

Assessing Conceptual Complexity and Compressibility Using Information Gain and Mutual Information

Fabien Mathy

Université de Franche-Comté

In this paper, a few basic notions stemming from information theory are presented with the intention of modeling the abstraction of relevant information in categorization tasks. In a categorization task, a single output variable is the basis for performing a dichotomic classification of objects that can be distinguished by a set of input variables which are more or less informative about the category to which the objects belong. At the beginning of the experiment, the target classification is unknown to learners who must select the most informative variables relative to the class in order to succeed in classifying the objects efficiently. I first show how the notion of entropy can be used to characterize basic psychological processes in learning. Then, I indicate how a learner might use information gain and mutual information –both based on entropy– to efficiently induce the shortest rule for categorizing a set of objects. Several basic classification tasks are studied in succession with the aim of showing that learning can improve as long as subjects are able to compress information. Referring to recent experimental results, I indicate in the Conclusion that these notions can account for both strategies and performance in subjects trying to simplify a learning process.

Information theory is aimed at quantifying data that needs to be stored or communicated (Shannon, 1948). It was extensively used in psychology in the 1950s and 1960s, especially for measuring the maximal amount of information that can be transmitted by subjects without error (the reader will find historical presentations of this approach in Attneave, 1959; Coombs, Dawes, & Tversky, 1970). In one of *Psychological Review*'s most cited articles, Miller (1956) presented absolute-judgment experiments with a large variety of tasks and showed that human subjects performing unidimensional categorizations could not discriminate a set of stimuli using more than about seven

categories. For instance, subjects could correctly match a set of audible sounds (varying continuously from 11 dB to 60 dB) to a set of five predefined categories (11-20 dB, 21-30 dB, 31-40 dB, 41-50 dB, and 51-60 dB). However, if the number of categories increased beyond seven, subjects were no longer able to perfectly match the stimuli to the categories.

In the terms used in information theory, transmitters (subjects) are not able to perfectly communicate the original message (the correct category, theoretically associated to a sound they receive) to a destination (experimenter) when the number of possible categories is too high. Any amount of transmitted information can be measured by computing a correlation between the correct categories and the subjects' responses. When there is no loss of information in the transmitted message, the correlation between the input message and the output message is equal to one; when a task is too difficult for a subject, the loss of information shows up as an imperfect correlation between the input and the output. However, the amount of information transmitted is more commonly measured in bits. This amount can be as

This research was supported in part by the Agence Nationale de la Recherche Grant #ANR-09-JCJC-0131-01. Correspondence concerning this article should be addressed to Fabien Mathy, Université de Franche-Comté - UFR SLHS, 30-32 rue Mégevand, 25030 Besançon Cedex, France. Email: fabien.mathy@univ-fcomte.fr

high as the amount of information present in the original message when there is no noise in the message transmitted.

The following sections show that information theory can be used as a broader perspective to shed light on learning processes. Some aspects of the theory have been used implicitly in the past to develop decision-tree models in psychology, but other aspects, such as mutual information, have received less attention but can be effectively brought to bear in accounting for some experimental results.

Entropy

Information theory computes basic probabilities to determine the quantity of data a set of messages contains. This quantity, called entropy (or information content), corresponds to the amount of uncertainty there is in a set of messages. For instance, let us imagine that someone wants to guess the color of the suit of a playing card picked from a regular deck (the suit is either black or red). Because the probability of guessing the color is $\frac{1}{2}$ (half of the cards have a black suit), the communication of a single piece of information (e.g., black) is equal to $-\log_2(\frac{1}{2}) = 1$ bit, a measure which is also called the surprisal. A single message ("black" or "red") corresponds to 1 bit of information in the sense that one needs 1 yes-no question to retrieve the color (Is the suit red or black?). Corollarily, this means that coding the suit color of the cards only requires one binary digit (e.g., 0 if black, 1 if red). A system limited to communicating the suit color of a card needs a capacity of 1 bit. If a system does not properly communicate the suit color of a card, it either may have a capacity of less than 1 bit or be subject to unwanted perturbation. The general formula of the surprisal is therefore simply $-\log_2(p)$, where p is the probability of obtaining a given response by chance. The surprisal is maximal when $p = .5$, because for any other situation where the probability of an event approaches 1 or 0, the receiver is less surprised by the information transmitted by the sender (if a box is filled with 100 red balls and 0 blue balls, nobody would be surprised to hear that the ball just randomly drawn from the box is red).

In the preceding example on absolute judgments about sounds, the channel-capacity limit is thought to be about 2.8 bits because no messages of more than $-\log_2(\frac{1}{7}) = 2.8$ bits can be perfectly transmitted ($\frac{1}{7}$ is the probability of transmitting the right category by chance). Because guessing the sound category is more difficult than guessing the color of a card, the surprisal is higher for sounds (2.8) than for cards (1). By describing human abilities in terms of channel capacity, Miller (1956) implied that overly demanding tasks (requiring the encoding of more than 2.8 bits) will not be handled by subjects. Simultaneously, Miller began contributing to the decline of information theory by making

a distinction between bits of information and chunks of information.¹

The formula for entropy is a little more general than the one proposed above, because some messages can be more probable than others. The uncertainty one has about a set of possible messages (e.g., black or red, or one of seven categories) is called entropy or H . It is determined by computing the expected value² of the surprisal of all possible pieces of information a message might contain:

$$H(X) = \sum_i -p_i \log_2(p_i) \quad (1)$$

In this formula, i indexes all possible messages. For instance, the entropy for colors in a deck of 52 cards is: $H(\text{color}) = -p(\text{red})\log_2(p(\text{red})) - p(\text{black})\log_2(p(\text{black})) = -(\frac{26}{52})(-1) - (\frac{26}{52})(-1) = .5 + .5 = 1$.

For 7 categories, we have:

$$H(\text{category}) = -p(\text{cat}_1)\log_2(p(\text{cat}_1)) - p(\text{cat}_2)\log_2(p(\text{cat}_2)) - \dots - p(\text{cat}_7)\log_2(p(\text{cat}_7)) = -(\frac{1}{7})(-2.8) - (\frac{1}{7})(-2.8) - \dots - (\frac{1}{7})(-2.8) = .4 + .4 + \dots + .4 = 2.8.$$

In the above two examples, note that the entropy is equal to the surprisal of a single message because all messages have the same outcomes.³ In what follows, I show how information entropy can be used to model concept learning.

Generalities on Concept Learning

Concepts are abstract ideas that can be used to classify objects on the basis of their functions, shapes, taste, composition, etc. Once acquired, concepts can be used to generalize from prior experience. The formation of concepts can be approached by studying child development, or by studying the history of ideas over longer periods of time.⁴ Another way of obtaining data on concept formation is to carry out microgenetic studies to get a finer grained picture of developmental change. By further reducing the time window, concept formation can also be scrutinized during a single learning session. In this case, which is of primary interest here, the cognitive processes involved in learning are inferred from measures such as failure vs success, number of trials to criterion, number of errors during task execution, learning time, response time per trial, etc. Even more interesting is the subjective complexity of a given task, which can be inferred from the above-mentioned variables, with high values most often denoting difficulty acquiring a concept.

Concepts can be viewed as categorization situations confined to two categories only (the category of positive examples versus the category of negative examples), no matter how many input dimensions (features) there are. For instance, the concept of zebra helps separate the positive examples (zebras = Equidae with black and white stripes all over the body) from the negative examples (all other

animals); the number of legs is not a critical feature in this case, whereas being a horse-like animal and having stripes are two relevant dimensions. A good definition of a concept (also called the intension) prevents one from having to memorize the list of all positive examples (called the extension).

This article deals with rule learning and artificial concepts, first studied in the 1950's by Bruner, Goodnow, and Austin (1956). Artificial concepts are built by experimenters who arbitrarily assign membership to a list of stimuli. Learning artificial categories imply quite different processes from learning natural categories. In artificial settings, the informativeness of the dimensions is expected to be nearly equivalent (and existing differences can be controlled by randomizing the relevant dimensions), whereas in natural categories, the dimensions are not equally informative. Subjects tend to assign different values to dimensions in their natural conceptualizations. For instance, in the twenty-questions game, a clever strategy is to start by asking whether the unknown thing is living, since the answer to this question eliminates lots of possibilities (when young children ask questions that are too specific, for instance, "Is it Mommy?", a "No" response leaves them with very many possibilities). Another difference is that evidence for peculiar perception can be shown in natural category-learning situations, provided the features are sufficiently continuous. For instance, categorical-perception behavior might indicate increased sensitivity to items of different categories (Goldstone, 1994) as well as decreased sensitivity to items of a similar category. In this vein, Wood (1976) showed that adults do not perceive continuously changing series of artificial sounds from *b* to *p* but they hear an abrupt switch from *b* to *p*. In contrast, artificial category learning can be linked to a reasoning process involving the inductive formation and deductive testing of logical rules (e.g., if the positive examples are big and red and a big red object is displayed, then that object is positive). This is especially true when the number of features is small, when features are discrete, and when categories are not fuzzy. For instance, J. D. Smith, Minda, and Washburn (2004) showed that humans transcend slow association-based learning whenever possible and exhibit very sudden learning through rule discovery and insight, in comparison to monkeys who learn the same tasks via conditioning.

However, even natural conceptualizations invoke some forms of abstraction akin to rule learning: it has been shown that children do not conceptualize the world simply by considering the physical characteristics of objects but also by abstracting theories, such as "essentialism" which they incorporate in their biological beliefs. Also, if-then rules can be used to form first-order logic in order to reason about

predicates (if $\text{parent}(z, y)$ and $\text{parent}(y, x)$, then $\text{grandparent}(z, x)$). This is why rule-based models have a long history (Murphy, 2002). However, other theories refrain from assuming such deliberate high-level processing and model category learning as an implicit associative learning process based on perceived similarities between objects that can result from the homogeneity of the categories. There is a bulk of evidence for each of these two forms of reasoning (that can sometimes work dually, in line with many hybrid models), depending on the design of the categorization experiment (Sloman, 1996). Note that the experimental conditions related in this paper are likely to produce data that appear to support deterministic rules, but similarity-based models are known to perfectly account for many results that would be obtained in such conditions in terms of pure exemplar-storage schemes (Nosofsky, Gluck, Palmeri, McKinley, & Gauthier, 1994).

Here, for the sake of simplicity, I focus on (1) stimuli built from Boolean dimensions (each taking on two different values only, like squares versus triangles for a shape dimension) and (2) separable dimensions (such dimensions can be consciously identified and separated by subjects, contrary to integral dimensions such as hue and brightness in colors, see Garner, 1974; for instance, shapes and colors are simple separable dimensions). When merging two separable binary-valued dimensions to build a set of simple stimuli, one can create four stimuli, that is, a blue square, a blue triangle, a red square, and a red triangle, forming what is called a training sample (also called a block of stimuli when the stimuli are displayed sequentially). Such canonical stimulus sets have been studied extensively since the pioneering work by Shepard, Hovland, and Jenkins (1961). Also, and again, for the sake of simplicity, I only focus here on supervised learning: the category label is provided to the learner whenever they are wrong, and participants gradually learn the appropriate classification by an error-driven process through which they adapt their responses to the feedback. Note that supervision necessarily takes place during a training phase (as opposed to a test phase, in which a new set of stimuli is given to subjects in order to test the generalizability of their concepts; in test phases, supervision is not mandatory). The model presented here aims to account for how subjects learn and use a concept during a training phase. In classical experimental settings, subjects are required to classify a set of stimuli displayed sequentially. The learner is presented with blocks of stimuli in random order, with each stimulus appearing once. Because learning is supervised, subjects progressively become able to correctly categorize the objects. In theory, computer simulations imply the same stimulus sets as in the experiments run on subjects, but in reality the modeling is

based on basic formalizations and alleviate the need to do simulations.

There are potentially many different Boolean concepts to be learned, depending on the number of dimensions, the number of positive examples, and the structure of the categories (Feldman, 2003). Our goal is to account for subjective complexity by investigating the learning mechanisms involved in such tasks. Most empirical studies evaluate the subjective complexity of tasks by measuring the number of trials to criterion or the proportion of correct responses for each block of stimuli in the task, and then using these measures to compare concepts with each other. However, more refined analyses of fit can be based on typicality judgments, response times, and proportion correct for each stimulus within a given concept (Lafond, Lacouture, & Cohen, 2009). All of these measures can be predicted from the model I develop here.

Some Examples of Boolean-Concept Learning Tasks

Let us begin with the example of the exclusive OR (called XOR) structure shown in Figure 1, first studied by Neisser and Weene (1962), and then used as a canonical example in how neural networks perform (Minsky & Papert, 1969; Rumelhart, Hinton, & Williams, 1986). XOR is presented below:

$$XOR = \begin{pmatrix} X & Y & Z \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Each column is a variable that can take on the value 0 or 1 (in such cases, the concepts are called Boolean). The columns are labelled X , Y , and Z . The last column represents the category variable and the other columns the input variables. This kind of truth table is convenient for describing the partitioning of objects (the positive examples are denoted by the 1's in column Z , and the negative examples are denoted by 0's in column Z). The input variables could describe color options or size options such as small vs big, blue vs red, etc. For instance, the first line of the table would indicate that the small (coded 0) blue (coded 0) object is not part of the positive category (coded 0). The entropy H of each of the variables is 1 bit because 1 bit of information is needed to store or communicate one of the two equally probable values (0 or 1) that can be taken on by each variable. More formally, we have:

$$\begin{aligned} H &= -\sum_i p_i \log_2(p_i) = -p(0) \log_2(p(0)) - p(1) \log_2(p(1)) \\ &= -\frac{2}{4}(-1) - \frac{2}{4}(-1) = .5 + .5 = 1 \end{aligned}$$

The question here is how can subjects make use of

information in the input variables to categorize the examples in the most efficient manner? I will focus here on strategies consistent with Occam's razor (1324). The idea is that formulating simpler hypotheses is preferable because such hypotheses generalize better (Blumer, Ehrenfeucht, Haussler, & Warmuth, 1987), not mentioning that simpler hypotheses are less memory-demanding.⁵ One strategy is to start by searching for the most diagnostic variable (if there is no totally diagnostic one). Then, the learner moves on to choose a second variable to complete the first one, and so on, until a minimal set of variables are ordered in the most efficient manner to categorize the objects. This gradual strategy is consistent with the idea of abstracting a rule and searching for exceptions. A second strategy consists of simultaneously considering the set of input variables in an attempt to discover some relationships that might be helpful for the categorization process. These two strategies can be modeled respectively using two notions developed in information theory: information gain and mutual information. I will show how these notions can account for complexity in terms of compressibility (I refer here to lossless compression rather than lossy compression of information). We will see that in that respect, XOR is a difficult concept according to information gain measures but a simple concept according to mutual information measures.

Information gain

Before describing information gain, a concept simpler than XOR should be examined:

$$SIMPLE = \begin{pmatrix} X & Y & Z \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

In the SIMPLE concept, the variable X is clearly correlated to the class Z (contrary to Y , and contrary to both X and Y in XOR). Therefore, subjects can use a basic rule such as "IF $X = 1$ THEN $Z = 1$; ELSE $Z = 0$ ". The only difficulty for subjects is to induce such a rule from the training sample. However, it should be pretty obvious for subjects that the two values of X are perfectly correlated to the category values, contrary to Y . In other terms, the probability of getting the right answers by focusing on X is maximal. Subjects only need to turn their attention to X to notice its relevance to the task.

Third example:

$$UNSPECIFIED = \begin{pmatrix} W & X & Y & Z \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Here, when $W = 0$, the situation is similar to the *SIMPLE* concept, that is, “IF $X = 1$ THEN $Z = 1$; ELSE $Z = 0$ ”. However, when $W = 1$, we simply have $Z = 1$. Intuitively, the simplest rule is “IF $W = 1$ THEN $Z = 1$; ELSE [IF $X = 1$ THEN $Z = 1$; ELSE $Z = 0$]”. This embedded rule corresponds to a decision-tree structure in which the value of W is tested first, and then X is tested whenever $W = 0$. We can then predict the following for stimuli for which $W = 1$: (1) They will be quickly and correctly categorized by subjects, given that subjects focus first on the most diagnostic features and postpone learning exceptions, (2) they will need only one step to be identified as positive examples (so response times should be short), (3) they should benefit from automatization as the task progresses, until the subject manages to learn the other stimuli in which $W = 0$ (so response times should improve over time), and (4) they should be perceived as more typical of the positive category.

The question is how can we model the induction of such efficient rules? First, note that simple statistical associations can help the learner select the most diagnostic dimension. For instance, by using the Y feature, the subject would get $\frac{4}{8}$ correct responses. However, using the W feature, the subject’s score would be better ($\frac{6}{8}$ correct). Indeed, Z and W do not match only for the third and fourth stimuli. The subjects might pick this dimension to start with, and then try to obtain more information about objects that are not correctly categorized when $W = 0$. The subjects might quickly notice that X can help identifying the correct category when $W = 0$.

Using information theory, this strategy might be modeled as follows (I refer here to ID3 developed by Quinlan, 1986, and summarized in Mitchell, 1997; Boden, 1996; Luger, 1994). For the *UNSPECIFIED* concept, the information content is:

$$\begin{aligned} H &= -p(0) \log_2(p(0)) - p(1) \log_2(p(1)) \\ &= -\frac{6}{8}(-.42) - \frac{2}{8}(-2) = .81 \end{aligned}$$

Note that the information content is less than one because 1’s and 0’s are not equiprobable, which gives the subjects a greater chance of guessing the right category of an object (more often equal to 1). The effectiveness of an attribute in classifying the objects can be measured by the information gain that this attribute provides. To obtain the information gain associated with using one variable, we compute the entropy for each attribute of that variable. Let’s begin with W . Given $W = 0$, the entropy is: $H_{W=0} = -(\frac{2}{4})(-1) - (\frac{2}{4})(-1) = 1$ (because half of the objects are positive when $W = 0$). Given $W = 1$, $H_{W=1} = -(\frac{4}{4})(-0) = 0$ (because all objects are positive when $W = 1$). This means that when using W , there is 1 bit of uncertainty left when $W = 0$ whereas there is no uncertainty

left when $W = 1$. Next we compute the mean entropy of the above two measures: $H_W = (H_{W=0} + H_{W=1}) / 2 = (1+0) / 2 = .5$, which indicates that on average, the learner is left with .5 bits of uncertainty when using W .

The information gain for W is: $H - H_W = .81 - .5 = .31$, which is larger than $H - H_Y$, but equal to $H - H_X$. Let us arbitrarily pick W instead of X (since W and X involve a similar gain in information) and let us focus on the objects for which some uncertainty is left (there is 1 bit of uncertainty left when $W = 0$). The learner might want to try other dimensions to fill in this amount of remaining uncertainty. The information gain can be computed in a similar fashion for the four objects left when $W = 0$. If $H_{X/W=0}$ is the entropy left by knowing X when $W = 0$, the information gain provided by X is $H_{W=0} - H_{X/W=0} = 1 - 0 = 1$, meaning that the information gain provided by X reduces the uncertainty to zero when $W = 0$. To sum up, ID3 first selects W as the best attribute for having the greatest number of correct answers and then uses X to make all answers correct. In psychological terms, this means that subjects must always pay attention to W ’s features and that they must focus in particular on X whenever $W = 0$.

Note that ID3 does not learn the list of examples and their categories by rote, but rather induces a short definition of the concept. The list of examples in the *UNSPECIFIED* concept could be represented by a tree made with 8 paths (“IF $W = 0$ and $X = 0$ and $Y = 0$, THEN $Z = 0$ ”, for the first path, and so forth). However, the rule abstracted by ID3 only has three paths: “IF $W = 1$ THEN $Z = 1$ (path 1); IF $W = 0$ THEN [IF $X = 1$ THEN $Z = 1$ (path 2); IF $X = 0$, THEN $Z = 0$ (path 3)]”. This is exactly the same as inducing a short definition of a zebra from a full description of a list of zebras.

When trying to learn a rule for XOR classification, computing the information gain is possible but of no help. XOR is not compressible with such a technique. Every time a value of a variable is kept constant, there is 1 bit of uncertainty left. The minimal rule for XOR is “IF $X = 1$ THEN [IF $Y = 1$ THEN $Z = 0$; ELSE IF $Y = 0$ THEN $Z = 1$]; ELSE IF $X = 0$ THEN [...]”. Table 1 shows the decision tree for this rule. The decision tree being made of four paths, this is not really more parsimonious than rote learning of the list of examples and their categories. Dropping the tests for negative examples would simplify the tree, but would not change the complexity ranking of such a concept. This impossibility of compressing XOR tends to argue in favor of its high degree of complexity. Nevertheless, I will later show that computing mutual information still allows some simplification in XOR.

Likewise, the Type-VI concept –originally studied by Shepard et al. (1961)– is not compressible, because it is an

extension of *XOR*.

$$TYPE - VI = \begin{pmatrix} W & X & Y & Z \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Type-VI is made up of two inverted *XOR* structures (one when $W = 0$; the other when $W = 1$). In human learning, *Type-VI* is most often judged to be the most complex of the six 3D Boolean concepts that have an equal number of positive and negative examples (Feldman, 2000; Shepard et al., 1961; Nosofsky, Gluck, et al., 1994). Its complexity has also been observed in animal learning (J. D. Smith et al., 2004). The difficulty people encounter in learning this concept can be explained by the complete heterogeneity of the categories. For this concept, ID3 produces a decision tree of 8 paths, which is shown in Table 1 along with other simpler decision trees. Consistent with *XOR*, each time a variable is kept constant, there is 1 bit of uncertainty left. This is also true when two variables are held constant: for instance, when $W = 0$ and $X = 0$, Z is either 1 or 0, so Y is necessary to reduce the remaining bit of uncertainty.

To conclude, learning in ID3 amounts to compressing a training sample into a short formula. Following this idea, it has sometimes been hypothesized that the compressibility of a concept is a measure of its subjective complexity (Feldman, 2000; Mathy & Bradmetz, 2004; Lafond, Lacouture, & Mineau, 2007; Vigo, 2006). Note, however that reduction technique has many options that have been investigated in psychology and in artificial intelligence. In artificial intelligence, the debate most often concerns the optimality of compression algorithms, but debates in psychology focus on the reason for the non-optimality of the rules induced by individuals (Bradmetz & Mathy, 2008; Lafond et al., 2007). In a previous study, we developed a model that accounts for the fact that the compression of information in individuals results from strictly serial verbal rules (Mathy & Bradmetz, 2004; Bradmetz & Mathy, 2008) and the limited capacities in working memory, options that would certainly not be chosen in artificial intelligence.

The next section shows that when considering the various dimensions simultaneously, the learner might find some relationships that can considerably simplify the categorization process for these apparently difficult concepts.

Mutual information

Mutual information is a measure of the amount of

information one can obtain from a given set of variables by observing another variable or by observing the relationships between other variables. Before describing the main formula, note that it is possible to compute the joint entropy and the conditional entropy of variables. Joint entropy is simply the entropy of the set of messages that can be created using several variables. For instance, using two binary variables X and Y , it is possible to generate four different messages {00, 01, 10, 11}. The messages being equiprobable here, there is entropy of 2 bits in this set of messages. In this case, where the variables are independent,⁶ the joint entropy is simply the sum of the individual entropies: $H(X, Y) = H(X) + H(Y) = 2$. The conditional entropy $H(X/Y)$ (the slash indicates the conditional statement “given”) is a measure of the amount of information in one variable when another is held constant. For instance, we have $H(X/Y) = 1$ in the preceding set, because there is 1 bit of uncertainty left on X for any value of Y . These notions were implicit when we computed the information gain above (mainly to make the equations more readable).

Mutual information simply quantifies the relatedness of two or more variables. Mutual information corresponds to the reduction in the uncertainty about one variable due to the knowledge of another variable (see Fass, 2006; Garner, 1962; and Duda, Hart, & Stork, 2001, pp. 630-632). The mutual information of the two variables is:

$$I(X; Y) = H(X) - H(X|Y) \quad (2)$$

where:

$$H(X|Y) = H(X, Y) - H(Y). \quad (3)$$


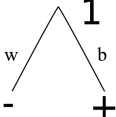
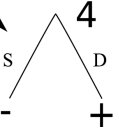




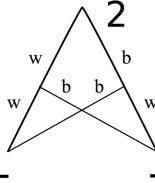




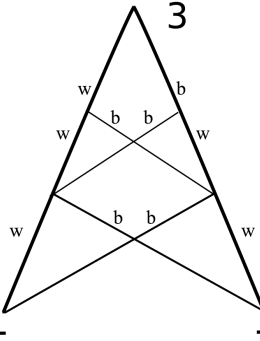
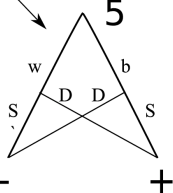
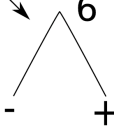







The variables used to denote the mutual information are separated by semicolons (e.g., $I(X; Y)$) to avoid confusion with the variables put in conjunction in the joint-entropy formula (e.g., $H(X, Y)$). For any pair of variables in the *XOR* truth table (e.g., X and Y), we get null mutual information. For instance: $I(X; Y) = H(X) - H(X|Y) = H(X) - (H(X, Y) - H(Y)) = 1 - (2 - 1) = 0$, meaning that these two variables are independent. By applying the same formula to any pair of variables, it can be shown that X , Y , and Z are independent.

Even if it looks much more complicated, the computation of mutual information is easily extendable to an arbitrary number of dimensions using alternating plus and minus signs over all subsets of variables. For three variables:

$$I(X; Y; Z) = -H(X) - H(Y) - H(Z) + H(X, Y) + H(X, Z) + H(Y, Z) - H(X, Y, Z) \quad (4)$$

Computed for the three variables in the *XOR* truth table, we have: $I(X; Y; Z) = -1 - 1 - 1 + 2 + 2 + 2 - 2 = 1$, which corresponds to the maximal amount of mutual information with three Boolean variables. This means that it is possible to get the value of a third variable by knowing the

Table 1. XOR and TYPE-VI concepts

	Stimuli	\mathcal{C}	Info. Gain	Mutual Info.	Num. Info.	
SIMPLE		-				
		-				
		+				
		+				
XOR		-				
		+				
		+				
		-				
TYPE VI		-				
		+				
		+				
		-				
		+				
		-				
		-				
		+				

Note. Each group of horizontally aligned balls represents a single stimulus. Column \mathcal{C} indicates the category membership. Mutual Info., mutual information. Num. Info, numerical information. Decision tree 1: “IF first ball is white, THEN -; IF first ball is black, THEN +”. Decision tree 2: “IF first ball is white, THEN [IF second ball is white, THEN -; IF second ball is black, THEN +]; IF first ball is black, THEN [IF second ball is black, THEN -; IF second ball is white, THEN +]”. Decision tree 3: similar to decision tree 2, except that the third level indicates the color of the third ball. Decision tree 4: “IF the color of the two balls is Same, THEN -; IF the color of the balls is Different, THEN +”. Decision tree 5: “IF first ball is white, THEN [IF the color of the other balls is Same, THEN -; IF the color of the other balls is Different, then +]; IF first ball is black, THEN [IF the color of the other balls is Different, THEN -; IF the color of the other balls is Same, then +]”. Decision tree 6: “IF the number of black balls is even, THEN -; ELSE +”.

relationship between the other two (or vice versa). Therefore, the category can be found by knowing the relationships between the two input values.

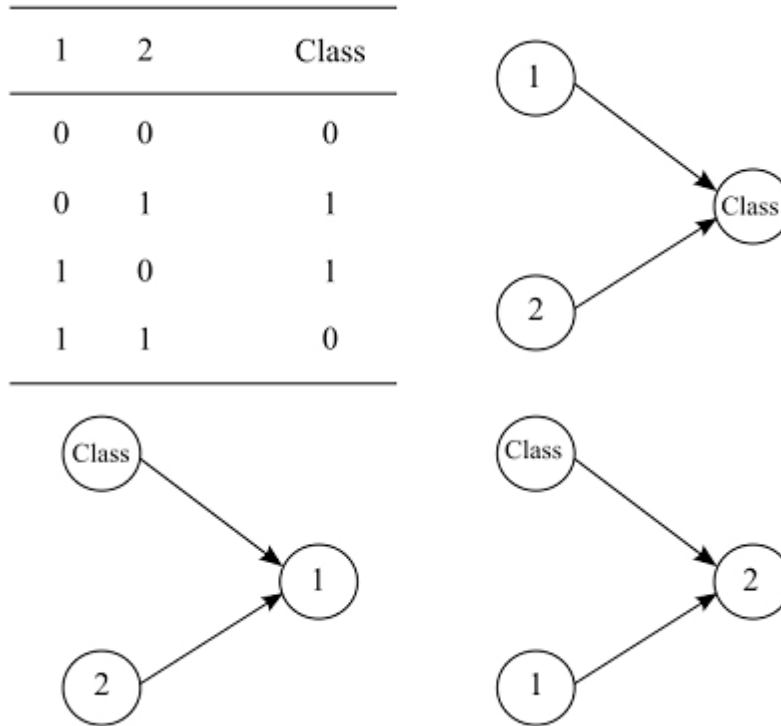
In Table 1, a set of stimuli was generated using combinations of black and white balls. Here, each group of horizontally aligned balls represents a single stimulus. I showed earlier that because no feature is characteristic of the class in this kind of concept, the learner must either learn a complex rule (computed, for instance, using information gain) or learn the examples and their respective categories by rote memory. However, by using mutual information, we can see that the stimuli are positive simply if the balls are of different colors, which is clearly more parsimonious. Here,

the features of one or the other ball are no longer important once one notices that the relation “different” between the two balls overrides the particular feature values.

This relational complexity can be demonstrated in a Bayesian network (Fig. 1), which indicates that variables 1 and 2 are two independent causes of the variable “Class” (see Glymour, 2001; Pearl, 2000).⁷ However, as specified in the network, the two variables are not independent, given the class. Clearly, knowing that the class is positive, it follows that variables 1 and 2 necessarily have different values. This is a case of a rather complex notion called conditional nonindependence.

Similarly, for the *Type-VI* concept, mutual information is

Figure 1. The exclusive disjunction and three possible corresponding Bayesian networks. Note. The truth table includes two input variables and one output variable. The output variable is called the class or the category in the categorization literature. The top right Bayesian network shows that input 1 and input 2 are independent. However, in this network, input 1 is NOT conditionally independent of input 2, GIVEN the class. Indeed, the value of input 2 is equal to the value of input 1 if the class is 0, whereas input 1 and input 2 are inversely correlated if the class is 1. Therefore, one can correctly classify the examples of an XOR by considering the relationships between the two input values. The same properties would follow if we permuted the variables in the network. In other words, if any two variables are the same, the third is equal to zero, whereas if any two variables are different, the third is equal to one.



also equal to 1. This means that it is possible to know the value of a given variable, given the relationship between the other three (or vice versa). Mutual information is maximal in *Type-VI* because there is an *XOR* structure with three variables for any value of the fourth variable. Therefore, the category can be determined by knowing the relationships between the three input values. Consistent with *XOR*, *Type-VI* apparently has no critical feature, thereby imposing rote memorization of the four positive examples. However, using mutual information, one can see that whenever the second and third balls are of different colors –when the first is white– or whenever they are of the same color –when the first is black– the example is positive. Such higher-order rules might facilitate learning of the *Type-VI* classification. The structure of information is even more intriguing in *Type-VI*: what was true for the first ball is also true for the second and third (for instance, if the third ball is white, the stimulus is positive whenever the other two balls are of different colors, and so on).

In sum, mutual information allows subjects to reduce the

complexity of the decision rules as follows: In *XOR*, subjects can use a two-path decision tree (IF balls are of different colors, THEN +; ELSE IF balls are the same color THEN -). In *Type-VI*, subjects can use a four-path decision tree (tree number 4 in Table 1: “IF first ball is white THEN [IF other balls are of different colors, THEN +; ELSE IF other balls are the same color THEN -]; ELSE IF first ball is black THEN [IF other balls are of different colors, THEN -; ELSE IF other balls are the same color THEN +”]. This rule can be reduced to “If first is white and other balls are of different colors, or if first is black and other balls are the same color” for positive examples. The diagonal arrows in Table 1 show that more compressed rules correspond to rules of a lower level.

It has been shown that correlations between features can be learned during classification tasks, even incidentally (Giguère, Lacroix, & Larochelle, 2007, although the subjects in this study only learned attributes that were perfectly correlated), but mutual information may be less apparent (Fass, 2006). For instance, when stimuli are more complex and have incommensurate dimensions (e.g., compound

stimuli such as big blue triangles, small red squares, etc.), mutual information can be available but needs to be used by subjects in a more complex manner than when the stimuli are like those shown in Table 1, where the relationships are obvious. For compound stimuli, mutual information can be used by reversing a subrule from one condition to another. For instance, a *Type-VI* concept can be simplified as follows: when the objects are small, the red squares and the blue triangles are positive, whereas when the objects are large, these objects are negative. The subject then has to memorize half of the decision paths and switch the leaves, given the size of the objects.

Necessarily, this assumes that the subject has already built a correct decision tree before noticing the symmetries, which explains why it has been shown that the utilization of mutual information in categorization tasks can only operate in the long run (see Mathy, In press). Finally, note that in Table 1, the stimuli are so peculiar that subjects can devise a much more abstract and efficient strategy using numerical information. For instance, by noticing that the stimuli are positive whenever the number of balls in the stimulus is odd, the subject can formulate a highly compressed two-path decision for both XOR and *Type-VI* concepts (trees 4 and 6 in Table 1). This strategy is also available when the stimuli are compound: knowing that the little blue triangle is a positive stimulus in a *Type-VI* structure, another stimulus is positive whenever a stimulus has only one feature in common with the little blue triangle. Again, subjects can reduce their decision rule to two paths (if there is one feature in common with the little blue triangle, then positive, else negative).

Conclusion

Decision-trees in concept learning

Entropy, information gain, and mutual information are basic notions in information theory. They can be used to model rule formation in concept-learning tasks, and to assess the subjective complexity of a given task and difficulty in classifying instances. The idea that learning involves the extraction of relevant information followed by gradual testing of hypotheses has considerable intuitive appeal. I suggest here that because rule formation in individuals functions by an information compression process (by starting small and seeking parsimony), individual performance can be measured by the compressibility of the information inherent in the conceptual structures, and compressibility itself can be expressed in terms of the minimal, ordered decision tree for a given concept. Similar modeling based on information reduction has been helpful in other domains, for instance, to explain

automatization in visual-memory search tasks via a process of information reduction achieved by ignoring features that are not diagnostic to the search (Cousineau & Larochelle, 2004).

Many hybrid models recognize the importance of specifying a mechanism for combining rule- and exemplar-based representations (Anderson & Betz, 2001; Erickson & Kruschke, 1998; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Rosseel, 2002; Pothos, 2005; E. E. Smith & Sloman, 1994). Indeed, the use of both processes is supported by recent advances in cognitive neuroscience (Ashby & Ell, 2001). There are also numerous other concurrent models of categorization –based on other paradigms– which also deserve serious consideration (Ashby & Maddox, 1993; Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Kruschke, 1992; Lamberts, 2000; Nosofsky, 1984; Nosofsky, Gluck, et al., 1994). In this paper, I advocate the development of rule-based models by focusing on putative relationships between subjective complexity and decision-tree complexity. This view is supported by recent experimental research in which learning times, error rates, and response times were used to assess subjective complexity. One possibility is that subjects need more time to discover and to form correct decision rules when the structure of the task-related information is more complex (Mathy & Bradmetz, 2004; Feldman, 2000; Lafond et al., 2007; Nosofsky, Palmeri, & McKinley, 1994). Another possibility is that the smallest number of steps required to reach a given leaf in a decision tree for a given stimulus is directly proportional to response time (Bradmetz & Mathy, 2008; Lafond et al., 2009; Trabasso, Rollins, & Shaughnessy, 1971). The principal difference between these recent studies and those of the last century is that (1) better characterization of reduction techniques has been sought to account for how subjects manage to learn⁸ (2) the plausibility of the information-reducing mechanisms and the tractability of the computations required during the reduction process have been better discussed, and (3) better measures of fit and broader experimental studies have been used to assess these different rule-based models.

Psychological plausibility

There are several classical methods of concept learning using rules based on information reduction. One might be reluctant to bring back to life an old-fashioned class of models based on simple decision trees (e.g., Hunt, Marin, & Stone, 1966) which have encountered many obstacles since their development. Firstly, such models often subsume verbal or consciously penetrable strategies, whereas recent research provides evidence of less-verbal, implicit processes in categorization in some cases. Secondly, these models may

appear quite crude in that they do not allow for fine tuning of parameters to achieve greater precision in modeling (a counter example is RULEX, which is presented below). A third reason is that certain predictions in rule-based models (e.g., assessing task complexity) can be easily mimicked by other models such as exemplar models, which are also very powerful in many other respects (Nosofsky, Gluck, et al., 1994). However, my goal was to show that perhaps more elaborate, rule-based models are worthy of serious consideration in psychology.

I will now discuss the plausibility of the notions presented in this paper and the optimality with which subjects abstract relevant information. I posit that subjects can compute relevant information and build a minimal decision rule using simple strategies, and that these strategies can be accounted for by computing information gain and mutual information. For subjects, computing information gain is equivalent to holding one dimension constant while globally assessing the number of correct responses obtained during the classification. Once subjects notice that they –statistically– get most responses if they hold a particular dimension constant (rather than some other dimension), they can start using this dimension systematically in order to separate the stimuli into two broad clusters. This process does not require much mental effort. The subject can simply try, for instance, to put all red pictures in Category A, and all blue ones in Category B. Assuming that most of their responses are correct, subjects select the dimension and then repeat the same strategy for the remaining incorrect classified instances. The dimensions are simply embedded by subjects until no uncertainty remains. It follows that the absence of trial-by-trial information for computing information gain is not necessarily a drawback for modeling subject strategies. Sayeki (1969) found, for instance, that subjects performed nearly optimally when required to ask as few questions as possible in order to identify one of the six classes assigned to 100 cards with classes and features that have different frequencies. All subjects used highly ordered efficient trees corresponding to optimal decision rules.⁹

Difficulty elaborating complex decisions about exceptions might explain why subjects struggle when trying to classify exceptions.¹⁰ Their difficulty may be caused by working memory limitations, which could prevent them from building or using an overly complex decision structure. In a previous paper, we argued that a decision-tree building process clearly depends on working memory capacity (Mathy & Bradmetz, 2004) (see also Lewandowsky, *in press*, who showed that working memory capacity is crucial to categorization). In the multi-agent model of working memory that we developed, there is no need for a

conductor (a central executive). Agents have very limited capacities for retaining information or organizing it, but they can nevertheless use basic communication processes to exchange information. A minimal decision tree can be built via an inter-agent communication process that promotes the elaboration of common knowledge (knowledge about stimuli and categories) from distributed knowledge (knowledge about features). The communication priorities simply depend on the information gain provided by the different agents. Our hypothesis was that the complexity of a decision tree can be determined by that of the multi-agent communication protocol. The number of agents that have to be held in working memory is limited simply because of the complexity of their interactions, which grows intractably with their number. This kind of gradual adaptation (i.e., agents, features are recruited one by one) captures the notions of learnability and the importance of starting small (Elman, 1993), in line with Gold's (1967) procedure of identification in the limit, the cascade-correlation algorithm for neural networks (Fahlman & Lebiere, 1990), the RULEX model (Nosofsky, Gluck, et al., 1994) and the SUSTAIN model of category learning (in which clusters are recruited gradually, Love, Medin, & Gureckis, 2004). These models are opposed to those involving pruning, which posit that learning progresses from the most specific and complex decisions to the simplest ones.

Extracting mutual information is also psychologically plausible, although it has been shown that subject's sensitivity to mutual information can be very limited (Fass, 2006). Mutual information, also called transinformation, measures the amount of information that can be obtained about one variable by observing other variables or relations between them. Other studies have shown that categorization learning can be guided by knowledge drawn from people's tacit understanding of causal relations, which explains why people can learn categories on the basis of feature correlations (Rehder & Hastie, 2001; Waldmann, Holyoak, & Fratianne, 1995; Waldmann, Meder, Sydow, & Hagmayer, *in press*). In some of the situations described in this paper, there are symmetries in two, apparently complex sets of decisions that can be used to form simpler rules. This can considerably simplify the learning process by halving the decision structure and resulting in performance gains. Mutual information is a non-metric tool enabling one to measure the complexity of relationships between features. Mutual information simply quantifies how certain features relate to each other. In various categorization models, Type-VI concepts are unanimously judged to be the most complex kind of 3D Boolean concepts. This has been largely confirmed by empirical data, but it is apparently inconsistent with the fact that this concept entails more

mutual information than any other 3D Boolean concept. Confirming this apparent paradox, I showed that subjects are able to gradually learn Type-VI concepts faster than other Shepardian concepts such as Type IV, which are supposedly less difficult than Type VI (Mathy, In press). Type-IV concepts have consistently been found to be easier than Type-VI ones (when the concepts are learned once). Using repeated measures, I pointed out that learning several classifications of the same type in succession has a substantial impact on how concepts are learned. By comparing Type-VI concepts (entailing a maximal amount of positive mutual information) with Type-IV concepts (entailing a minimal amount of positive mutual information), the results showed that Type-VI concepts gradually became more learnable than Type-IV ones (Shepard et al., 1961, made the same observation, but in a less controlled experimental setting). Mutual information hence emphasizes the peculiar status of Type VI and might also account for other Type-VI effects that have been noted in studies on inductive biases and cultural evolution (Griffiths, Christian, & Kalish, 2008). Another study showed that when initial learning pertains to a Type-II structure (XOR with a third irrelevant dimension, which also entails some mutual information in the two relevant dimensions), a reversal shift is easier than shifts based on a single, previously relevant dimension, or shifts based on a single, previously irrelevant dimension (Kruschke, 1996). Such an effect also tends to support the hypothesis that individuals can easily derive new rules based on reversed decisions. Because individuals can effectively use mutual information in the long run to devise easier strategies for category learning, categorization models should include every possible way for subjects to simplify a categorization task. Another example of cases where people easily reverse rules has been studied using the intra-extra dimensional set shift paradigm. When children have to switch from one sorting rule to another, they have more trouble when they need to change the relevant dimension (the classification is based on shape, then on color) than when they have to change the value of the relevant dimension (the classification is based on shape only, first with rabbits in category A and cars in category B, and then with rabbits in category B and cars in category A) (Perner & Lang, 2002).

A comparison with RULEX

There are always alternative ways of describing any category structure in terms of rules, logic formulae, or decision trees, and controversy certainly exists about exactly what form of abstraction is the one subjects are most likely to adopt (Mathy & Bradmetz, 2004; Feldman, 2000; Lafond et al., 2007). Still, there is agreement on the fact that subjects

perform more or less optimally in reducing the total amount of information. Overall, subjects have less trouble learning homogeneous categories that can be covered by simple rules, which can be tracked by fast and accurate learning. Learning in RULEX (rule-plus-exception learning model, Nosofsky, Palmeri, & McKinley, 1994) differs in several ways from the notions introduced in the present paper.

1. In RULEX, categorization decisions are made by sequentially verifying stored one-dimensional rules, conjunctive rules, and exceptions, whereas in a decision tree, there is a single verification process which tests the features ordered in a given decision tree. The distinction between rules and exceptions breaks down in a decision tree. Even though it is possible to consider that the difference between rules and exceptions merely depends on the number of feature tests, it is better to simply consider that one decision tree represents one rule, which can be simple or complex depending on its structure.

2. A more radical difference in the classification decision process is that RULEX first checks for all exceptions stored in memory, and if no exceptions apply, a check is made on simpler rules (going from the most specific to the most general). However, the sequence of hypothesis-testing stages is reversed: first there is a search for perfect single-dimension rules, imperfect single-dimension rules, and conjunctive rules, and then there is a search for exceptions if each of the other steps failed. In more classical decision trees, the classification decisions follow the tree-building process. First, simple dimensions are tested to induce a simple rule. If one dimension is not sufficient, other dimensions can be added to gain precision. Similarly, when a stimulus is presented, the decision process follows the same order, testing for the first dimension, then the second if the first dimension leads to uncertainty.

3. RULEX is quite ad hoc, since it is intended to be psychologically plausible. Because it allows for the tuning of numerous parameters, the space of predictions can contain a substantial number of possibilities, and predictions can only be found through time-consuming simulations. Navarro (2005) outlined a formalization of RULEX using basic probability theory and simple combinatorics to calculate its predictions faster. Still, he showed, for instance, that there can be 12 different patterns of results for ranking the six Shepardian concepts from Type I to Type VI. In contrast, a given decision-tree model predicts analytic expressions most often resulting in single patterns for ranking a given set of concepts. It is only by considering different decision-tree models that it is possible to obtain different patterns for ranking a given set of concepts. These patterns can then be compared to subjects' performance to gain insight into the mechanisms that underpin the categorization process. The

one-to-one function facilitates the rejection of a model, whereas in RULEX, the probability of accepting the model is greater. For instance, in two previous studies (Mathy & Bradmetz, 2004; Bradmetz & Mathy, 2008), we tried to show that the computation of entropy combined to a multi-agent model of categorization (which can be parameterized more or less in a sequential versus parallel way) can help determine the degree of optimality reached by individuals in learning decision rules. Given that the different parameterizations of the multi-agent model led to different rankings of complexity, we were able to determine that subjects tended to use information in an overly strict sequential manner, a process which slows down the decision process and sometimes make subjects form longer rules than necessary.

However, RULEX shares one interesting property with the decision-tree model presented in Bradmetz and Mathy (2008), also designed to be psychologically plausible. Both models predict that individuals vary in the particular rules they form and that averaged classification data are presumed to represent a mixture of idiosyncratic rules. For instance, let us imagine that *blue triangles*, *blue squares*, and *red triangles* are positive, whereas *red squares* are negative. The minimal formula is known to be *blue OR triangle*, which is isomorphic to a decision tree such as "IF *blue* then positive; IF *triangle* THEN positive" (Feldman, 2000; Lafond et al., 2007). The corresponding decision tree is *polythetic*, in that multiple attribute values can label each tree branch. In this case, the branch for categorizing the positive instances can be labelled *blue* or *triangle*. This minimal formula implies that the three positive objects require the same amount of computation to be categorized, so the response times are expected to be similar for those objects. The reason why most psychological models are not worried about polythetic decisions is certainly because psychological experiments do not use a large number of dimensions, but such trees are almost never used in artificial intelligence, for complexity reasons (Duda et al., 2001). In one of the multi-agent models we developed, which is based on the computation of information gain, the decision trees are *monothetic* because it is hypothesized that the order in which information is used in working memory is constant. In this case, given that the shape dimension and the color dimension are equally informative about the class, subjects could either induce the rule "IF *blue* THEN positive; IF *red* THEN [IF *triangle* THEN positive]", or the rule "IF *triangle* THEN positive; IF *square* THEN [IF *blue* THEN positive]". Again these decision trees are less optimal than polythetic ones because subjects are thought to use information in an overly strict sequential manner. Here, the *blue triangles* could be categorized using one piece of information by all the subjects, whereas the

other two positive instances were categorized by some subjects using a single piece of information and categorized by the rest of the subjects using two pieces of information (the mean response time for these two instances was therefore predicted to depend on 1.5 pieces of information). Our results showed that the time of access to the categories was related to these numbers of pieces of information. The *blue triangle* therefore acquired a prototype-like status, and overall, the mean classification pattern of the positive and negative instances was typical of the one predicted by the exemplar model. However, the prediction was made without relying on similarity as an explanatory principle.

Directions for future research

The present model proved useful both for accounting for subjective complexity across concept learning-tasks and for modeling response-time variability within concept-learning tasks. A couple of unsolved problems should be addressed in future research. Although attention allocation to dimensions can be inferred from the rules induced by subjects (since the most attractive dimension is generally chosen as the first dimension to test in a tree, whenever several dimensions are equally informative), decision-tree models would benefit from incorporating a typical form of dimension weighting to indicate the importance of the features. In addition, the absence of trial-by-trial information in the notions I presented poses a problem. Information gain and mutual information were computed here on whole sets of stimuli. Because the testing phase in concept-formation tasks appears to involve trial-by-trial tests (subjects seek evidence to test their hypothesis for every single presentation of a stimulus), a better approximation of subjects' performance could be targeted by defining how information is used trial by trial.

A critical problem for future research is integrating all possible strategies that subjects might use to facilitate the categorization task. I have shown here that despite the fact that mutual information is available to subjects, it can mostly only be used in the long term, once subjects have acquired expertise in the task. As the categorization process progresses, similarity can certainly also be used by subjects to exhaust any available strategy or to make faster decisions. Pruning could also be used once subjects notice analogies in the decisions, provided non-optimal decisions have been reached at one point. Overly monolithic models must therefore be proscribed in order to incorporate all possibilities available to subjects for efficient learning. One cannot obtain a real measure of the complexity of a concept until every possible strategy for reducing redundant information has been taken into account, assuming that subjects are as versatile as they can be.

References

- Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review*, 8(4), 629–647.
- Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, 105, 442–481.
- Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Science*, 5(5), 204–210.
- Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37(372–400).
- Attnave, F. (1959). *Applications of information theory to psychology: a summary of basic concepts, methods, and results*. New York, NY: Holt, Rinehart, and Winston.
- Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam's razor. *Information Processing Letters*, 24, 377–380.
- Boden, M. (1996). *Artificial intelligence*. San Diego, CA: Academic Press.
- Bradmetz, J., & Mathy, F. (2008). Response times seen as decomposition times in Boolean concept use. *Psychological Research*, 72, 211–234.
- Bruner, J., Goodnow, J., & Austin, G. (1956). *A study of thinking*. New York: Wiley.
- Coombs, C. H., Dawes, R. M., & Tversky, A. (Eds.). (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.
- Cousineau, D., & Larochelle, S. (2004). Visual-memory search: An integrative perspective. *Psychological Research*, 69, 77–105.
- Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. New York, NY: John Wiley and Sons.
- Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, 48, 71–99.
- Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, 127, 107–140.
- Fahlman, S. E., & Lebiere, C. (1990). *The cascade-correlation learning architecture* (Tech. Rep.). PA: Carnegie-Mellon University.
- Fass, D. (2006). *Human sensitivity to mutual information*. Unpublished doctoral dissertation, Rutgers University.
- Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, 407, 630–633.
- Feldman, J. (2003). A catalog of Boolean concepts. *Journal of Mathematical Psychology*, 47, 75–89.
- Garner, W. (1962). *Uncertainty and structure as psychological concepts*. New York: John Wiley and Sons.
- Garner, W. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.
- Giguère, G., Lacroix, G. L., & Larochelle, S. (2007). Learning the correlational structure of stimuli in a one-attribute classification task. *European Journal of Cognitive Psychology*, 19, 457–469.
- Glymour, C. N. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Massachusetts, MA: MIT Press.
- Gold, E. (1967). Language identification in the limit. *Information and Control*, 10, 447–474.
- Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123, 178–200.
- Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32, 108–154.
- Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, 32, 68–107.
- Hélie, S. (2006). An introduction to model selection: tools and algorithms. *Tutorials in Quantitative Methods for Psychology*, 2, 1–10.
- Hunt, E. B., Marin, J., & Stone, P. J. (1966). *Experiments in induction*. New York: Academic Press.
- Kruschke, J. K. (1992). Alcové: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22–44.
- Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, 8, 225–247.
- Lafond, D., Lacouture, Y., & Cohen, A. L. (2009). Decision-tree models of categorization response times, choice proportions, and typicality judgments. *Psychological Review*, 116, 833–855.
- Lafond, D., Lacouture, Y., & Mineau, G. (2007). Complexity minimization in rule-based category learning: Revising the catalog of boolean concepts and evidence for non-minimal rules. *Journal of Mathematical Psychology*, 51, 57–74.
- Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review*, 107, 227–260.
- Lewandowsky, S. (in press). Working memory capacity and categorization: Individual differences and models.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: A network model of category learning. *Psychological Review*, 111, 309–332.

- Luce, R. (2003). Whatever happened to information theory in psychology. *Review of general psychology*, 7, 183-188.
- Luger, G. (1994). *Cognitive science: the science of intelligent systems*. San Diego, CA: Academic Press.
- Mathy, F. (In press). The long term effect of relational information in Type-VI concepts.
- Mathy, F., & Bradmetz, J. (2004). A theory of the graceful complexification of concepts and their learnability. *Current Psychology of Cognition*, 22, 41-82.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.
- Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. Massachusetts, MA: Cambridge University Press.
- Mitchell, T. M. (1997). *Machine learning*. New York, NY: Mc Graw-Hill.
- Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
- Navarro, D. J. (2005). Analyzing the rule model of category learning. *Journal of Mathematical Psychology*, 49(4), 259-275.
- Neisser, U., & Weene, P. (1962). Hierarchies in concept attainment. *Journal of Experimental Psychology*, 64, 640-645.
- Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1), 104-114.
- Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Gauthier, P. (1994). Comparing models of rules-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, 22, 352-369.
- Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101, 53-79.
- Pearl, J. (2000). *Causality: models, reasoning, and inference*. New York, NY: Cambridge University Press.
- Perner, J., & Lang, B. (2002). What causes 3-year-olds' difficulty on the dimensional change card sorting task? *Infant and child development*, 11(2), 93-105.
- Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, 6(10), 421-425.
- Pothos, E. M. (2005). The rules versus similarity distinction. *Behav Brain Sci*, 28(1), 1-14.
- Quinlan, J. (1986). Induction of decision trees. *Machine learning*, 1, 81-106.
- Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of experimental psychology. General*, 130(3), 323-360.
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological Review*, 107, 358-367.
- Rossee, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, 46, 178-210.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations* (pp. 318-362). Cambridge, MA: MIT Press.
- Ruth, M., & Ryan, M. (2000). *Logic in computer science*. Cambridge, UK: University Press.
- Sayeki, Y. (1969). Information seeking for object identification. *Organizational Behavior and Human Performance*, 4, 267-283.
- Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379-423, 623-656.
- Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, 75, 13, whole No. 517.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3-22.
- Smith, E. E., & Sloman, S. A. (1994). Similarity- vs. rule-based categorization. *Memory & Cognition*, 22(4), 377-386.
- Smith, J. D., Minda, J. P., & Washburn, D. A. (2004). Category learning in rhesus monkeys: A study of the shepard, hovland, and jenkins (1961) tasks. *Journal of Experimental Psychology: General*, 133, 398-414.
- Trabasso, T., Rollins, H., & Shaughnessy, E. (1971). Storage and verification stages in processing concepts. *Cognitive Psychology*, 2, 239-289.
- Vigo, R. (2006). A note on the complexity of Boolean concepts. *Journal of Mathematical Psychology*, 50, 501-510.
- Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology. General*, 124(2).
- Waldmann, M. R., Meder, B., Sydow, M. von, & Hagmayer, Y. (in press). The tight coupling between category and causal learning. *Cognitive Processes*.
- Wood, C. (1976). Discriminability, response bias, and phoneme categories in discrimination of voice onset time. *Journal of the Acoustical Society of America*, 60, 1381-1389.

Manuscript received April 2nd, 2009.

Manuscript accepted March 30th, 2010.

Footnotes follow

¹ His argument was that the memory span is limited in the number of chunks (7) it can retain, but not in the number of bits, thus showing that short-term memory does not fit with a model of channel capacity. Luce (2003) gives an interesting historical account of the shift away from information theory in psychology after the 1960s.

² The expected value (i.e., the “mean”) for a random variable is $E(X) = \sum_i p_i x_i$, where i stands for all values that can be taken on by the random variable. For instance, if someone flips a coin to either win 2 dollars or lose 2 dollars, the expected value for the variable is $E(X) = (.5)(2) + (.5)(-2) = 0$.

³ Also, again, H corresponds to the number of binary questions one would need to ask to retrieve the content of a single message. For the categories, there would be three questions (rounding 2.8 to 3): Is the category less than or equal to 4? If so, Is the category less than or equal to 2? If so, Is the category equal to 1?

⁴ The study of when and what concepts are formed, and via what process at the ontogenetic or sociogenetic levels is called epistemology

⁵ Occam’s razor also applies to models (Hélie, 2006; Pitt & Myung, 2002; Roberts & Pashler, 2000), which should be as simple as possible, an argument that is often put forward by those who advocate rule-based models.

⁶ I am referring here to the notion of independence of probabilities, such as $p(X/Y) = p(X)$.

⁷ A peculiar property of *XOR* is that relational complexity is maximal, which means that the three variables can be permuted, and any of them can act as the class, because as long as the other two are different, the third is equal to one. As a result, three Bayes nets could be drawn from the truth table depicted in Fig. 1.

⁸ Different techniques such as C4.5, or C5 –also developed by Quinlan, for instance, to improve algorithms for continuous or missing data– or OBDDs have been developed in artificial intelligence, and some of them may have inspired psychologists (for an overview of such techniques, see Duda et al., 2001, Chap. 8, and Ruth & Ryan, 2000, Chap. 6). However, such adjunctions to ID3 are not so useful for the simple tasks described in this paper.

⁹ In Sayeki’s experiment, in which the categories were not equiprobable, the Shannon-Fano encoding theory was used to evaluate the optimality of the decision trees used by subjects. The Shannon-Fano encoding procedure is usually helpful when a set of symbols (e.g., the class labels) have different frequencies but no underlying defining features. However, the procedure may not be helpful when the objects are already encoded by a set of features. Sayeki experimented with objects whose features perfectly matched the codes produced by the Shannon-Fano algorithm. This gave the illusion that subjects could easily transpose the optimal set of codes to the set of features that characterized the objects. For instance, 68 objects labelled F could be recoded 0 while the other 32 objects could be recoded 1. Because all F objects were green and other objects were red (by construction), the decision tree for recoding the class labels according to frequency paralleled the feature codes (see Experiment 1, Deck 1, p. 272). However, this procedure is not helpful for the examples developed in this paper. For instance, in an *XOR* structure where a white square and a black circle are labelled A, and a black square and a white circle are labelled B, the Shannon-Fano encoding procedure would code both As as 0 and both Bs as 1, since the labels are equiprobable. However, such encoding would not be of any help when the learner needs to identify the class of an object given its features. In the *XOR* case, the minimal encoding of objects requires an average bit number of 2 bits per object, not 1. Whenever objects are already encoded by features, the computation of information gain is required.

¹⁰ Exemplar models simply state that exceptions are more difficult to handle because of their few similarities to the majority of objects belonging to the same category.