We investigate human category learning from partial information provided by equivalence constraints. Human participants learned to classify stimuli on the basis of either positive or negative equivalence constraints, that is, when informed that two exemplars belong to the same category or to different categories, respectively. We discovered that when provided with positive constraints, participant categorization performance is distributed normally, but when provided with negative constraints performance distribution is bimodal. A constrained EM algorithm was used with identical constraint information to simulate the experimental setup. Results of the EM clustering algorithm showed surprising qualitative similarity to human results, with bimodality for negative but not positive constraints. Taken together, these results indicate that positive constraints provide information that may be used intuitively for categorization, while use of negative constraint information may require a less natural rule-based strategy, which most participants failed to implement. These results are consistent with the view that humans naturally use similarity-based representations (prototypes or exemplars) as opposed to rule-based strategies (e.g., strategies that use decision boundaries).

1 Introduction

Theorists generally agree that similarity serves as a guiding principal in category formation (e.g. [8, 9]), but there is still considerable debate concerning the nature and sources of perceived similarity. A related debate concerns the way in which we learn features that are relevant for classification in a specific environment ([5, 1, 6]).

We focus here on category learning from partial similarity information provided in the form of equivalence constraints ([2]). Classically, supervised category learning by both machines and humans has been studied in a context where a set of exemplars (training data) is given with the correct category label. In the scenario we study here, instead of labeled exemplars, subjects are informed that pairs of exemplars originate from the same category (Positive Equivalence Constraints, PEC) or from different categories (Negative
Equivalence Constraints, NEC). From these constraints, participants are asked to derive which of a set of exemplars belong to the same class as a single standard.

Why equivalence constraints? Clearly labels are more informative, since equivalence constraints can be extracted from labeled data, but not necessarily vice-versa. The problem is that in many real life situations, labeled exemplars are unavailable or costly to achieve. In contrast, constraint information can be automatically deduced in a variety of contexts, or obtained more cheaply from a group of un-coordinated supervisors ([2]). For example, successive views of a face or object from gradually changing viewpoints may be assumed to reflect the same - though unknown - individual (PEC), while faces simultaneously seen in separate locations can be assumed to belong to different individuals (NEC).

This scenario of learning from constraints has been studied mostly in the context of machine learning [2, 7, 4, 10] and to our knowledge was never directly addressed with regard to human category learning. The computational studies show that learning from constraints can be as effective as learning from labels. Furthermore, learning from constraints and assuming a shared covariance matrix among the different classes permits generalization to unseen categories, i.e. categories about which no constraints have been given.

One computational issue that immediately arises is the difference between positive and negative constraints. Positive constraints can be more readily incorporated into learning algorithms for two reasons: First, positive constraints are transitive, while negative constraints are not. Second, since classes are usually assumed to be sampled from compact regions of space, positive constrains are typically useful in characterizing these compact regions; negative constraints should be most useful in those rare occasions when they involve points lying close to the boundary between adjacent classes. Not surprisingly, therefore, results of machine learning with currently existing clustering algorithms [7, 2] indicate that learning on the basis of equivalence constraints indeed makes better use of positive than of negative constraints.

Recently, a number of clustering algorithms which incorporate equivalence constraints have been suggested in the machine learning literature. Among these are constrained complete-linkage ([4]), constrained K-means (COP K-means) ([10]) and the constrained EM algorithm ([7]). All of these algorithms show improved clustering performance when incorporating equivalence constraints. The constrained EM algorithm, which we had used in our simulations, is an algorithm that estimates a Gaussian Mixture Model (GMM) which can also make use of equivalence constraints. The algorithm can easily handle an unlimited number of PEC, and can be solved in closed form. However, the algorithm can only deal with \( O(n) \) NEC (where \( n \) is the data size) and requires the use of a Markov Network, which dramatically increases computation time. Therefore this algorithm uses negative constraints much less efficiently than positive constraints.

In this paper, we ask if humans, too, can learn categories from constraint information, and whether they, too, show the difference between positive and negative constraints. We created an experimental setup in which participants are asked to learn new categories in a multi-category learning scenario (as opposed to classical binary classification tasks), using only equivalence constraints to guide them.

We tested classification performance of participants who learned from positive or negative constraints, respectively, hypothesizing that these types of constraints would be used differently. More specifically and for the reasons outlined above, we expected that when provided with positive constraints (PEC), participants would be able to extract the relevant dimensions for categorization more "intuitively" than when provided with negative constraints (NEC). Therefore, in order to make these alternatives more equal, we provided constraints

---

\[1\text{Transitivity implies that if we know that exemplars 1 and 2 come from the same category and 2 and 3 come from the same category - then we also know that 1 and 3 come from the same category.}\]
in the form of pairs of exemplars with only a single feature difference between them. As a result, in both cases participants could ideally derive that this dimension was relevant (for NEC) or irrelevant (for PEC). Thus, in our setup, positive and negative constraints provide the same amount of information about the structure of classes, and ideal performers would be able to compute the exact classification rules that lead to perfect performance with both PEC and NEC.

Our main result is that this is not the case, suggesting that human participants perform the task more like our machine learning algorithms, and less like a rule based ideal classifier. In order to support this conclusion, we present an EM-based computational model that learns from both positive and negative constraints. Performance of this model bears some surprising similarity to human performance, and thus reveals useful insights for the source of disparate contributions of positive and negative constraints for category learning in human classifiers.

2 Methods

2.1 Materials

3D computer-generated pictures of “alien creature faces” were used as stimuli, as demonstrated in Figure 1. Each face was characterized by a unique combination of 5 features: chin, nose and ear shape, and skin and eye color. All (32) combinations of these 5 binary dimensions were presented in each of the 20 experiment trials. Stimuli were presented on a 22 inch, high-resolution computer screen, using specially designed software that enabled both simultaneous presentation of many stimuli and recording of participant reactions.

2.2 Participants and Procedure

Forty four university students were randomly assigned to two experimental conditions: negative-constraints (NEC) and positive-constraints (PEC). Participants were told that during the experiment they would have to learn which of the 32 “alien creatures” (test stimuli) presented on the computer screen belonged to the same tribe as the pictured “chief” (standard). Participants did not receive specific instructions about the dimensions of the faces, and they were asked to perform the task intuitively, using the “clues” provided during the task. Participants were also instructed that each trial in the experiment was independent and would necessitate learning a new way of discriminating between tribes.

Clues were provided as colored frames around pairs of aliens, indicating that the members of the pair belong to different tribes (NEC condition) or the same tribe (PEC condition). Each relation provided information on the relevance of one dimension for tribe classification on that trial. Some clues provided redundant information so that the number of clues was not indicative of the number of relevant dimensions. Two or three (of the 5 possible) dimensions were relevant for category definition on each trial. Thus, on average, half of the dimensions were relevant and half were not, and the objective amount of information provided by the NEC was identical to the amount of information provided by the PEC. As mentioned above, NECs directly indicated which dimensions were relevant for classification on the current trial, while PECs indicated the irrelevant dimensions. Still, the PEC group could derive that the rest of the dimensions were relevant. Thus, for both groups, in each trial, the clues were sufficient to derive explicitly all relevant dimensions needed for classification.

\[\text{2The full set of stimuli used had 3 discrete values for each dimension (e.g. red, green or blue eyes), with two used for each trial. This served to minimize direct memory from trial to trial.}\]

\[\text{3In some cases, constraints were given as the tribe category name(s) of four faces - either all the same tribe (PEC) or all different tribes (NEC). Thus, these names gave constraint information rather than typical label information.}\]
perfect categorization. Nevertheless, no participant performed perfectly (see Results and Figure 2).

On each trial, the clues appeared together with the alien faces for 20 seconds, then they were removed and the alien faces were shuffled. At this stage, the participant could select (by drag-and-drop) those aliens that he or she thought belonged to the chief’s tribe. At the end of 50 seconds the next experimental trial began.

Figure 1: Examples of stimuli and experimental setting. (a) Examples of alien faces that differ in all 5 dimensions. (b) A simplified example of the experimental setting: Participants decided which of the 32 test images (exemplars; 30 are shown here) belong to the same tribe (category) as the "chief" (category standard) on the basis of the equivalence constraints. Constraints were indicated by colored frames surrounding pairs of images. In the NEC condition, participants were told that the two images in each frame were from different tribes, and in the PEC condition that the two images were from the same tribe. In the example shown, participants could infer that ear shape and eye color were relevant (NEC) or irrelevant (PEC), respectively. Using information from all the presented positive constraints, participants would know all the irrelevant dimensions and could infer that the remaining dimensions were the relevant ones, i.e. in the example shown, skin color, nose and chin shape.
3 Results

3.1 Psychophysical results

We found that human participants were able to categorize novel images on the basis of information embedded in equivalence constraints. Performance was measured as Z-score, where,

\[
Z = \frac{2 \cdot Purity \cdot Accuracy}{Purity + Accuracy} = \frac{2 \cdot Hits}{2 \cdot Hits + Misses + False Alarms}
\]

Figure 2a,b demonstrates the distribution of Z-scores for the participants receiving positive (PEC) or negative (NEC) constraints, respectively. Mean Z-Scores were significantly higher than that expected from chance selection of stimuli (\(Z = 0.1875\)) both in the NEC condition (\(M = 0.53, SD = 0.20\); one-sample t-test \(t(21) = 8.08, p < .001\)), and in the PEC condition (\(M = 0.60, SD = 0.15\); \(t(21) = 12.70, p < .001\)). We also found that performance could not be explained by participant use of a categorization strategy whereby tribe membership was determined by high overall similarity between test stimuli and the trial standard (see below).

We hypothesized that even when identical objective information is provided by negative or positive constraints, participants would perform better with positive constraints. At first sight, there was no significant difference between the mean Z-Score in the PEC and NEC conditions (\(F(1,41) = 1.62, p = n.s.\)), but there was a highly significant difference between their standard deviations (Levene’s test of homogeneity of variances: \(W(1,42) = 3.55, p = 0.006\)). Furthermore, the distribution of participant performance in the PEC condition was much closer to normal than in the NEC condition (Shapiro-Wilk test of normality: \(W(22) = 0.99, p = 0.99\) vs. \(W(22) = 0.93, p = 0.18\); see also Figure 2a,b). This suggested that the NEC group of participants actually comprised two distinct subgroups.

By using the K-means clustering algorithm, we separated the participants from the NEC condition into two highly distinct groups: Low-NEC group with 13 participants, and High-NEC group with 9 participants. Mean Z-scores for these groups were not only different from each other, they were also significantly different from that of the PEC group (ANOVA \(F(2,41) = 22.45, p < 0.001\) and post-hoc t-tests). The performance distributions of these separate groups were much closer to normal than that of the group as a whole, confirming the justification of the separation procedure. (Shapiro-Wilk test of normality, Low-NEC group \(W(13) = 0.97, p = 0.81\); High-NEC group \(W(9) = 0.94, p = 0.49\)). Participant performance was better than chance (Low-NEC: \(M = 0.39, SD = 0.09\), \(t(12) = 7.98, p < .001\); High-NEC: \(M = 0.74, SD = 0.09\), \(t(8) = 17.36, p < .001\)). Furthermore, their performance could not be explained by use of the overall similarity strategy mentioned above (Low-NEC: \(t(12) = -4.63, p < .005\); High-NEC: \(t(8) = 7.49, p < .001\); PEC: \(t(21) = 3.00, p < .01\)).

In summary, the above findings reveal that while in the PEC condition classifier categorization performance distributes normally, in the NEC condition the distribution is bimodal. Performance of the low-NEC and high-NEC groups are lower and higher, respectively, than that of the PEC group. The higher performance suggests that negative constraints have the potential for providing information more efficiently than positive constraints. The fact that most subjects are unable to achieve this efficiency suggests that using NEC information may be less intuitive. Performance for these NEC groups is also respectively lower and higher than that expected for a participant using the overall similarity strategy, confirming that participants do in fact extract information from the constraints provided. The difference between the two groups - in contrast to the presence of only a single group for the PEC condition - presumably stems from the greater dependence on prior tendencies or...
biases (e.g. classifier starting strategy). In the next section we demonstrate computational findings suggesting a possible source of such biases.

Figure 2: Frequency histograms of Z-scores, and their best fitted normal distribution curves. a. Human performances in the NEC condition. b. Human performances in the PEC condition. c. EM algorithm performances in the NEC condition. d. EM algorithm performances in the PEC condition.

3.2 Clustering simulation results

In the experiment described above participants were required to perform a clustering task, after a short learning phase from equivalence constraints. In order to gain some insight into the empirical results obtained from human participants, we ran a simulation of the experiment using the constrained EM algorithm as a clustering algorithm.

Our simulation was designed to perfectly replicate the experimental setup described above: Each of the 32 different alien faces was represented by a binary 5-dimensional vector. The constraint information provided to the algorithm was identical to that presented to human participants. It is widely known that the EM algorithm converges to a local maximum of the data logLikelihood. The algorithm is also very sensitive to its initial conditions, and each choice of initial conditions implicitly determines the local minimum to which the algorithm will converge. In order to address this issue, we ran 35K different simulations for each of the groups (PEC and NEC).

Figure 2c,d displays performance histograms for the constrained EM algorithm using negative and positive constraints, respectively. On average the constrained EM which used PEC achieved better clustering Z-scores than those obtained by the constrained EM base
on NEC. This was not surprising, considering the constrained EM algorithm’s performance on real world datasets. This result may also be intuitively explained by the inherent differences between PEC and NEC mentioned above. However, we were more than surprised to discover that in the NEC simulations, the algorithm’s performance histograms were bimodal. This bi-modal distribution may be explained as follows: Despite the fact that in our experimental setup the amount of information in PEC and NEC was equally balanced, the way in which NEC are incorporated into the constrained EM algorithm cannot easily make use of such information. Thus in the NEC condition, performance critically depends on the choice of initial conditions. The bi-modal distribution of the algorithm’s Z-scores suggests that on some initial conditions the constrained EM converges to good solutions, while in most cases, it converges to relatively poor local maxima.

4 Discussion

The goal of the current study was to assess the effect of equivalence constraints on human category learning. The most important conclusion from our findings is that equivalence constraints are important sources of information for human categorization. All participants performed above chance, and all had performances that differed significantly from that expected from disregard of the constraints, and basing categorization only on overall similarity.

Inspired by both previous computational studies and cognitive studies, we hypothesized that category learning by equivalence constraints should be a natural task for human classifiers when provided with positive constraints (PEC) but not necessarily when provided with NEC. Although it is not possible to claim from the current findings that category learning by PEC is always easier than by NEC, it is clear that these two methods of providing information have different effects on both human and machine category learning, as follows: While in the PEC condition participant performance was normally distributed - as might be expected from human performance in most cognitive tasks, in the NEC condition this was not the case. The bimodal distribution observed in the NEC condition suggests that category learning by NEC is either easy or hard, depending on prior classifier tendencies or biases - that emerge only in the NEC condition. The disparity between the NEC and the PEC conditions further suggests that while category learning by PEC is relatively intuitive and natural for most humans, category learning by NEC is not necessarily so. Our explanation for this disparity is that humans are probably oriented to learning categories by first identifying a category center or its typical exemplars. Only after this initial learning stage, when in possession of a good representation of a category prototype, can humans separate well between different categories. Doing the opposite, i.e. learning category centers from information provided about differences between categories, seems not to be always possible for a human classifier - even when this information objectively enables perfect classification, as in the current experimental setting.

If this is the case, that is, if NEC information is not easily accessible by human classifiers, then why did we not observe a total failure or at least poorer average performance in the NEC (compared to the PEC) condition? The high similarity between performance of the human participants to that of the EM algorithm may provide some insight into this issue. The EM algorithm we used is designed to retrieve category representations by calculating the Gaussian mixture that best describes the data. In this sense, it can be claimed that it defines categories by the way they distribute around centers, similar to human classifiers that define categories by prototypes or typical exemplars. For this reason, the EM algorithm can more easily extract the best Gaussian mixture when provided with PEC than with NEC information, and it will be less dependent on initial conditions with PEC information. In the case of NEC, the algorithm converges to a satisfying solution only when the initial conditions are already close to this solution. We suggest that this was the source of the
bimodal performances of the EM algorithm.

If human classifiers are truly oriented to defining categories by their centers, using principles that are similar to those embedded in the EM algorithm, then we should expect them to learn new categories better when provided with PEC than with NEC. With NEC information, human performers will have to adopt a new, less intuitive, categorization strategy, that may or may not enable good performance. This new, less intuitive categorization strategy might be based on similarity, or may be a rule-based strategy. In the former case, using NEC might be too demanding in computational terms for satisfactory performance in a limited time, but the later strategy should enable rapid convergence to a satisfactory solution. This rapid convergence would mimic fortuitous initial conditions. Thus, use of an intuitive cluster center strategy for PEC information and a non-intuitive rule-based strategy by the minority of participants receiving NEC information, might explain the disparity between the success of category learning by these groups, as described in this paper.

We note in conclusion that the current experimental setting is certainly not without limitations, and therefore it can not provide the whole picture of the role of NEC and PEC information for human category learning. It might be suggested that although PEC may play a more intuitive and central role in the initial stages of category learning, NEC may be perceived by human classifiers as a more adequate source of information for later refining of category representation. Such a two-stage process for categorization would be a natural outcome of the Reverse Hierarchy Theory recently proposed for conscious perception [3]. Thus, the current findings provide important insight into a classical problem in cognitive science that has yet to be solved.

References