Focusing Attention for Observational Learning: The Importance of Context

Joel Martin*
Georgia Institute of Technology, Atlanta, GA 30332-0280
joel@gatech.edu

Abstract

A significant component of human observational learning is the ability to focus attention toward important or relevant input features. Amechanism with this capability can serve as an inductive bias to facilitate learning in both humans and machines. Past attempts to model attentional focus for human learning have postulated a single salience value for each feature, such that features with greater salience command more attention. These models, however, assume that the feature's salience is not dependent on context, whereas studies of human attention show sensitivity to context. This paper presents a mechanism for contextually focused attention in observational learning.

1 Introduction

Observational learning is a form of inductive knowledge acquisition in which there is no external guidance, such as explicit feedback. However, some guidance or learning bias is required to make general induction tractable (eg. Rendell et al., 1987). Since humans do engage in some observational learning (Billman et al., 1987), there must be a method for *internally* guiding this learning. Discovering such methods will prove useful both for understanding human learning and for designing computer programs that learn from observation (eg. Fisher, 1987). Zeaman and House (19C3) and Billman and licit (1988) have argued that attention directed by learnable feature saliences may provide some internal guidance for human learners. They each proposed a mechanism for doing this and were able to confirm the approach for simple learning. Neither method used *context* — what already is known about an example — to help focus attention. Other researchers, however, have found that human attention and other cognitive processes vary with context (Loftus & Mackworth, 1978; Barsalou & Medin, 1986). As well, the use of context can allow learning of more complex examples. The noncontext approach assumes that there is only one important subset of features that are always salient. When this assumption is violated, learning is not facilitated.

Given that attention is useful for human and machine observational learning, it is important to ensure that proposed attentional mechanisms support a useful type of learning. Human observational learning most clearly is useful for natural language and concept based predictions. Additionally, many machine learning studies of observational learning have concentrated on concept based prediction (Schlimmer, 1986; Fisher, 1987). These types of knowledge have frequently been described as capturing correlational feature structure (Rosch, 1978; Medin & Schaffer, 1983; Fisher, 1987). In other words, category structure and linguistic structure can be represented partially by correlational rules or conditional probabilities of the form: *P(feature} — valuel | feature'2 = value2).* Thus, a "rule" such as, *recovering = feathers \ locomotion = wings),* records the frequency with which 'feathers[1] occur given that 'wings' is true. Anderson (1988) has further argued that even if human category structure is not implemented using conditional probabilities, it and other phenomena are best described and explained by such probabilities. Similarly, recent machine learning models of concept acquisition have proposed that categories can be best learned by maintaining conditional probabilities (Schlimmer, 1986; Fisher, 1987). These psychological and machine learning studies suggest that an adequate model of human and machine attentional learning should demonstrate how the attention mechanism can facilitate learning of conditional probabilities or estimates thereof.

This paper presents a new model of the use of attention for observational learning. This model, called Contextually Focused Sampling, introduces a context controlled attention mechanism, and is proposed as a method for both human and machine observational learning. This use of context partially was motivated by the need for dynamic learning biases (eg. Rendell et al., 1987) and machine learning studies of the use of probabilistic context for generalization (eg. Fisher, 1987). CFS is compared to an important non-context alternative, Focused Sampling (Billman & Heit, 1988), to demonstrate its similar behavior for simple learning and its superior behavior for more complex learning in

*The author wishes to thank Dorrit Billman for providing constructive comments on all aspects of this research, and Janet Kolodner an anonymous reviewer for assistance with earlier versions of the manuscript. This research was supported by the Army Research Institute for the Behavioral and Social Sciences under Contract *No.* MDA-903-86-C-173.

which there are multiple important subsets of features.

2 Focused Sampling

Billman and Heit's (1988) CARI implementation of the Focused Sampling (FS) method describes how attention can be used to facilitate observational learning. Put simply, FS allocates more attention to those features that participate in strong correlations or rules. The 'rules' and 'correlations' referred to by Billman and Heit are simply the conditional probability relationships between features.

1: Choose two features, Fl and F2. The probability of choosing any given feature is that featured salience divided by the sum of the all saliences (Luce, 1959).

2: Sample Fl: Observe the value, v1, for Fl.

3: Given vl, predict the value for F2. The probability of predicting a value, v2, is equal to the probability from vl to v2 divided by the sum of all rule strengths between vl and values of F2.

4: Sample F2: Observe the value, v2, for F2.

5: - If predicted matches the true value, increment the rule strength from vl to v2 and increment the salience of Fl and F2.
- If they do not match, decrement rule strength and saliences.

Figure 1: Focused Sampling Algorithm

In CARI (Figure 1), two features, such as color and size, are sampled, and a prediction of the value of the second feature is made on the basis of the value for the first feature. All training examples are assumed to be collections of feature/value pairs, and sampling a feature reveals that feature's value. For example, the color feature, when sampled, might be found to have the value 'green'. If the prediction of the second value is correct, then the saliences of the features and the strength of the prediction are incremented. Otherwise, these values are decremented. The adjustment of the values is based on an estimator of conditional probabilities called the delta rule. This estimator, in similar forms, has been used in many psychological learning models (Rescorla, 1972; Rumelhart, Hinton, & Williams, 1986). CARI updates the rule strength and feature saliences by,

$$S_n = S_{n-1} + a[T - S_{n-1})$$

S_n = Salience or Strength; α = learning rate
T = 1, if prediction is correct; T = 0, otherwise.

FS is an attentional learning mechanism that supports learning of correlational structure (Billman et al., 1987; Billman & Heit, 1988); and there are two major learning behaviors of the model that any viable alternative must also demonstrate.

- First, FS produces a facilitation in learning as compared to random sampling of features (Billman & Heit, 1988);
- Second, particular rules are learned faster when they are part of a system of interrelated rules than when they occur in isolation. Billman and her colleagues term this effect *clustered feature facilitation.* Human subjects have demonstrated this effect for observational learning of a novel language (Billman et al., 1987).

3 Contextually Focused Sampling

There are many reasons to suspect that context is important for attention. First, humans are able to use information that they already know about an example to direct their attention to unusual aspects of the same example (Loftus & Mack worth, 1978). Second, some multiple-look attention models (Trabasso &; Bower, 1968) suggest an averaging method for using what is known about an example when generating a response. Finally, algorithms like FS would not allow a human or machine learner to focus on different cohesive subparts of an example. For instance, there are many subsets of animal features that internally cohere, like food-type and size or habitat and means-of-locomotion. FS, though, assumes that there is only one important subset. An alternative model will be proposed that introduces a limited form of context for feature sampling. The use of the word 'context' in this work refers specifically to known feature values of a particular example. Using context for attention therefore refers to using those feature values that have already been observed to help choose other features to which to attend.

The method proposed by this paper, Contextually Focused Sampling (CFS), samples attributes based on their estimated predictability. It calculates those estimates using estimates of conditional probability between feature values. In CFS (Figure 2) then, choosing a feature depends upon that feature's predictability given what values are already known. This method allows the probability of sampling a particular feature to vary with the context.

The CFS algorithm, like the FS, uses the delta rule to update the estimates of conditional probabilities and no-context feature saliences. The no-context feature saliences are used for sampling when nothing is yet known about an example. Feature saliences *in* context are based solely on the estimates of the conditional probabilities. An important difference between CFS and FS is that CFS allows multiple samples to be taken from each example in order to provide context.

CFS requires an algorithm for estimating predictiveness in context, i.e., when several features have already been sampled. It is not reasonable to maintain all such higher order probabilities, because there are exponentially many of them. The most straightforward alternative is to use a Bayesian estimate assuming independence (Martin, 1988). However, pilot studies have shown that an arithmetic average (Trabasso fc Bower, 1968) is better correlated to actual higher order conditional probabilities for the types of training example being used. Davis

(1985) gives an argument for the use of a geometric average in a similar machine learning system.

1: Choose starting feature, Fl. Probability of sampling is featured no-context salience divided by sum of all no-context saliences.

2: Sample Fl: Observe the value, vl, for Fl.

Loop for i = 1 to n

3: Choose feature, Fi, based upon an estimate of probability of F's values given known values.

4: Using an estimate of joint conditional probability, predict the value of Fi.

5: Sample Fi: Observe the value, vi, for Fi.

6: - If predicted value matches vi, increment strengths between all previously sampled values and vi. If i=2, increment saliences of Fl and Fi.
- If the predicted value does not match, decrement rule strengths. If i=2, decrement the saliences.

Figure 2: Contextually Focused Sampling

4 Experimental Tests of CFS

CFS was compared to FS in three experiments. The first two experiments were performed to demonstrate that CFS is a viable alternative to FS. Experiment III was conducted to determine whether CFS is superior to FS for more complex inputs.

4.1 Algorithms

The experiments performed comparisons between three algorithms, Random Sampling (RS), FS, and CFS. In general, it is difficult to compare algorithms because they often differ by more than one characteristic. For instance, CFS and FS differ not only by how a feature is sampled but also by how many features are processed per example. FS samples exactly two features, while CFS can sample several. These extraneous differences can confound a comparison on the characteristic of interest. It is therefore important to remove as many extraneous differences as possible before comparisons are made. The FS algorithm was modified to incorporate the loop from the CFS algorithm. The only difference between the FS and CFS algorithms was that the former always used salience to select features for sampling. The RS algorithm was like the CFS algorithm, except that it selected features independently of salience and estimates of conditional probabilities. These versions of the RS, FS, and CFS algorithms were used in all experiments.

4.2 General Method

The general method used for all three experiments was very similar to that used by Billman and Heit (1988). The input was provided as lists of digits in the form, (1 2 3 2), to represent that the features 1 through 4 have the values 1, 2, 3, 2 respectively. These number vectors are used for simplicity but are meant to represent vectors such as, (covering=fur, habitat=land, size=big, locomotion=legs). In all three experiments, the inputs consisted of eight features (Figure 3). These inputs were presented one at a time as examples. Each trial consisted of presenting one example that was selected randomly from all available inputs. These trials were divided into blocks of 50. As in Billman and Heit (1988), the strengths between the values of the first two features were averaged to measure learning after each block of trials. In all sets of training examples, the first two features were related strongly. The test strengths were, {f1 = 1 —> f_2 = l,f_2 = 1 - f1 = 1,fi = 2 ->f_2 = 2,f_2 = 2 -> f_1 = 2}. Statistical comparisons were made based on this average target strength after a criterion number of trial blocks. The criterion was set for each experiment when the mean strength for IIS was equal to 0.50 ± 0.02, as in Billman and licit (1988).

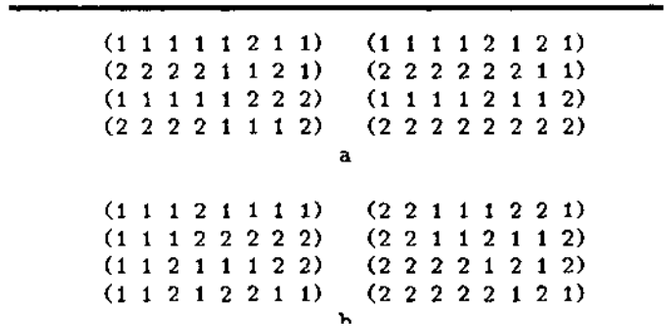

Figure 3: Input Vectors.

The variable parameters were set to the values used by Billman & Heit (1988) and were held constant for all experiments. The initial strength values were set to 0.01, initial feature saliences were set to 0.125, and delta learning rates were set to 0.02. CFS and the modified FS and RS algorithms have one additional parameter, the number of features sampled per example. Pilot studies demonstrated that as this parameter is increased, learning facilitation increases. This parameter was set at 3 samples per example for all algorithms throughout the experiments to reflect the limited capacity of attention and to achieve some benefit of context for CFS.

Fifteen simulated subjects were run in each condition. These subjects varied due to probabilistic sampling and random example selection.

5 Experiment I & II

CFS should be able to demonstrate the significant behaviors of Focused Sampling. The first of the two important FS behaviors is a facilitation of learning as compared to random sampling. In Experiment I, CFS was predicted to produce a learning facilitation because, like FS, CFS's focusing mechanism leads it away from irrelevant features.

CFS should also show increasing facilitation for a larger set of interrelated features. In Experiment II, CFS is predicted to show clustered feature facilitation because the CFS sampling is biased toward features that are interpredictive.

5.1 Method

In experiment I, the subjects received the inputs presented in Figure 3a. These inputs were chosen to maximize FS benefit by interrelating half the features (Billman & Heit, 1988, experiment 3). The remaining four features were termed irrelevant because they each are related randomly to every other feature. Experiment I compared the different attention methods, random sampling, FS, and CFS.

The method for Experiment II was the same as for Experiment I, except that both sets of inputs from Figure 3a and Figure 3b were used. Fifteen simulated CFS subjects received inputs with two clustered features and 15 received inputs with four clustered features. The learning measure used for each set of inputs was the difference in learning between using the CFS and RS algorithms.

5.2 Results

The learning rates depicted in Figure 4 show clear effects of attention method.

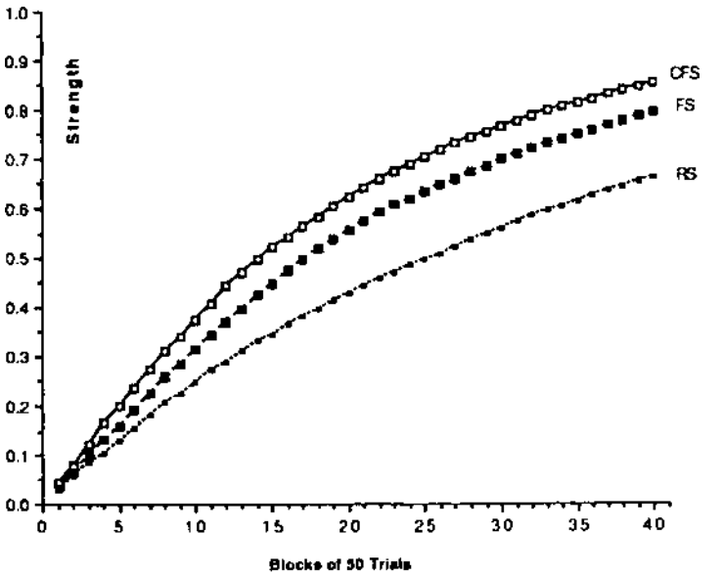

Figure 4: Learning with Different Attention Methods.

An ANOVA was performed using the strength values at the criterion number of trials as defined above. Attention method showed a significant effect, $F(2,42) = 33.66, p < 0.01$. Tukey's USD was used to compare means for the attention method to determine significant differences: $HSD(3, 42) = 0.079$, $p < 0.01$. The comparisons revealed that both CFS and PS showed facilitation over the random method. Although CFS produced a greater facilitation than FS, this difference was not significant. These data demonstrate that CFS produces the same type of learning facilitation as FS.

In Experiment II, there was a greater facilitation for clusters of four features rather than two. A t-test was performed at the learning criterion for the four clustered feature condition, $£(28) = 2.73, p < 0.01$. The test showed that, as with FS and in accord with human data, CFS demonstrated clustered feature facilitation.

6 Experiment III

As predicted, CFS produces two findings which motivated the FS model. It was assumed that because it was sensitive to context, CFS would predict greater learning facilitation for more complex inputs than would FS. Both Experiment I and II and the experiments of Billman and Heit (1988) have used inputs in which there is only one important cluster of features. That is, there are some relevant features and some irrelevant features, and all relevant features are intercorrelated. However, more realistic inputs would allow for multiple clusters of relevant features. For example, in humans, hair-color and eye-color are somewhat intercorrelated as are arm-length and height. All four of these features are relevant to feature clusters, but are not *all* intercorrelated.

Context is important for attentional learning in domains with multiple clusters because it allows the human or machine learner to concentrate on a single subcluster at a time. The non-context approach used by FS would set the saliences of the features independently of the cluster, permitting sampling across clusters. For example, if hair-color is the most salient and height the second most salient then the most frequent sampling pair would be across clusters. CFS can help alleviate this problem, because it allows the feature saliences to vary depending on what has already been sampled. In the above example, after hair-color is sampled, then the most salient feature becomes eye-color. The saliences in CFS are modified to have the learning focus on one cluster at a time.

Because of these considerations, it was predicted that CFS would be found to be superior to FS when there were multiple unrelated clusters of features in the input. As well, FS was expected to have a decreased learning facilitation as compared to random sampling.

6.1 Method

The method was the same as for Experiment 1 and II, except that the inputs had three clusters of three features each and three irrelevant features. The three clusters of features were independent of each other. All three algorithms were compared on these inputs. The learning measure, as in Experiment 1, was the average strength of the rules relating features one and two; and statistical comparisons were made at the criterion number of trials.

G.2 Results

Figure 5 shows the learning curves for each block of 100 trials. There was an increased facilitation for CFS over FS and RS.

An ANOVA was performed that indicated a significant difference between algorithms: $F(2,42) = 31.21, p <$

O.OL Tukey's HSD, *HSD(3,*42) = 0.082, p < 0.01, revealed a significant difference between CFS and both other groups. FS was not found to be significantly different from random sampling.

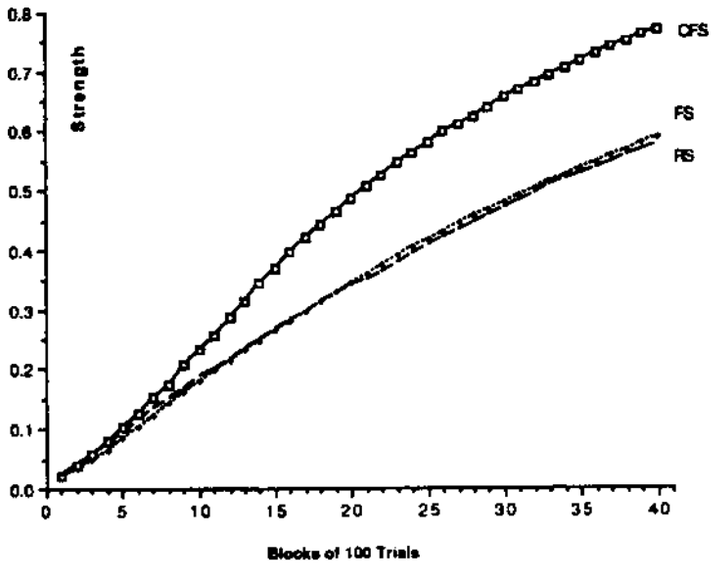

Figure 5: Learning for Examples with Multiple Clusters.

As predicted, CFS was superior to FS and RS for inputs with multiple clusters. As well, the multiple clusters prevented a significant learning facilitation for FS over RS.

7 Discussion

Attention can serve as an important learning bias for learning by observation. The type of attention mechanism used, however, should be sensitive to context if the training examples are complex, i.e., have multiple clusters. Contextually Focused Sampling (CFS) was proposed to be a better match to human attentional processes than earlier models that were not sensitive to context (Billman & Heit, 1988). CFS is also an important candidate for one type of attentional bias in machine learning systems.

These computational experiments suggest several interesting predictions about human learning. First, if humans use either FS or CFS, then they must show faster learning than random sampling (RS) would permit. Second, if humans use CFS and not FS, they should demonstrate faster learning of multiple clusters than either FS or RS would permit. Finally, CFS would predict that humans show different probabilities of sampling particular features depending upon what they have already sampled.

For machine learning, the results imply that CFS can be used to make induction more feasible. A learning bias is some restriction or ordering on what can be learned, and a good bias is one that allows faster learning. CFS represents one simple type of bias that gradually comes to ignore certain features that are irrelevant and thereby accelerates learning of informative conditional probabilities. One important aspect of CFS is that if irrelevant features become relevant, then those features gradually will come to be sampled more and more often.

Also of interest for machine learning, the results of Experiment III suggest how CFS might be used to divide a set of training examples into appropriate categories of interpredictive features. Because CFS is capable of finding and learning about multiple subclusters of interrelated features, it can provide a method for constructing a hierarchy of probabilistic concepts. The use of focused attention to easily isolate these concepts may result in an algorithm that is more efficient and more powerful than current concept learning methods (Fisher, 1987). Such concept acquisition also would allow a CFS system to learn higher order conditional probabilities to improve its inference capabilities (Chalnick & Billman, 1988; Davis, 1985).

An important extension of contextually focussed sampling is to augment the attribute value lists with structure, such as predicate-style relationships between values. A naive approach would be to maintain conditional probabilities between values and other values, values and relationships, and relationships and relationships. This, however, results in an unreasonable growth in the size of storage. Future research must determine how optimal context dependent attention can be approximated for structured knowledge without resorting to complete interconnectivity.

References

[Anderson, 1988] J. R. Anderson. The place of cognitive architectures'in a rational analysis. In *Proceedings of Tenth Annual Conference of the Cognitive Science Society,* Montreal, Canada, 1988.

[Billman and Heit, 1988] D. Billman and E. Heit. Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science,* 12:587-625, 1988.

[Billman *et al,* 1987] D. Billman, E. Heit, and J. Dorfman. Facilitation from clustered features: Using correlations in observational learning. In *Proceedings of Ninth Annual Conference of the Cognitive Science Society,* Seattle, Washington, 1987.

[Chalnick and Billman, 1988] A. Chalnick and D. Billman. Unsupervised learning of correlational structure. In *Proceedings of Tenth Annual Conference of the Cognitive Science Society,* Montreal, Canada, 1988.

[Davis, 1985] B. R. Davis. An associative hierarchical self-organizing system. *IEEE Transactions on Systems, Man, and Cybernetics,* 15:570-579, 1985.

[Fisher, 1987] D. Fisher. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2:139-172, 1987.

[Loftusand Mackworth, 1978] G. It. Loftus and N. H. Mackworth. Cognitive determinants of fixation location during picture viewing. Journal of Experimental Psychology: Human Performance and Perception, 4:565-572, 1978.

[Martin, 1988] J. D. Martin. Cora: A best match memory for case storage and retrieval. In Proceedings of AAAI Case-Based Reasoning Workshop, St. Paul, Minnesota, 1988.

[Medin and Schaffer, 1978] D. L. Medin and M. M. Schaffer. A context theory of classification learning. Psychological Review, 85:207 238, 1978.

[Rendell et al, 1987]
L. Rendell, R. Seshu, and D. Tcheng. More robust concept learning using dynamically-variable bias. In Proceedings of the Fourth International Workshop on Machine Learning, UC Irvine, 1987.

[Rescorla, 1972] It. A. Rescorla. Informational variables in pavlovian conditioning. In G. II. Bower, editor, The Psychology of Learning and Motivation. Academic Press, New York, 1972.

[Rosch, 1978] E. H. Rosch. Principles of categorization. In E. II. Rosch and B. B. Lloyd, editors, Cognition and Categorization. Erlbaum, Hillsdale, N.I, 1978.

[Rumelhart et al, 1986] I). E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propogation. In D. E. Rumelhart and J. L. McClelland, editors, Parallel Distributed Processing, Vol 1. MIT Press, Cambridge, Massachusetts, 1986.*

[Schlimmer, 1987] J. C. Schlimmer. Incremental adjustment of representations for learning. In Proceedings of the Fourth International Workshop on Machine Learning, UC Irvine, 1987.

[Trabasso and Bower, 1968] T. Trabasso and G. II. Bower. Attention in Learning. Wiley, New York, 1968.

[Zeaman and House, 1963] D. Zeaman and B. J. House. The role of attention in retardate discrimination learning. In N. R. Ellis, editor, Handbook of Mental Deficiency. McGraw Hill, New York, 1963.