

Why Concept Learning is a Good Idea

Chris Thornton

Cognitive and Computing Sciences
University of Sussex
Brighton
BN1 9QH
UK

Email: Christopher.Thornton@firenet.uk.com

Tel: (44)1273 678856

May 21, 2003

Abstract

The paper provides an information theoretic account for similarity-based concept learning. It shows that the process can be understood as the attempt to minimise both the information content and equivocation (noise) of inputs from the environment.

1 Introduction

In general, an *implementation* of a concept c is any mechanism which partitions some set of possible data elements into two subsets: the subset covered by the concept and the subset not covered. Any mechanism which produces such black-boxes can be construed as a concept learner. The class of all such mechanisms is large. It encompasses concept learners proper, e.g. symbolist methods such as Candidate-Elimination [1, 2], Focussing [3,4], Classification [5] and Conceptual Clustering [6,Fisher, Learning from 7,8,9]. It also encompasses mechanisms which do not have concept learning as an explicit goal but which nevertheless produce the requisite black-boxes, e.g. connectionist mechanisms such as Back-Propagation, Competitive Learning [10] and numerical taxonomy methods such as hierarchical clustering [11,12].

In connectionist learning, the candidates for concept implementations are ‘feature detectors’. These produce high levels of output for certain classes of input so can be thought of as *probabilistic* concepts covering the classes in question. Clustering mechanisms produce dendrograms which can be thought of as disjunctive concepts in the manner of the decision trees produced by symbolist mechanisms such as ID3 [5,13].

The fact that so many computational mechanisms do (or can be seen as doing) concept learning seems to suggest that the process must be rather significant cognitively speaking. But what is its significance? Putting the question another way: What is the *point* of concept learning? Why is it a good idea? Some authors argue that concept learning is useful because it helps to make the world simpler and more easily comprehended. For example, Smith and Medin begin their book on ‘Categories and Concepts’ as follows.

Without concepts, mental life would be chaotic. If we perceived each entity as unique, we would be overwhelmed by the sheer diversity of what we experience and unable to remember more than a minute fraction of what we encounter. And if each individual entity needed a distinct name, our language would be staggeringly complex and communication virtually impossible. Fortunately, though, we do not perceive, remember, and talk about each object and event as unique, but rather as an instance of a class of concept that we already know something about [14].

This sort of argument makes a compelling appeal to intuition but it does not provide an answer in terms of any existing theories. In contrast, the present paper provides an *information theoretic* account of the processes of and effects achieved by concept learning. The account shows that concept learning can usually be viewed¹ in terms of a cognitive agent’s attempt to minimise both the information content and equivocation of messages (inputs) from the environment. Moreover, different outcomes in a learning process can usually be understood in terms of the discovery of locally optimal positions in a tradeoff between information-reduction and equivocation-increase.

1.1 Motivation

There are three main arguments which justify this work. They are as follows.

- Computational mechanisms for empirical concept learning are known to be quite limited in power. No existing mechanism will continue to acquire new concepts indefinitely. There will come a point after which it will cease to produce any worthwhile results. A clear information theoretic account for the process may shed new light on this limitation.
- Interest in concept learning extends across a number of different paradigms (e.g. symbolism, connectionism, statistics, psychology). Typically the results, ideas and methods associated with one paradigm are difficult to assimilate into the framework of another. The development of an information theoretic account might provide a framework via which to make cross-paradigm links.

¹The model developed is applicable only in the case where the computational abilities of the learner are subject to certain constraints.

- Many researchers in both the symbolist paradigm and the connectionist paradigm argue that there is a need to identify the information theoretic foundations for the computational methods developed in the cognitive sciences [15, 16]. This work is precisely an attempt to bring out one such foundation.

2 Similarity-based concept learning

A substantial proportion of mechanisms which generate (or can be seen as generating) concepts do so in a particular way. In machine learning the method is called *similarity-based learning* (SBL) [17]. Its central aim is to identify and define groupings of similar data elements. (These are normally called ‘descriptions’ in symbolist learning, ‘input vectors’ in connectionist learning, and ‘data points’ in numerical taxonomy.)

Let us take as an example the symbolist learning mechanism called Focussing. This is an algorithm which effectively manipulates the boundaries of two regions in the dataspace: an outer region enclosing all elements which are not known to be excluded from the concept and an inner region enclosing all elements which are known to be included in the concept.²

The algorithm attempts to expand the inner region and shrink the outer region until they are identical. The definition of the region forms the target concept; i.e. it forms a rule which generalises all positive instances but no negative instances [18]. The point to note is that the method is effectively a way of exploiting the fact that the positive instances of a concept will tend to be more *similar* to (and therefore closer in dataspace to) each other than to negative instances.

Connectionist mechanisms frequently employ the delta learning rule [10, Chapter 2]. This is a way of manipulating a weight-vector so as to ensure that the inner product of some input vector with the weight vector is maximised for some particular subset of inputs. If vectors are normalised then this operation effectively involves ‘defining’ groups of similar data elements in terms of their centroid point, cf. [10, Chapter 5]. In Competitive learning for example, the potential centroids (weight vectors) are moved around in dataspace until they lie at the centre of groups of similar elements. Again, the point to note is that the method is a way of exploiting the notion that the instances covered by a concept will tend to be more similar to each other than to negative instances.

Many more examples could be provided here. Suffice it to say that there is a large class of mechanisms which produce, either as a main result or as a side-effect, computational components which can be classified as concepts — and that they do so using a technique which effectively seeks to capture groups of *similar* data elements. The effective goal of this sort of mechanism is to produce concepts which maximise intra-class similarity and minimise inter-class similarity, cf. [19].

²The algorithm actually manipulates marks in generalisation trees.

3 A generic model

The ultimate aim of the paper is to provide an *information theoretic* explanation for similarity-based (concept) learning. To do this we need to solve various problems and answer a number of questions. An initial problem is the fact that the information theoretic (IT) paradigm is somewhat at odds with the similarity-based learning (SBL) paradigm in its view of the world. Information theory decomposes the world into entities such as ‘source’, ‘receiver’, ‘channel’ whereas most SBL models assume a decomposition in terms of ‘learner’, ‘environment’ etc. Our approach will be to set up a generic model which can be reconciled with both viewpoints. The model involves the following entities.

- K - a cognitive agent
- U - a universe
- D - a data or input language (a set of possible data elements)
- M - a total set of inputs generated by some universe
- f - a differentiation function.

In an instantiation of the model, K is a particular cognitive agent in a particular universe U. U is characterised purely in terms of D, M and f, with M being a subset of D. Two elements of D may be more or less distinct: their differentiation is given explicitly by the function f (whose range is assumed to be some sequence of values). Note that U may be a continuous or discrete universe. Its continuous properties are captured in f. If U is discrete, f always returns one of two values—indicating ‘same’ or ‘different’.

We can use the model to map the entities which are of significance in the information theoretic view of the world into the entities which are of significance in similarity-based concept learning. In information theoretic terms, K is the ‘receiver’ and U is the ‘source’. A message passing from source to receiver corresponds to the appearance of a new input in K. In SBL terms, K is the learner or classifier, D is the description language and M is the source of the training instances. In connectionist terms, K is a network, D is the data or input language, M is the set of potential training vectors.

4 Prediction and information

The framework allows us to pose a number of questions. For example, given some particular instantiation of the model (involving a given K and a given U), we might ask how much knowledge K has about U. In general we expect that the more knowledge an agent has about some universe the better the agent can predict that universe. In terms of the model, predicting U means being able to accurately identify M. But what does *this* mean?

In the case where U is perfectly discrete, it means enumerating M ; or, in general, assigning high probabilities to all members of M and relatively low probabilities to all members of $D - M$. In the case where U is not necessarily discrete, it means that for every element m in M , K identifies a relatively high probability to an element of D which is very similar to m .

In information theory we find this relationship between knowledge and predictive ability turned on its head. The situation in which K assigns relatively high probabilities to real inputs (i.e. members of M) from U is viewed as a situation in which K receives relatively little information from U . This is a complementary rather than a contradictory view of the situation. Our assumption is that being able to predict U means being able to assign relatively high probabilities to members of M . But if every real input is assigned a relatively high probability then every real input contains little surprise for K and the information content of inputs is therefore low. Hence, being able to predict U means receiving little information from U , and vice versa.

5 Optimising predictive accuracy means minimising information

Given the remarks above, we can see that there is a relationship between acquiring knowledge about U and reducing the information content of messages from U . In particular, it is clear that K can *reduce* input information by acquiring knowledge about U . But can K acquire knowledge of U by attempting to reduce input information?

Let us consider the special case where K is initially completely knowledgeableless. The information content of each new input to K is

$$- \log_2 p$$

where p is the probability assigned to the input by K . To minimise the average amount of information content of new inputs it is necessary to assign the highest possible probability to all members of M . This means assigning high probabilities to the members of the set M . But in the absence of any knowledge about U , there is no way to decide which members of D are in M . Therefore, all members must be assigned the same probability value. Moreover, this value must be related to the size of D . If n is K 's estimate of the size of M , then K must assign a probability of

$$n / |D|$$

to each element of D in order to minimise the information content of new inputs [20,21]. In this context, the information content of each new input is

$$- \log_2 (n / |D|)$$

6 Forwarded messages

Does the situation change if we allow K to process the received inputs in some way? In particular, is there any way in which K can process new inputs so as to further reduce their information content? We can think of the situation in which K processes inputs in terms of the application of some function to the inputs received at any given point. But there are two quite different cases to be considered. K might be able to process inputs one at a time or K might be able to process n inputs all in one go. Putting it more precisely, K might be able to apply a 1-place ('context-free') function to input data. Alternatively, K might be able to apply a n -place ('context-sensitive') function to input data.³

The first case here—involving the 1-place function—is the simpler one; it can be given a fairly straightforward informational account. Our approach will be to think of the 1-place function as a kind of black-box which accepts inputs direct from U and 'forwards' them to K. This allows us to work out what the information content of the *outputs* of f will be in given situations.

Let us denote the set of all possible outputs as D' . In the case where f implements a genuine many-to-1 mapping (where there is at least one group of inputs which are mapped onto the same output), the size of D' is less than the size of D . But if $|D'| < |D|$ then K is justified in assigning higher probabilities to members of D' than to members of D . K therefore potentially receives less information from forwarded inputs than from direct inputs.

But this seems a little absurd. If K can reduce the amount of information received from U simply by mapping unique inputs onto non-unique outputs then surely all that has to be done to minimise the information received (and therefore maximise predictive ability) is to map all inputs onto *one* output. This would mean that $|D'| = 1$; the probability of the single output would be 1 and the amount of information received would always be zero (implying that K has perfect knowledge of U). Well, obviously, there is a catch. And it is called *noise*.

7 Noise and equivocation

If K maps all inputs onto one output, that output tells K absolutely nothing about what the input was: it is completely ambiguous. On the other hand, if K maps almost all inputs onto unique outputs, then a given output can give a completely accurate indication of what the input was. In reducing the absolute number of outputs K reduces their information content but also, inevitably, increases their ambiguity. In information theoretic terms, this introduction of ambiguity is understood in terms of the introduction of noise, or more precisely, *equivocation* [22].

In the case where inputs are perfectly discrete entities, information theory states that the ambiguity can be quantified in terms of the conditional entropy

³We assume that in either case the function is deterministic; i.e. that it always produces the same output for the same input. It may, of course, produce a unique output for unique inputs; alternatively, some inputs may evoke the same output.

of the outputs. This is just the entropy of the set of probability values derived when we work out what is the probability — given a particular input — of the output taking on each of the entire range of possible values.

In the continuous case, equivocation is a more complex beast. However, we can think of it as working just like discrete equivocation provided we imagine that the space of possible outputs is divided up into cells such that the conditional probability of getting any output in the cell for some given input is the same. Intuitively, we can think of the equivocation in the continuous case as the average perturbation or inaccuracy of messages (inputs) from U [22, Chapter 4].

8 Minimising equivocation

Clearly, by mapping unique inputs onto non-unique outputs K reduces their information content but pays a price in terms of increased equivocation. Ideally, K must find some way of minimising both the information content of inputs *and* their equivocation. To minimise input information it is necessary to make D' as small as possible; i.e. to map as many inputs as possible onto the same output. To minimise the equivocation of outputs, K must minimise their ambiguity; i.e. must minimise the degree to which messages are perturbed as they are forwarded (via the 1-place function).

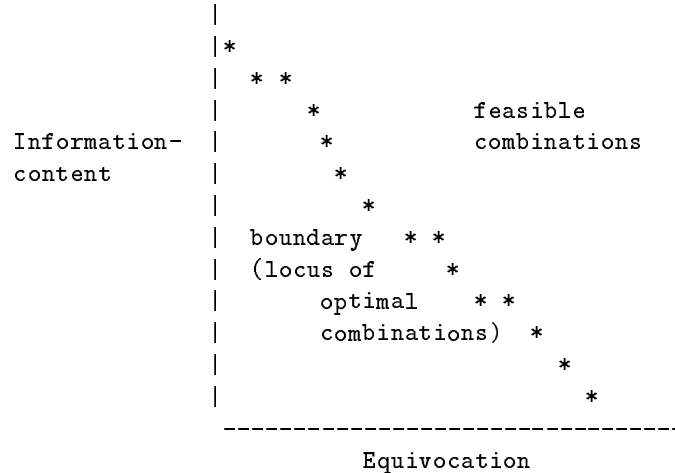
To achieve this minimisation it is only necessary to ensure that any set of unique inputs which are mapped onto the same output are as similar (according to the differentiation function f) as possible and that the output is as similar to all of the inputs as possible. Provided the similarity of any set of inputs evoking a unique output is maximised (for any given level of information-reduction), the average perturbation (inaccuracy) of inputs from U is minimised too.

9 The information/equivocation tradeoff

K , then, is confronted with a *tradeoff* between information and equivocation. K can reduce the information content of inputs but only by allowing their equivocation (ambiguity) to increase. K can reduce equivocation but only by allowing the information content of inputs to increase. For some fixed reduction in input information, there is some *minimum* price that K must pay in terms of equivocation. The price is kept to a minimum only if maximally similar inputs are mapped onto unique outputs. For some fixed equivocation, there is some maximum reduction in input information which can be achieved. Again, this is only achieved if maximally similar inputs evoke unique outputs from f . Thus, there is a set of optimal ways of combining information-reduction and equivocation-increase.

We can visualise the situation in terms of a 2-dimensional graph in which information content is plotted on the vertical axis and equivocation is plotted on the horizontal axis; see Figure 1. A given point corresponds to a partic-

ular information/equivocation combination. Points corresponding to feasible information/equivocation combinations form a region whose boundary represents the locus of optimal information/equivocation combinations; i.e. the set of points for which the ratio of information reduction to equivocation increase is maximised.



10 Is equivocation such a bad thing?

Recall that our initial notion was that by minimising the information content of inputs, K becomes better able to make accurate predictions about U and therefore, in some sense, acquires knowledge about U . We have seen that K can reduce input information by allowing equivocation to increase and also that the information-reduction/equivocation ratio is maximised in the case where the average pairwise similarity for input-groups (i.e. set of inputs mapped onto given outputs) is maximised. But what does all this mean for predictive accuracy?

As we might expect, maximising the information-reduction/equivocation ratio effectively optimises predictive accuracy. In the case where inputs are forwarded via the many-to-1 mapping, being able to predict U implies assigning high probabilities to some subset of elements of D which are, collectively, relatively similar to M . Since we are assuming that K is knowledgeable, we know that assigned probabilities will always be the same. Therefore, predictive accuracy will be related only to the average similarity between inputs and their corresponding outputs.

Now, in minimising equivocation at some fixed level of information-reduction, we minimise both input-group differences and the average difference between an arbitrary input-group member and the corresponding output. Thus we necessarily optimise predictive accuracy. We effectively ensure that for any arbitrary input, there is an element of D — to which K assigns an equal probability — which is as similar to the input as possible.

11 SBL = knowledge maximisation

We can summarise the points made above thus. In the special case where K has no prior knowledge of U and is only allowed to process inputs using a 1-place (context-free) function, K maximises the accuracy of predictions about U (and therefore optimises knowledge about U) by maximising the information-reduction/equivocation ratio. This is done by identifying groups of maximally similar inputs and mapping them onto maximally representative outputs (i.e. outputs which are as similar to members of the input group as possible).

Now, as was noted above, similarity-based learning is a process which attempts to map groups of data elements onto concepts in such a way as to maximise intra-group similarity and minimise inter-group similarity. This is precisely the optimal course of action for any agent who is (1) subject to the constraints mentioned above who (2) attempts to maximise predictive accuracy for an arbitrary universe. The identification of groups of different grain-size — a common property of SBL mechanisms [17] — corresponds to the identification of particular positions along the information-reduction/equivocation tradeoff.

12 Concluding comments

The paper has argued that the process known as similarity-based learning in symbolist work — which also plays a role in connectionist work and in numerical taxonomy — can be understood as an attempt on the part of a cognitive agent to maximise predictive accuracy in the training domain. More precisely, it has shown that similarity-based learning maximises predictive accuracy (and an information-reduction/equivocation ratio) in any agent who can only process inputs using a context-free (i.e. 1-place) function.

The question of how predictive accuracy might be maximised in an agent who can process inputs using *arbitrary* context-sensitive (i.e. n-place) functions has not been addressed. Indeed, there are reasons for thinking that there will be no simple answer to this, much more general, question. If n-place functions can be applied to inputs then the agent is, in principle, able to detect arbitrary types of relationship between inputs and, therefore, abstract patterns in the data. In a worst-case scenario, agents would exploit each different type of abstract pattern (for the purposes of making predictions) in a different way. Thus there would be no single answer to the question of how an agent might exploit context-free input processing, just a myriad of special cases. How closely this worst-case scenario approximates to reality is an important issue for further research.

References

- [1] Mitchell, T. (1977). Version spaces: a candidate elimination approach to rule learning. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* (pp. 305-310).

- [2] Mitchell, T. (1982). Generalization as search. *Artificial Intelligence*, 18 (pp. 203-226).
- [3] Young, R., Plotkin, G. and Linz, R. (1977). Analysis of an extended concept-learning task. In R. Eddy (Ed.), *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* (p. 285).
- [4] Bundy, A., Silver, B. and Plummer, D. (1985). An analytical comparison of some rule-learning programs. *Artificial Intelligence*, 27, No. 2 (pp. 137-81).
- [5] Quinlan, J. (1983). Learning efficient classification procedures and their application to chess end games. In R. Michalski, J. Carbonell and T. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*. Palo Alto: Tioga.
- [6] Fisher, D. and Langley, P. (1985). Approaches to conceptual clustering. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence: Vol II* (pp. 691-697). Los Altos: Morgan Kaufmann.
- [7] Fisher, D. (1987). Conceptual clustering, learning from examples and inference. *Proceedings of the Fourth International Workshop on Machine Learning* (June 22-25 University of California, Irvine) (pp. 38-49). Los Altos: Morgan Kaufmann.
- [8] Michalski, R. and Stepp, R. (1983). Learning from observation: conceptual clustering. In R. Michalski, J. Carbonell and T. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*. Palo Alto: Tioga.
- [9] Stepp, R. and Michalski, R. (1986). Conceptual clustering: inventing goal-oriented classifications of structured objects. In R. Michalski, J. Carbonell and T. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach: Vol II*. Los Altos: Morgan Kaufmann.
- [10] Rumelhart, D., McClelland, J. and the PDP Research Group, (Eds.) (1986). *Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Vols I and II*. Cambridge, Mass.: MIT Press.
- [11] Romesburg, H. (1984). *Cluster Analysis for Researchers*. London: Wadsworth.
- [12] Anderberg, M. (1973). *Cluster Analysis for Applications*. New York: Academic Press.
- [13] Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1 (pp. 81-106).
- [14] Smith, E. and Medin, D. (1981). *Categories and Concepts*. Cambridge, Mass.: Harvard University Press.
- [15] Wolff, G. (1989). Information and redundancy in computing and cognition. *AISB Quarterly*, No. 68 (pp. 14-17).

- [16] Thornton, C. (1988). Links between content and information-content. *Proceedings of the Eighth European Conference on Artificial Intelligence* (Munich, 1-5 August). Pitman (also available as CSRP 109, School of Cognitive Sciences, University of Sussex, 1988).
- [17] Kodratoff, Y. (1988). *Introduction to Machine Learning*. London: Pitman.
- [18] Thornton, C. (1987). Hypercuboid-formation behaviour of two learning algorithms. CSRP 067, University of Sussex (Extended version of paper in IJCAI-87).
- [19] Gluck, M. and Corter, J. (1985). Information, uncertainty, and the utility of categories. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 283-287). Irvine, California: Lawrence Erlbaum Associates.
- [20] Brillouin, L. (1962). *Science and Information Theory*. New York: Academic Press.
- [21] Baierlein, R. (1971). *Atoms and Information Theory*. San Francisco: W.H. Freeman.
- [22] Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Information*. Urbana: University of Illinois Press.