations: (a) those resulting from his/her own actions, and (b) those arising from the behavior of other agents. In this third situation the contexts surrounding observed agents' actions are likely to be underspecified because the reasoning of other agents may be based on unobservable factors.

Learning from indifferent agents involves a wide-spread class of situations that arise whenever there are multiple agents with comparable, but independent resources and tasks. Like the steering wheel example, people frequently have the opportunity to observe the behavior of

others without knowing exactly which factors lead to that behavior. This can happen because they are competing on a task or because the overhead of communication does not benefit the observed person. How do people make use of these observations? We hypothesize that learners may use the behavior of another agent for two ends: (a) as a source of novel, potentially useful operator applications ("Why did he do that?"), and (b) as a source of evaluation for current operator use ("She does that too"). We further hypothesize that learners will in fact make use of this information whenever they are able to adequately specify the situation surrounding the behavior. If too little is known about why the observed agent behaved as they did, then learners will be constrained to behave as if in a learning by doing situation. Conversely, if it is relatively easy to determine factors leading to the observed agent's behavior, then learners will learn higher quality knowledge more quickly.

2. A sample learning task

Card games are relatively simple task environments that afford opportunities for one agent to learn from another. Each player has comparable resources and goals. In some ways, card games relax unrealistic assumptions made in other games. Specifically, because players cannot see each other's cards and the cards are shuffled, card games do not involve the assumptions that agents have complete information or that the environment is completely deterministic. Chess, for instance, satisfies both of these assumptions, and Backgammon satisfies only the former.

As a domain of study, we have chosen the card game of *Mille Bornes* (1962, 1981) (or simply MB), which is modeled after a road race. Though the rules of the game are complex, and expert play is difficult, legal play in MB is relatively simple. Players begin with six cards, and on each turn they draw and then either discard or play a card. There are three types of cards: distance cards played to advance along the road (25, 50, 75, 100, and 200 miles), hazard cards played to temporarily stop an opponent (Flat Tire, Out of Gas, Accident, Stop, and Speed Limit), and remedy cards played to recover from hazards (Spare Tire, Gasoline, Repairs, Roll, and End of Limit, respectively). There is also a class of super-remedy cards called safeties that both remedy current hazards and protect from future ones (Puncture Proof tire, Extra Tank of gasoline, Driving Ace, and Right of Way, respectively). The primary object of the game is to travel either 700 or 1000 miles exactly, but the score depends on a number of factors, including: finishing first, using only distance cards less than 200, playing safeties, extending the play to 1000 miles, and keeping an opponent from playing any miles.

In this initial study, we chose two subjects who had played many card games before but had never played MB. They played MB against a computerized, expert player for about 1.5 hours each, their time consisting of three complete games of three or four hands each.[1] We asked the subjects to think aloud as they played, and we recorded their moves. One subject was instructed at the beginning of the second game to try to pay attention to the opponent's play and see what he could learn.

[1]Note that each play in a card game conveys information, but because subjects played against a computer implementation they correctly assumed that the opponent was neither trying to teach them nor trick them with its plays.

3. Evidence for learning from indifferent agents

While playing MB, subjects had the opportunity to learn by doing and to learn from the indifferent agent, i.e., the opponent. The subjects were free to choose if and when to use either technique. We are interested in how often learning from the opponent was used, in which situations it was used, and how it took place.

Learning from the opponent is possible when the opponent uses a strategy that the subject can infer from the opponent's overt action. In MB the only action that conveys information from the opponent to the subject is the opponent's move, usually a single card either played or discarded. In order to explain the reasons behind the opponent's move, the subject must deal with two unknowns: the cards in the opponent's hand and the set of strategies that the opponent is using to choose a move. One way of dealing with these multiple unknowns is to assume that the opponent has the same set of move strategies as oneself. Using this assumption and the opponent's move, aspects of the opponent's hand can be inferred. As long as the probability that the opponent has such a hand is high, one may conclude that the opponent is using a known strategy. If, however, the probability that the opponent has a hand that would lead to the observed move is low, one may consider alternative strategies that the opponent may be using.

For example, suppose the subject has a strategy to play the highest mile card available when able to play. If the opponent plays a 100 mile card, the subject can explain this move by hypothesizing that the highest mile card held by the opponent was a 100, which is very likely. However, if the opponent discards rather than playing a mile card even when able to play, the move can only be explained by hypothesizing that the opponent has no mile cards. Although this is possible, it is not common, so it may be useful to consider other strategies that could lead to this behavior. If the opponent is close to the 700 mile mark, one alternative explanation for the discard is that the opponent is saving mile cards until he has enough to reach 700 exactly.

Revising strategies in light of another agent's actions requires information about the context of those actions. If expert strategies require reasoning about information that is unobservable (e.g., contents of the opponent's hand), then it will be difficult for subjects to learn from their opponents. The converse is not in general true, though: if expert strategies require reasoning about observable information (e.g., cards already played), it may still be difficult for subjects to learn new strategies because of the relative difference between their strategies and optimal ones (see related work on expert/novice differences, e.g., Chi, Feltovich, & Galser, 1981).

Using this as an initial theory of the process of learning from an indifferent agent, we expect that subjects may learn from the opponent's behavior in two ways. When a move is explained by a strategy known to the subject, the subject may use this information as a positive evaluation of the strategy, because the opponent is known to be an expert player. When the move cannot be satisfactorily explained by current strategies, a new strategy may be learned if the context of the expert strategy is observable. Subjects learned from the opponent in both of these ways.

3.1 General observations

To test the initial theory just presented, we first examined the comments made by the subjects after each of the opponent's moves. Comment sequences referring to *why* the opponent made a particular move make up 3% of the total number of protocol statements, showing that the subjects are interested in the opponent's strategies and consider them a possible source of useful information. The majority of the remaining statements concerned the state of the game and the subject's own strategies.

Table 1. Learning from an indifferent agent: hypothesized factors and responses exhibited by subjects.

Strategy Name	Strategy Features		Type of Analysis			Total
	Observ.	Times Used	Confirm	Novel	Attempt	
Discard-useless	high	30	5	2	2	9
Extension	medium	9	1	0	1	2
End-limit	medium	4	0	1	0	1
Discard	low	195	1	1	6	8
Safety	low	26	0	0	0	0
End-game	low	13	0	0	0	0
Totals	—	277	7	4	9	20

We then categorized analyses made by the subjects into three groups. Analyses that explained the opponent's behavior in terms of a strategy currently used by the subject were classed *confirming*. Analyses that explained the opponent's behavior in terms of strategies not yet considered by the subject were classed *novel*. Finally, analyses that attempted to explain an opponent's move but were unable to do so satisfactorily were classed *attempted*. Specific examples of these types of statements appear in the next section.

According to our theory of the process of learning from the opponent, some strategies should be easier to infer from the opponent's move than others. Strategies with conditions that depend little on the cards in the opponent's hand will be more observable to the subject and therefore easier to identify. To investigate this hypothesis, we identified six, relatively complex strategies that subjects failed to use correctly at some point in the games. The six strategies are:

- Discard-useless. How to identify useless cards to discard.
- Extension. How to decide whether to extend the game to 1000 miles.
- End-limit. How to decide when to end a Speed Limit hazard.
- Discard. How to choose a card to discard when none are useless.
- Safety. How to decide when to play a safety card.
- End-game. How to decide which miles to play as the final mileage point (which must be reached exactly) is approaching.

Table 1 lists for each strategy: (a) the observability of the strategy, (b) the number of times the opponent used the strategy,[2] and (c) the number (by type) of analyses performed by subjects. At this level of detail, strategies are composed of a number of sub-strategies. We have initially encoded these using in a form similar to Siegler's (1976) encoding of strategies for the balance scale task. Each analysis made by the subject pertains to a particular sub-strategy. Because the observability of the strategy varies from sub-strategy to sub-strategy, Table 1 lists an overall qualitative assessment of observability.

[2]The number of times a strategy is used is approximated by applying a rational model to the observed agent's behavior; exact information is unavailable because the observed agent's hand is unknown.

The table shows that in general the observability of a strategy does increase (a) the number of analyses attempted by the subjects, and (b) the number of cases where a novel strategy is learned. For the high observability Discard-useless strategy, subjects display considerably more analyses than for the low observability Safety strategy along both of these dimensions (total analyses 9 to 0, novel analyses 2 to 0). However, the Discard strategy does not follow this pattern. Although it is a low observability strategy, subjects still analyze it eight times. Another factor appears to be operating here: the frequency of exposure to a strategy. As the table shows, the Discard strategy is used by the opponent extremely often. Perhaps with repeated exposure to the aspects of the strategy which are observable, the opponent's hand can be successfully inferred. This frequency must be very high to overcome the hindrance of low observability; subjects give no analyses of the low observability Safety strategy in spite of the fact that it occurs three times as often as the medium observability strategy Extension, which receives two analyses. Another potentially relevant factor, which we have not considered, is the complexity of the strategy compared to the subject's skill level.

We now turn to a specific class of strategies to illustrate the process of learning from indifferent agents in more detail.

3.2 A case study: The discard strategies

To supplement the high-level, quantitative analysis in the previous section, in this section we focus on one subject's specific analyses of his opponent's discard behavior. Because the discard action is a part of many card games, a person familiar with other card games (e.g., Gin Rummy) brings discard strategies from these games to MB. These become the initial set of discard strategies, which are modified during play as the person learns about the unique features of MB. Studying the protocols reveals that the subjects' simple initial discard strategies are:

1. Discard useless cards first.
2. If no useless cards, discard card of least expected utility:
   (a) Discard duplicate cards by rank.
   (b) Discard duplicates of a category by rank.
   (c) Discard singles by rank.

These initial strategies are expanded and modified by the subjects as the game progresses. The opponent's fixed set of discard strategies corresponds to the fully expanded forms of these two strategies.

The first strategy, the Discard Useless (DU) strategy, requires the identification of useless cards. Because this feature is unique to MB, the subjects must learn it as they play. There are several cases in which a card becomes useless. These are listed below along with the game and hand in which the strategy is first used by a subject, if at all. Note that most of the cases do not require information about the opponent's hand and thus are highly observable. The opponent uses a complete version of the DU strategy throughout the games.

1. If the player has played a safety card, the corresponding remedy is useless (Game 1, Hand 1).
2. If the opponent has played a safety card, the corresponding hazard is useless (Game 1, Hand 1).

3. If all hazard cards of a particular type have been played, the corresponding remedy is useless (not used).

4. If the player has played two 200 cards, more 200 cards are useless (not used).

5. If the player is within 50 miles of his final point, the End of Limit (EOL) card is useless (Game 3, Hand 2).

6. If the opponent is within 50 miles of his final point, the Speed Limit (SL) card is useless (Game 3, Hand 2).

7. If the player is approaching a final point, mileage cards that put the player over that point are useless (Game 3, Hand 2).

Because the DU strategy is highly observable and often used by the opponent, we expect that subjects will be able to acquire it by observing the opponent's actions. One subject first identifies an instance of a useless remedy by observing the relationship among his own cards. When he needed to discard for the first time, he used Case 1 of the DU strategy. Later in the same hand, the opponent discarded a useless remedy, and the subject was quick to notice and explain this behavior, perhaps confirming his own strategy in the process (Game 1, Hand 1, *confirming*):

> He discarded a Gasoline, obviously, cause he's got an Extra Tank.

Thus the subject acquired the first case of the DU strategy by doing rather than from the opponent, using the opponent's moves only to confirm the strategy.

An interesting example of learning a novel DU strategy from the opponent occurred in the subject's acquisition of the 5th and 6th cases of the strategy. When the subject had 675 miles, he observed the opponent discard a Speed Limit and attempted to explain the opponent's behavior using the DU strategy (Game 3, Hand 2, *novel*):

> He discarded a Speed Limit. Wonder why he did that? Oh, because he knows that I only need a 25 to win is that why?

The subject inferred that the opponent discarded the Speed Limit because it was a useless card. Another possible explanation, and in fact the correct one, is that the opponent had two Speed Limit cards. The subject then deduced a version of the fifth instance of the DU strategy:

> So I get rid of the End of Limits obviously.

However, these versions of the DU strategy are slightly incorrect. They test whether the player is within 50 miles of 700, not whether the player is within 50 miles of the *final point*, which may be 1000 miles if the player elects to extend the game. In fact, the subject did go on to extend this game, and after doing so discovered his earlier error.

This episode illustrates some of the pitfalls of learning from indifferent agents. It seems likely that strategies with complex conditions are prone to such errors (especially if the strategies are completely new to the subject). However this episode also shows that the opponent's moves can be used to acquire strategies that might otherwise be missed.

The DU strategy is highly observable. If a play is not possible, then its use does not depend on other cards in the hand. This is not true of the second initial strategy (discard the card of least expected utility), making it more difficult to learn by observing the opponent.

However, after observing the opponent discard a 25 mile card, the subject is successful at modifying a discard strategy. Up to this point the subject had been ranking all mile cards above remedy and hazard cards. He noticed the opponent discarding a 25, and considered the possibility of changing his ranking of low mileage cards in certain situations (Game 2, Hand 1, *novel*):

> He had to discard, which is good. He discarded his lowest point value. So it seems like one idea is to ...25s don't seem that important, especially if you have two of them like I do ...because you may need a 25 to finish ...to get exactly to 700, but ...

More often, however, a subject remarks on the opponent's discard and gives an unsatisfactory explanation, unable to infer the correct strategy from the observation (Game 2, Hand 1, *attempted*):

> He just discarded a Roll Card, boy, why did he do that? He must be pretty desperate to discard a Roll Card. He must have a lot of those.

These specific examples of learning discard strategies from observing the behavior of the opponent illustrate two important features of our theory. First, subjects attempt to explain opponent's moves first in terms of their own strategies and then, if necessary, in terms of novel strategies. Second, they are more often successful in their explanation when the conditions of the strategy are highly observable. Other factors influencing their success include the number of opportunities to observe the strategy in use and the complexity of the strategy itself.

4. Conclusions

Whether playing a game or living their daily lives, people pay attention to the behavior of other agents, and in some cases, they learn from those observations. In terms of the information that the learner sees, the task of learning from indifferent agents is a composite of learning by doing, where the learner generates learning opportunities, and learning from examples, where another agent generates learning opportunities.

Our initial theory postulates a specific relationship between the nature of an observed strategy and the ease of learning it from an indifferent agent: the less observable the factors measured by the strategy, the more difficult it will be to learn. This has a direct application for the design of instructional situations. In a field-classroom setting, to facilitate student learning by watching an expert perform a task, ensure that the factors weighted by the expert are clearly identified for students. It may not be as important to identify the expert's interpretation of those factors explicitly. Applying this principle to the card game we studied here, we hypothesize that the failure to learn one of the low observability strategies (e.g., Safety) may be facilitated by the inclusion of even small amounts of additional information (e.g., the number of safeties that the opponent has).

Our initial theory is far from a complete theory predicting the efficacy of learning from an indifferent agent. The protocols reveal that subjects do not always learn from their opponent even when the optimal strategy measures only observable features, that they do not consistently apply strategies they have verbalized, and that they do not consistently notice behavior that is inconsistent with their verbalized strategies. A more complete theory would require accounting for the semantic distance between the learner and the observed agent's strategies, practice effects, and attentional mechanisms. If an accurate and comprehensive theory could

be formulated, it would have a significant impact on the design of instructional materials and artificial learning methods.

Acknowledgements

We would like to thank Randy Jones for his helpful suggestions and careful reading of an earlier draft of this paper. This research is supported by the National Science Foundation under grant IRI-8740522 and by a grant from Digital Equipment Corporation.

References

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). An overview of machine learning. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach.* San Mateo, CA: Morgan Kaufmann.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–152.

Mille Bornes rules (1962, 1981). Beverly, MA: Parker Brothers.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*, 481–520.

Simon, H., & Anzai, Y. (1989). Theory of learning by doing. In *Models of thought.* Yale University Press.