Proceedings of the 18th Annual Conference of the Cognitive Science Society, La Jolla, CA, July 12-15, 121-125 (1996)

Dynamics of Rule Induction by Making Queries: Transition Between Strategies

Iris Ginzburg
Terrence J. Sejnowski
Howard Hughes Medical Institute
The Salk Institute
10010 N. Torrey Pines Rd.
La Jolla, CA 92037
e-mail: iris@salk.edu

Abstract

The induction of rules by making queries is a dynamical process based on seeking information. Experimenters typically look for one dominant strategy that is used by subjects, which may or may not agree with normative models of this psychological process. In this study we approach this problem from a different perspective, related to work in learning theory (see for example Baum 1991, Freund et al. 1995). Using information theory in a Bayesian framework, we estimated the information gained by queries when the task is to find a specific rule in a hypothesis space. Assuming that at each point subjects have a preferred working hypothesis, we considered several possible strategies, and determined the best one so that information gain is maximized at each step. We found that when the confidence in the preferred hypothesis is weak, "Confirmation Queries" result in maximum information gain; the information gained by "Investigation Queries" is higher when the confidence in the preferred hypothesis is high. Considering the dynamical process of searching for the rule, starting with low confidence in the preferred hypothesis and gradually raising confidence, there should be a transition from the "Confirmation Strategy" to the "Investigative Strategy", as the search proceeds. If we assume that subjects update their beliefs regarding the task, while performing, we would expect that the "Positive Confirmation Strategy" would yield more information at low confidence levels while the "Negative Confirmation Strategy" (simple elimination) would be more informative at higher confidence levels.

We tested subjects performance in such a task, using a paradigm introduced by Wason (1960). All subjects first assumed a hypothesis and then made positive confirmation queries. Upon receiving confirmation, half the subjects presented negative confirmation queries and later, half switched into investigative queries before attempting to guess the experimenter's rule. Also, the frequency of queries in the more 'advanced' strategies went down as the confidence level required to evoke the strategy went up. We conclude that subjects appear to be using different strategies at different stages of the search, which is theoretically optimal when queries are guided by a paradigm that maximizes information gain at each step.

Introduction

Investigating the induction of rules by making queries is an active area of research. One of the pioneering studies in this field was by Wason (1960) who presented subjects with a rule induction task and found that subjects tended to make queries that conformmed to the rule they have in mind, counter to the 'normative' approach suggesting that subjects should try to disprove their hypotheses. More recent studies (Klayman & Ha, 1987), (Oaksford & Chater, 1994) suggested that under some conditions it may be better to use a confirmation strategy rather than the disproving one.

In general, the information gained at each step of the search depends on the hypothesis space in which the search is conducted and on the apriori beliefs regarding the probability for each hypothesis in this space to be the target hypothesis. An additional factor is how the beliefs are updated given new information. In this study, we present a theoretical framework for a search in a large hypothesis space and study the information gained by several search strategies. Although the space is large, we study strategies that consider explicitly a small number of working hypotheses, while making very general assumptions regarding the rest of the space.

We predict that different strategies will be adopted at different stages of the search, as the confidence in the working hypotheses increases. We compare our theoretical results with behavior of subjects in a modified experiment along the lines of the 'triples guessing game' suggested by Wason (1960).

Theory

Here we present and analyze a theoretical model for the rule induction problem. We suggest an optimal behavior based on maximization of the information gained by individual queries. We define four types of queries and find under which conditions it is best to use which strategy. We begin with definitions. The search is conducted in a hypothesis space $\{H\}$ which we assume contains N independent hypotheses. Each hypothesis, h_n, defines a unique subsets over a data space $\{D\}$. Each hypothesis can be represented by a binary function which has value 1 at data points that lie within the specific subset, and value 0 at other data points in $\{D\}$. One of the hypotheses, h_t, is chosen to be the target hypothesis which an 'optimal seeker' needs to find by making queries. With each query, the seeker chooses a point d

in $\{D\}$ and is then informed whether this data point belongs to the target subspace or not. At each point in the search, the knowledge that the seeker has about the solution is represented by assigning each one of the hypotheses a probability that it is the target hypothesis, $p[h_n] \equiv p[h_n = h_t]$. The result is a probability distribution over the hypothesis space which reflects the knowledge or uncertainty of the seeker. The entropy S of the probability distribution over the hypotheses space is a measure of uncertainty:

$$S(P) = -\sum_{n=1}^{N} p[h_n] \log(p[h_n]) \quad (1)$$

where P is the probability distribution $P \equiv \{p(h_n)\}$. The entropy increases with the uncertainty, which can be easily demonstrated by two extreme cases: (a) When the target hypothesis is known and there is no uncertainty, $p[h_t] = 1$ and all other probabilities are 0. It is easy to see that in this case the entropy is 0. (b) When the uncertainty is at its maximum, all hypotheses are equiprobable, $p[h_n] = 1/N$. The corresponding entropy is $\log(N)$, which is the maximum value the entropy can have in this case. Normally, there is some knowledge regarding the probability distribution and the entropy will have intermediate values. Prior to choosing a query d, there is a prior probability distribution over the hypotheses space, $P^o \equiv \{p^o[h_n]\}$ with a corresponding entropy $S(P^o)$ as defined by (1). Upon receiving the information $h_t(d)$, the seeker updates the probabilities and the result is a set of posterior probabilities $P^p \equiv \{p^p[h_n]\}$. Basically, any hypothesis h_n such that $h_n(d) \neq h_t(d)$ is ruled out, i.e. $p^p[h_{n'}] = 0$. We assume that the probabilities of the remaining (surviving) hypotheses, are updated according to Bayes rule. The uncertainty changes to $S(P^p)$. This change depends on the value that will be obtained in reply to the query, which is not known in advance. Thus, the expected uncertainty after making the query is given by the expected entropy

$$\begin{aligned} ES(P^p) \quad &= p[h_t(d) = 1] S(P^p | h_t(d) = 1) \\ &+ p[h_t(d) = 0] S(P^p | h_t(d) = 0) \end{aligned} \quad (2)$$

The difference between the prior entropy and the posterior expected entropy is the information gained by the query:

$$EIG = S(P^o) - ES(P^p) \quad (3)$$

The posterior probabilities may be updated using different rules, but in this study we choose to update the beliefs using Bayes' rule. Therefore,

$$p^p[h_n | h_t(d)] = \frac{p[h_t(d) | h_n] p[h_n]}{p[h_t(d)]} \quad (4)$$

where $h_t(d)$ can be equal to 1 or 0. Under these conditions it is easy to see that the expected information gain is equal to the entropy of the information source $h_t(d)$, since the hypotheses are deterministic functions over the data set $\{D\}$:

$$\begin{aligned} EIG = \quad & -p[h_t(d) = 1] \log(p[h_t(d) = 1]) \\ & -p[h_t(d) = 0] \log(p[h_t(d) = 0]). \end{aligned} \quad (5)$$

In order to compute the expected information gained by any query, one needs to estimate the prior probability of the possible values of $h_t(d)$ and the best strategy would be to choose queries that are expected to yield 0 or 1 with probability of 0.5. In other words choosing the query that is least predictable will yield the maximal information gain. In order to do so, however, one needs to know the values of many hypotheses at each data point, an ability that requires a large memory capacity when dealing with a large space, see (Freund et al. 1995). Since humans have a limited capacity for working memory, the strategies used by subjects will probably not be optimal, leading perhaps to the use of more than one strategy at different times, in a way that approximates the optimal strategy.

We now apply this for the special case of rule induction. One of the main limiting factors is the memory capacity, which translates in this case to the number of hypotheses subjects can consider simultaneously. We consider strategies that involve a small number of explicit working hypotheses; other hypotheses are considered implicitly by assuming that there is some average probability for a 'random' hypothesis to have a value 1 at any specific data point

$$r = p[h_n(d) = 1]. \quad (6)$$

This is equivalent to assuming that there is a typical size of the subsets corresponding to the hypotheses in the space. We define four different strategies, but our analysis could be easily expanded to more.

The first two types of strategies are the "Confirmation Strategies", which can be positive or negative. When subjects have one working hypothesis, and do not consider any alternatives, they can make two types of queries. The "Positive Confirmation Strategy" (PCS) queries are those data points that conform to the working hypothesis, in other words: theses are the 'classical' confirmation queries (see [Wason 1960]). The "Negative Confirmation Strategy" (NCS) queries are those which yield a negative reply with respect to the working hypothesis. In both cases, subjects consider only THE WORKing hypothesis, and no alternative hypotheses, as a guide to the queries that they make.

Let h_w be the working hypothesis, then this implies that

$$p^o[h_w = h_t] > p^o[h_n = h_t], \qquad n \neq w. \quad (7)$$

Since we assume that all other hypotheses are equiprobable, the probability of all other hypotheses is

$$p_n = (1 - p_w)/(N - 1), \qquad n \neq w \quad (8)$$

There are two possible types of queries: the Positive Confirmation query, with a data point d_{pc} for which $h_w(d_{pc}) = 1$ and the Negative Confirmation query, with a point d_{nc} for which $h_w(d_{pc}) = 0$. When choosing a Positive Confirmation query, d_{pc}, we can estimate the prior probability $p[h_t(d_{pc}) = 1]$. If the working hypothesis is correct, $h_w = h_t$, the reply will be 1. There is also a fraction r of the hypotheses for which $h_{n'}(d_{pc}) = 1$ defined in (6). The probability that the reply will be 1 is:

$$p[h_t(d_{pc}) = 1] = p_w + (1 - p_w) r \quad (9)$$

A similar analysis holds for the negative confirmation query, for which

$$p[h_t(d_{nc}) = 0] = (1 - p_w)r \qquad (10)$$

Two other types of strategies considered here are "Investigative Strategies" in which the seeker considers a working hypothesis and an alternative hypothesis simultaneously. Any query which conforms to one hypothesis but does not conform to the other, will enable the subject to rule out one of these hypotheses. We define the "Positive Investigation Strategy" (PIS) when a query conforms to the working hypothesis and does not conform to he alternative hypothesis. Similarly, define the "Negative Investigation Strategy" (NIS) when a query conforms to the alternative hypothesis and does not conform to the working hypothesis. Note that the only difference between the Confirmation Strategies and the Investigative Strategies is the existance of an alternative hypothesis.

Formally, we define the alternative hypothesis, h_a, as the hypothesis which is less favorable than h_w but is preferable to other hypotheses in the space.

$$p^o[h_w] > p^o[h_a] > p^o[h_n] \qquad n \neq w, a. \qquad (11)$$

The Positive Investigation query d_{pi} is defined by $h_w(d_{pi}) = 1, h_a(d_{pi}) = 0$. The probability that the reply will be 1 is:

$$p[h_t(d_{pi}) = 1] = (1 - p_w - p_a)r + p_w \qquad (12)$$

The Negative Investigation is similar to the Positive Investigation only it has the opposite values with respect to h_w and h_a and the probability that the reply will be 1 is:

$$p[h_t(d_{ni}) = 1] = (1 - p_w - p_a)r + p_a \qquad (13)$$

One can generalize our analysis to the case where there is more than one alternative; the results do not change significantly.

In summary, the strategies are defined by two criteria: First, how many favorable hypotheses are being considered simultaneously? A working hypothesis is always considered, but subjects may or may not consider an alternative hypothesis. Secondly, Is the favorable hypotheses positive or negative at the query point?
The factor that determines the informativeness of the query is the predictability of the reply, which we can estimate for each strategy. The average information gained by the different queries, using eqs. (3) & (5) is:

$$\begin{aligned} EIG \quad &= p[h_t(d) = 1] \log(p[h_t(d) = 1]) \\ &+ (1 - p[h_t(d) = 1]) \log(1 - p[h_t(d) = 1]) \end{aligned} \qquad (14)$$

where $p[h_t(d) = 1]$ is the probability of the reply to query d to be 1, and the closer $p[h_t(d) = 1]$ is to 0.5, the more informative the query is. We summarize the results in the following table:

Query Type	Data Point	$p[h_t(d) = 1]$
Pos. Conf.	$d = d_{pc}$	$(1 - p_w)r + p_w$
Neg. Conf.	$d = d_{nc}$	$(1 - p_w)r$
Pos. Inv.	$d = d_{pi}$	$(1 - p_w - p_a)r + p_w$
Neg. Inv.	$d = d_{ni}$	$(1 - p_w - p_a)r + p_a$

Table 1. Predictability of queries: probability of reply equals 1.

These conditions lead to a number of predictions, assuming that the optimal strategy is adopted. Regardless of what r is, the confirmation strategies are always preferable to the investigative strategies at low condifence $p_w < 0.5 + p_a r/2$ and the other way around when the confidence is high. Up to this critical value, $|p[h_t(d_{pc,nc}) = 1] - 0.5| < |p[h_t(d_{pi,ni}) = 1] - 0.5|$, and then the relation is reversed. From a similar comparison of the positive and negative confirmation strategies, when $r < 0.5$ the positive confirmation is always better than the negative confirmation and when $r > 0.5$ the negative confirmation is preferable to the positive confirmation. Although we have no direct measure of the subjective value of r using the current experimental paradigm, we assume that subjects update their subjective estimate of r along the search, from a low value at the start to a higher value after receiving frequent replies of 1, due to the specific design of the Wason test. In summary, we predict:

- Different strategies will be used by subjects according to the following order: positive confirmation, negative confirmation and last, investigation
- The later strategies, corresponding to higher confidence levels, will be less frequent since subjects guess the rule at different subjective confidence levels,
- Confirmation strategies should be correlated with low confidence, and investigative strategies with high confidence.

Experimental Paradigm

Subjects were asked to discover a mathematical rule that applies to triples of numbers by writing down sets of three numbers along with the reasons for the choice. They were also asked to write down their best guess of the unknown rule at this point and their confidence that this may be the correct rule. In addition, they were asked to note their prediction of what will be the experimenter's reply and their confidence in their best guess as well as the predicted reply. The confidence was rated as Low, Medium, High or Very High. Subjects were given one confirming example to begin the process, and all replies were kept on a form (Figure 1). Subjects were given only one chance to explicitly guess the rule, after which the process terminated. We analyzed the data according to the definitions of the strategies given in Table 1. which are demonstrated in Figure 2. We assumed that the best guess at each point was the working hypothesis of subjects.

Results

We analyzed data from 20 subjects, generating a total of 99 queries. We consideried the best guess to be the working hypothesis.

Confidence levels were transformed to numbers between 1-4. The most frequent strategy was the positive confirmation: 19/20 subjects used it at some point. The secondary most frequent was the negative confirmation: 10/20 subjects used it. The least frequently used is the investigation strategy: 5/20. This perfectly matched our theoretical predictions, although other explanations can be given for the same data. Since the Investigative strategies were much rarely used, we pooled data from the Positive and Negative Investigation queries together.

We consider each query to be an independent event. The confidence level was found to be correlated with the different types of strategies that are used (Figure 3). However, the two confirmation strategies were not significantly different, probably due to a small sample size. The investigative strategies were found to be used at significantly higher confidence levels with $p < 0.05$.

A similar analysis was performed for the ordering of the different strategies. Each query was given an ordering label within each subject's game, according to its sequential numbering normalized by the number of queries asked by each individual. That is, each query received a label between 0 and 1. There is a correlation between the type of strategy and the ordering (Figure 4). The Positive Confirmation strategy which is more commonly used in the beginning of the game, was found significantly different from the Negative Confirmation which is used at more advanced sequencing, with $p < 0.06$. The investigative strategy was found to be used at higher sequencing, and is significantly distinct from the Confirmation strategies with $p < 0.05$.

It was interesting that 16/20 subjects started their guesses by a positive confirmation query to which they expected a negative reply. This suggests that subjects initially expect that the probability of a random query to yield a positive reply is low, in accordance with our assumption that $r \ll 0.5$. 2/8 subjects who found the correct rule used only positive confirmation queries. These subjects performed a series of spontaneous changes in their working hypothesis.

Summary

We have presented a theoretical analysis of information search in a large hypothesis space. We have shown that subjects used different search strategies at different stages of the search, in a way that was correlated with confidence in their working hypothesis. The confidence level appears as a significant parameter in theoretically determining the best search strategy, as well as in predicting the behavior of subjects.

An important factor in our analysis is the assumption that subjects updated their subjective beliefs regarding the reply to a random query, a direction we intend to explore in the future.

In addition, one can easily show, using the paradigm we have presented, that as the hypothesis space becomes larger, it is less valuable to consider alternative hypotheses. Evidence for this notion was found by (Van Wallendaeland and Hastie, 1990) who showed that when the number of possible alternative hypotheses was large, subjects tended to update their beliefs regarding one working hypothesis only. As the space size was reduced (in that study the space is quite small), subjects began to update their belief in more than one hypothesis.

In our theoretical analysis we assumed that the subjects updated their beliefs about the likelihood of receiving positive (or negative) answers during the task, which is equivalent to the presumed size of the target subset. This correlates with the anchoring effect (Kahneman and Tversky, 1974): information that is explicitly not relevant to the task subjects are required to perform still affects the behavior of subjects.

Acknowldgments

I.G. is grateful to W. Bialek for motivating this study and to C. McKenzie for interesting and helpful discussions.

References

Baum E. (1991). Neural net algorithms that learn in polynomial time from examples and queries. IEEE Trans. Neural Networks, 2:5-19.

Freund Y., Seung H.S., Shamir E. and Tishby N. (1995). Information, prediction, and query by committee. preprint.

Kahneman D. and Tversky A. (1974). Judgment under uncertainty: Heuristics and biases. Science, V185:1124-1131.

Klayman J. and Ha Y-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. Psychological Review, V94. No 2.211-228.

Oaksford M. and Chater N. (1994). A rational analysis of the selection task as optimal data selection. Psychological Review, V101 (n4):608-631.

Van Wallendaeland L. R. and Hastie R. (1990). Tracing the footsteps of Sherlock Holmes: Cognitive representations of hypothesis testing. Memory & Cognition, V18 (n3):240-250.

Wason P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology, V12 129-140.

Three Numbers	Predicted Reply	Confidence in reply: L, M, H, V	Reasons for Choice	Best guess of rule at this point	Confidence in best guess: L, M, H, V	REPLY
example: 2, 4, 6						YES

Figure 1: The form used in the experiment.

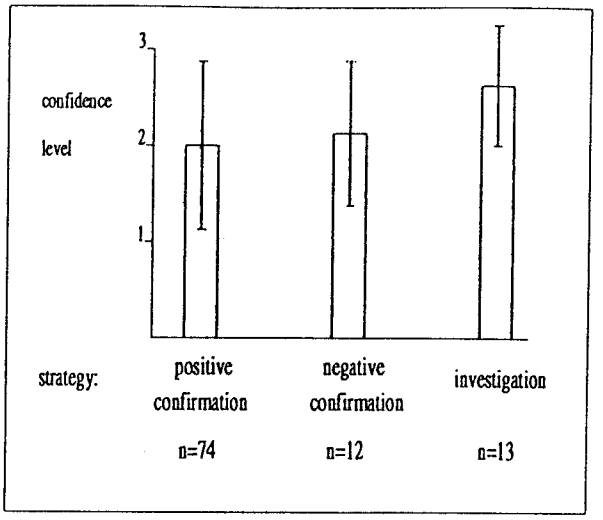

Figure 3: Correlation between strategy and confidence level. n is the number of queries that correspond to each strategy. The error bars represent the standard deviation of the distributions and not the deviations of the means.

Query Type	Three Numbers	Reasons for Choice	Best guess of Rule at this point
1. Positive Confirmation	6, 8, 10	confirm	even numbers increasing by 2
2. Negative Confirmation	1, 3, 5	test	even numbers increasing by 2
3. Positive Investigation	8, 10, 12	rule out possibility that numbers are multiples of the first number	even numbers increasing by 2
4. Negative Investigation	2, 6, 10	perhaps all triples of increasing evens	even numbers increasing by 2

Figure 2: Definition of four different types of queries.

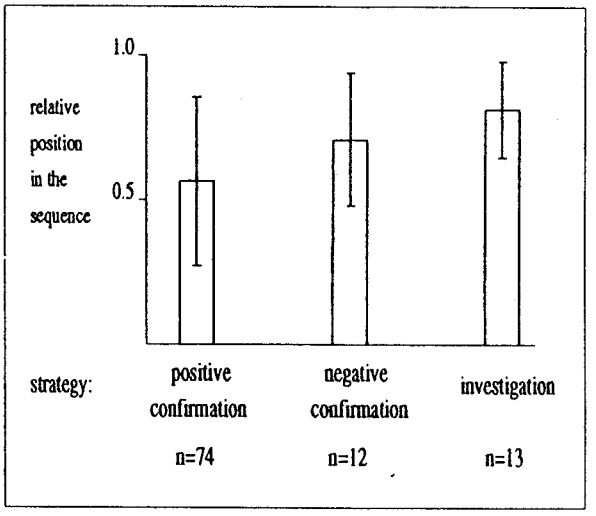

Figure 4: Order of strategies. n is the number of queries that correspond to each strategy. The error bars represent the standard deviation of the distributions and not the deviations of the means.