Modeling Unsupervised Learning with SUSTAIN

Todd M. Gureckis and Bradley C. Love
{gureckis, love}@love.psy.utexas.edu
Department of Psychology - MEZ 330
University of Texas at Austin
Austin, TX 78712 USA

Abstract

SUSTAIN (Supervised and Unsupervised STratified Adaptive Incremental Network) is a network model of human category learning. This paper extends SUSTAIN so that it can be used to model unsupervised learning data. A modified recruitment mechanism is introduced that creates new conceptual clusters in response to *surprising* events during learning. Two seemingly contradictory unsupervised learning data sets are modeled using this new recruitment method. In addition, the feasibility of using a unified recruitment method for both supervised and unsupervised learning is discussed.

Introduction

The process of learning categories from examples can take many forms. Sometimes learning is supervised and explicit feedback directs category formation. Other times learning is unsupervised and no explicit feedback is available from the environment. For example, we are commonly asked to categorize incoming email as belonging to the "junk mail" category or to the "interesting mail" category. We are not explicitly taught to identify members of the either category and we do not receive specific feedback on each example. Nevertheless, we acquire and use categories to sort our mail on a daily basis.

Traditionally, researchers interested in categorization have focused on modeling human performance in supervised learning tasks. This may be motivated in part by the additional constraints that feedback can play in the design of an experiment. However, given the pervasiveness of unsupervised learning in our daily life, there is potentially quite a bit to gain from expanding our understanding of this type of learning.

This paper presents a model of human category learning called SUSTAIN (Supervised and Unsupervised STratified Adaptive Incremental Network). SUSTAIN has been successfully applied to an array of challenging human data sets spanning a variety of category learning paradigms including supervised learning and inference learning (Love, Markman, & Yamauchi 2000; Love & Medin 1998).

This work was supported by AFOSR Grant F49620-01-1-0295.

This paper will specifically address how SUSTAIN can be modified to model human performance in unsupervised learning tasks. We will begin our discussion with an overview of SUSTAIN which serves to highlight some of the important features of the model and introduces the motivation for the later sections. Next, we discuss the challenges of modeling unsupervised learning and explore how SUSTAIN can be modified to use a flexible and intuitive notion of *surprise* as a cluster recruitment method. We fit a version of SUSTAIN that uses this generalized recruitment method to a series of unsupervised learning data sets. Finally, we evaluate the prospect of using this new recruitment rule for both unsupervised and supervised learning.

An Overview of SUSTAIN

Before discussing the issues involved in modeling unsupervised learning with SUSTAIN, we will present an overview of the operation of SUSTAIN and discuss some of the major principles and psychological motivations of the model.

SUSTAIN is a clustering model of human category learning. The model takes as input a set of perceptual features that are organized into a series of independent feature dimensions. Like other models of category learning (e.g. Kruschke, 1992), SUSTAIN maintains an attentional tuning mechanism which allows it to selectively weight stimulus feature dimensions. During the process of learning, SUSTAIN updates these attentional weights to place emphasis on stimulus dimensions that are most useful for categorization.

The internal representations in the model consist of a set of clusters. Categories are represented in the model as one or more associated clusters. Initially, the network only has only one cluster that is centered upon the first input pattern. As new stimulus items are presented, the model attempts to assign new items to an existing cluster. This assignment is done through an unsupervised procedure based on the similarity of the new item to the stored clusters. When a new item is assigned to a cluster, this cluster updates its internal representation to become the average of all items assigned to the cluster so far. However, if SUSTAIN discovers through feedback that this similarity based assignment is incorrect, a new cluster is created to encode the exception. Classification decisions are ultimately based on the cluster to which an instance is assigned.

Principles of SUSTAIN

With this general understanding of the operation of the model, we now examine what we consider to be the five key principles of SUSTAIN.

Principle 1, SUSTAIN is biased towards simple solutions SUSTAIN is initially directed towards simple solutions. At the start of learning, SUSTAIN has only one cluster which is centered on the first input item. It then adds clusters (i.e., complexity) only as needed to accurately describe the category structure. Its selective attention mechanism further serves to bias SUSTAIN towards simple solutions by focusing the model on the stimulus dimensions that provide consistent information.

Principle 2, similar stimulus items tend to cluster together In learning to classify stimuli as members of two distinct categories, SUSTAIN will cluster similar items together. For example, different instances of a bird subtype (e.g., sparrows) could cluster together and form a sparrow cluster instead of leaving separate traces in memory for each instance. Clustering is an unsupervised process because cluster assignment is done on the basis of similarity, not feedback.

Principle 3, SUSTAIN learns in both a supervised and unsupervised fashion In learning to classify the categories "birds" and "mammals", SUSTAIN relies on both unsupervised and supervised learning processes. Consider a learning trial in which SUSTAIN has formed a cluster whose members are small birds, and another cluster whose members are four-legged mammals. If SUSTAIN is subsequently asked to classify a bat, it will initially predict that a bat is a bird on the basis of overall similarity (bats and birds are both small, have wings, fly, etc.). Upon receiving feedback from the environment (supervision) indicating that a bat is a mammal, SUSTAIN will recruit a new cluster to represent the bat as an exception to the mammal category. The next time SUSTAIN is exposed to the bat or another similar bat, SUSTAIN will correctly predict that a bat is a mammal. This example also illustrates how SUSTAIN can entertain more complex solutions when necessary through cluster recruitment (see Principle 1).

Principle 4, the pattern of feedback matters As the example used above illustrates, feedback affects the inferred category structure. Prediction failures result in a cluster being recruited, thus different patterns of feedback can lead to different representations being acquired. This principle allows SUSTAIN to predict different acquisition patterns for different learning modes (e.g., inference versus classification learning) that are informationally equivalent but differ in their pattern of feedback.

Principle 5, cluster competition Clusters can be seen as competing explanations of the input. The strength of the response from the winning cluster (the cluster the current stimulus is most similar to) is attenuated in the presence of other clusters that are somewhat similar to the current stimulus (see Sloman's, 1997, account of competing explanations in reasoning).

SUSTAIN and Unsupervised Learning

In the formulation of SUSTAIN described above, the network adapts its architecture in response to external feedback. Only when SUSTAIN predicts an incorrect response does it recruit a new cluster to capture the exception. We might say that SUSTAIN changes its architecture in response to a *surprising* event, which in this case is a misclassified item. Unfortunately, this recruitment rule leaves SUSTAIN unable to model unsupervised learning data. In unsupervised learning, there is no feedback and we assume that each stimulus item is a member of the same category (the global category). This disables SUSTAIN's supervised recruitment process because prediction errors do not occur.

This deficiency necessitates a modification to SUSTAIN to accommodate unsupervised learning. While it is possible to model unsupervised and supervised learning with separate recruitment procedures (Love, Medin, & Gureckis 2002), a unification of the two procedures provides a more parsimonious account. We propose that a more general notion of surprise may be sufficient to model both unsupervised and supervised learning.

A More General Notion of Surprise

What criteria should be used to create a new cluster in an unsupervised learning task? A reasonable approach might be to store items in memory that are not sufficiently similar to existing clusters. This notion of surprise can be extended to supervised learning situations by creating a new cluster when the most similar cluster predicting the *correct* response (or category) is not sufficiently activated. The two schemes for supervised and unsupervised learning are actually one in the same because we assume that unsupervised learning involves only one nominal category (thus each cluster always predicts the correct response).

This recruitment strategy has a number of virtues over the traditional recruitment rule for supervised learning. For example, with the prediction-error based recruitment rule, if a novel and unusual stimulus item is encountered that marginally activates an existing cluster, a new cluster may not be recruited even when the item is very dissimilar to the winning cluster (having only weakly activated it). The winning cluster might then undergo catastrophic change by shifting too far away from its current position in order to accommodate the unusual item.

Unified Recruitment Rule

In the cluster recruitment strategy described above, a new cluster is created when the current input pattern is not sufficiently similar to existing prototypes and exceptions. In particular, a new unit is recruited when the activation of the winning cluster is below a fixed threshold. A simple unsupervised recruitment equation displaying the requisite characteristics is:

$$\text{if } (A_{H_j} < \tau), \text{ then recruit a new cluster} \qquad (1)$$

where A_{H_j} is the activation of the most highly activated cluster that belongs to the same category as the current input

stimulus and τ is a constant between 0 and 1 (a parameter). We would like to stress that in unsupervised learning, all items belong to the same category, thus A_{H_j} refers to the most activated cluster overall. In supervised learning, the most activated cluster predicting the correct category may not in fact be the most activated cluster overall.

Modeling Unsupervised Learning with SUSTAIN

To evaluate SUSTAIN's promise in the domain of unsupervised learning, we provide results of SUSTAIN's application to Experiments 2 and 3 from Billman and Knutson's (1996) unsupervised learning study and to unsupervised category construction (i.e., sorting) data from Medin, Wattenmaker, and Hampson (1987).

Modeling Billman and Knutson's (1996)

Billman and Knutson's experiments tested the prediction that category learning is easier when certain stimulus attributes are predictive of other attributes by way of a correlation (e.g., "has wings", "can fly", "has feathers" are all correlated features of birds). Their studies evaluate how relations among stimulus attributes affect learning in an unsupervised task.

Experiment 2 Experiment 2 consisted of an Isolating and a Structured condition. Stimulus items in both conditions depicted imaginary animals that were made up of seven attributes: type of head, body, texture, tail, legs, habitat, and time of day pictured. Each attribute could take on one of three values. For example, the time of day could be "sunrise", "nighttime", or "midday".

Training items in the Isolating condition preserved only one pairwise correlation between stimulus attributes. All of the stimulus items thus conformed to one of the following patterns: 11XXXXX, 22XXXXX, or 33XXXXX (a 'X' means that the dimension was free to assume any of the three possible values). If the first stimulus dimension encoded the head of the animal and the second stimulus dimension encoded the body, then knowledge about the type of head an animal possessed would allow prediction of what type of body it had and vice versa. The remaining five dimensions were not correlated so that they were not useful for prediction.

Data items in the Structured condition had six of these pairwise correlations. The first four dimensions of these items were constrained to vary together like the first two dimensions in the isolating condition (e.g, 1111XXX, 2222XXX, or 3333XXX). Since four dimensions were involved and because the correlations were interrelated, there were 6 pairwise correlations in the training items for the Structured condition (e.g., cor(A,B), cor(A,C), cor(A,D), cor(B,C), cor(B, D), cor(C, D)).

In the learning phase for both conditions, subjects were told that they were participating in a visual memory experiment and viewed the stimulus items for four blocks (four passes through all of the training items). Each item appeared once per block in a random order.

In the test phase of the experiment, subjects viewed a novel set of 45 stimulus item pairs. Each member of the pair had two obscured attribute values (e.g., the locations where the tail and head should have been were blacked out) so that in the Structured condition information about only one correlation was available from each test item. The purpose of blocking dimensions was to query learning on only one correlation at a time.

Subjects were asked to evaluate the remaining five attributes that were visible and to choose the stimulus item in the pair that seemed most similar to the items studied in the learning phase (a forced choice procedure). One of the test items was considered the "correct" test item because it preserved the correlations present in the items viewed during the study phase and the other was considered "incorrect" because it did not preserve the correlations.

The basic result from Experiment 2 was that the "correct" item was chosen more often in the Structured condition than in the Isolating condition (73% vs. 62%). This finding supports the hypothesis that extracting a category's structure is facilitated by intercorrelated dimensions.

Experiment 3 An alternative explanation of the results from Experiment 2 is that a larger number of pairwise correlations in the Structured condition (relative to the Isolating condition) facilitated learning. To test this explanation, the number of pairwise correlations in the Isolating and Structured conditions were equated in Experiment 3.

In the Isolating condition, the items had three isolated pairwise correlations. The abstract structure of the items constrained the first six dimensions into three orthogonal pairs of correlated dimensions. Example stimulus items had the following structure: 112233X, 113322X, 111122X, etc...

Items in the Structured condition had three interrelated correlations. The first three dimensions were correlated which created three pairwise correlations (e.g. cor(A,B), cor(B,C), cor(A,C)). Thus, the number of pairwise correlations in the Isolating and Structured condition was equal, but the relationship between these pairs was varied between conditions.

Experiment 3 used the same training and test procedure as Experiment 2. The basic result from Experiment 3 is that the "correct" item was chosen more often in the Structured condition than in the Isolating condition (77% vs. 66%).

Modeling Results SUSTAIN was trained in a manner analogous to how subjects were trained by using four randomly ordered learning blocks. No feedback was provided and all stimulus items were encoded as being members of the same category. New clusters were recruited according to the new recruitment rule. In order for SUSTAIN to mimic the forced choice nature of the test phase, a response probability was calculated for each of the two items. The ultimate response of the network was towards the item in the forced choice that had the strongest response probability.

SUSTAIN was run numerous times on both conditions in both experiments and the results were averaged (see Table 2). The best fitting parameters for both Experiment 2 and 3 (one set of parameters was used to model both studies) are shown in Table 1 under the unsupervised column. For both

Table 1: SUSTAIN's best fitting parameters for the studies considered.

function/adjusts	symbol	unsupervised	six types
learning rate	η	0.0966	0.0923
cluster competition	β	6.40	0.25
decision consistency	d	1.98	16.9
attentional focus	r	10.0	9.01
threshold	τ	0.5	0.7

Table 2: The mean accuracy for humans and SUSTAIN (shown in parentheses) for Billman and Knutson's (1996) Experiment 2 and 3.

	Isolating	Structured
Experiment 2	.62 (0.67)	.73 (0.79)
Experiment 3	.66 (0.60)	.77 (0.77)

experiments, SUSTAIN correctly predicts greater accuracy in the Structured condition than in the Isolating condition (see Table 2).

In both Experiment 2 and Experiment 3, SUSTAIN's most common solution in the Isolating condition was to partition the studied items into three clusters. However, in Experiment 3, the nature of the three partitions varied across runs. SUSTAIN tended to focus on one of three correlations present in the Isolated condition and ignored the other two. For instance, during training SUSTAIN might create three clusters organized around the first two input dimensions (one cluster for each correlated value across the two dimensions) and largely ignore the correlation between the third and fourth dimensions and the fifth and sixth dimensions.

The same dynamics that lead SUSTAIN to focus on only one correlation in the Isolating condition leads SUSTAIN to focus on all of the interrelated correlations in the Structured conditions in Experiments 2 and 3. When SUSTAIN learns one correlation in the Structured condition, SUSTAIN necessarily learns all of the pairwise correlations because of the way cluster weights are updated (i.e., a prototype is formed). This type of learning facilitation is what lead to the higher accuracy levels.

SUSTAIN's solution to Experiment 3 suggests some novel predictions: 1) Learning about a correlation is more likely to make learning about another correlation more difficult when the correlations are not interrelated. 2) When correlations are interrelated, either all of the correlations are learned or none of the correlations are learned. Both of these predictions are currently under investigation.

Modeling Sorting Behavior with SUSTAIN

Billman and Knutson's (1996) studies found that subjects preferred stimulus organizations in which the perceptual dimensions were intercorrelated. Interestingly, category construction studies reveal a contrasting pattern — subjects tend to sort stimuli along a single dimension. This behavior persists despite the fact that alternate organizations exist that respect the intercorrelated nature of the stimuli (Medin, Wattenmaker, & Michalski 1987).

Table 3: The logical structure of the perceptual dimensions in Medin et al. (1987) sorted according to family resemblance.

Category A	Category B
1 1 1 1	2 2 2 2
1 1 1 2	2 2 2 1
1 1 2 1	2 2 1 2
1 2 1 1	2 1 2 2
2 1 1 1	1 2 2 2

SUSTAIN was applied to the sorting data from Medin et al.'s (1987) Experiment 1 in hopes of reconciling the apparently contradictory findings. In Experiment 1, subjects were instructed to sort stimuli into two equal sized piles. Stimuli were cartoon-like animals that varied on four binary-valued perceptual dimensions (head shape, number of legs, body markings, and tail length). The logical structure of the items is shown in Table 3. The basic finding is that subjects sort along a single dimension as opposed to sorting stimuli according to their intercorrelated structure (i.e., the family resemblance structure shown in Table 3).

When SUSTAIN was applied to the stimulus set from Experiment 1 it was constrained to create only two piles (i.e., clusters) like Medin et al.'s subjects. This was accomplished by not allowing SUSTAIN to recruit a third cluster. This modification proved to be unnecessary as an unmodified version of SUSTAIN recruited two clusters in 99% of simulations. SUSTAIN was presented with the items from Table 3 for 10 random training blocks to mirror subjects' examination of the stimulus set and their ruminations as to how to organize the stimuli. To evaluate the performance of the model, we looked at how SUSTAIN's two clusters were organized. Using the same parameters that were used in the Billman and Knutson (1996) studies listed in Table 1, SUSTAIN correctly predicted that the majority of sorts (99%) will be organized along one stimulus dimension.

SUSTAIN's natural bias to focus on a subset of stimulus dimensions (which is further stressed by the selective attention mechanism) led it to predict the predominance of unidimensional sorts. Attention is directed towards stimulus dimensions that consistently match at the cluster level. This leads to certain dimensions becoming more salient over the course of learning. The dimension that develops the greatest salience over the course of learning becomes the basis for the unidimensional sort.

Supervised Learning

An interesting challenge for models of category learning is to provide a unified account of both unsupervised and supervised data without modification. A model which can accurately account for supervised and unsupervised learning within a single framework may provide considerable insight into the subtle relationships between the two types of learning. Anderson's rational model (Anderson 1991) is able to operate in both a unsupervised and supervised fashion

Table 4: The logical structure of the six classification problems tested in Shepard et al. (1961) is shown.

Stimulus	I	II	III	IV	V	VI
1 1 1	**A**	**A**	**B**	**B**	**B**	**B**
1 1 2	**A**	**A**	**B**	**B**	**B**	**A**
1 2 1	**A**	**B**	**B**	**B**	**B**	**A**
1 2 2	**A**	**B**	**A**	**A**	**A**	**B**
2 1 1	**B**	**B**	**A**	**B**	**A**	**A**
2 1 2	**B**	**B**	**B**	**A**	**A**	**B**
2 2 1	**B**	**A**	**A**	**A**	**A**	**B**
2 2 2	**B**	**A**	**A**	**A**	**B**	**A**

within a unified framework. However, the rational model is unable to account for some of the human learning data sets that SUSTAIN has successfully modeled (see Love, Medin, & Gureckis, 2002). The cluster recruitment rule introduced in this paper is general enough, in principle, to allow it to perform in both a supervised and unsupervised context.

To examine the effectiveness of the new recruitment rule with supervised learning data, we fit SUSTAIN to Shepard et al.'s (1961) classic experiments on human category learning as replicated by Nosofsky, et al (1994). Shepard had subjects learn to classify eight items that varied on three perceptual binary dimensions into two categories (four items per category). On every trial, subjects assigned a stimulus to a category and feedback was provided. Six different assignments of items to categories were tested with the six problems varying in difficulty (Type I was the easiest to master, Type II the next easiest, followed by Types III-V, and Type VI was the hardest). The abstract structure of the six problems are shown in Table ref6types.tab.

For reference, the parameters used to fit this study are shown in Table 1 under the six types heading. While SUSTAIN was able to get the correct orderings of problem difficulty, the overall quality of the fit was degraded due to a number of subtle difficulties. For certain problems (particularly problems III-V), SUSTAIN could never achieve 100% accuracy in learning, even with an infinite number of learning trails. In these cases, SUSTAIN was in principle able to learn the problem completely before the end of the 32 blocks, but the model continued to make errors. Certain stimulus items were seen as more similar to a cluster of another category, but still activated a cluster in their own category above the threshold. Because the model could not create a new cluster like the traditional recruitment rule to handle the exception, it could not ever correctly classify the item and acheive 100% accuracy.

Discussion

SUSTAIN has been successfully applied to unsupervised learning sets using a more general notion of surprise as a method of cluster recruitment. The combined fits of Billman and Knutson's (1996) studies and Medin et al. (1987) suggest that the saliency of stimulus dimensions changes as a result of unsupervised learning and that the correlated structure of the world is most likely to be respected when there are numerous intercorrelated dimensions that are strong. Though not tested, SUSTAIN predicts that the intercorrelated structure of a stimulus set can be discovered when the intercorrelations are imperfect (as in Medin et al., 1987) if the correlations are numerous. In cases where the total number of correlations is modest, and the correlations are weak and not interrelated, SUSTAIN predicts that stimuli will be organized along a single dimension.

The formulation of a cluster recruitment method that would, in principle, allow SUSTAIN to fit both supervised and unsupervised studies did not achieve the desired quality of fit with Shepard's (1961) problem set. While the ability of SUSTAIN to master the correct problem difficulty orderings is encouraging, the results of our preliminary work suggest that our unified recruitment rule is insufficient for modeling both unsupervised learning problems and supervised data sets like the six Shepard problems. These results may suggest that while a unified model of human category learning may be a desired simplification it may, in fact, be necessary to provide separate accounts of supervised and unsupervised learning.

References

Anderson, J. 1991. The adaptive nature of human categorization. *Psychological Review* 98:409–429.

Billman, D., and Knutson, J. 1996. Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 22(2):458–475.

Hartigan, J. A. 1975. *Clustering Algorithms*. New York: Wiley.

Kruschke, J. K. 1992. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* 99:22–44.

Love, B. C., and Medin, D. L. 1998. SUSTAIN: A model of human category learning. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 671–676. Cambridge, MA: MIT Press.

Love, B. C.; Markman, A. B.; and Yamauchi, T. 2000. Modeling classification and inference learning. *Proceedings of the Fifteenth National Conference on Artificial Intelligence* 136–141.

Love, B. C.; Medin, D. L.; and Gureckis, T. 2002. SUSTAIN: A network model of human category learning. under review.

Medin, D. L.; Wattenmaker, W. D.; and Michalski, R. S. 1987. Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science* 11(3):299–339.

Nosofsky, R. M.; Gluck, M. A.; Palmeri, T. J.; McKinley, S. C.; and Glauthier, P. 1994. Comparing models of rule based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition* 22:352–369.

Sloman, S. A. 1997. Explanatory coherence and the induction of properties. *Thinking & Reasoning* 3:81–110.