The Rex Leopold II model: Application of the Reduced Set Density Estimator to Human Categorization

Maarten De Schryver  
Department of Data Analysis  
Ghent University  
Ghent, Belgium  
Maarten.DeSchryver@Ugent.be

Bjorn Roelstraete  
Department of Data Analysis  
Ghent University  
Ghent, Belgium  
Bjorn.Roelstraete@Ugent.be

Abstract—Reduction techniques are important tools in machine learning and pattern recognition. In this article, we demonstrate how a kernel-based density estimator can be used as a tool for understanding human category representation. Despite the dominance of exemplar models of categorization, there is still ambiguity about the number of exemplars stored in memory. Here, we illustrate that by omitting exemplars categorization performance is not affected.

Keywords—human categorization; kernel density estimation

I. INTRODUCTION

A well-known problem in pattern recognition and machine-learning consists of selecting a subset from a large dataset in order to reduce both the storage requirements and the number of computations that are needed to make a classification decision (See [19] for an overview). The challenge is to find a representative subset of instances or exemplars without losing useful information [6]. That is, selecting these instances should be done such that categorization performance is not affected.

Many methods have been developed to succeed in this. For large datasets with a relatively easy structure, random sample methods appear to work well. A downside of these techniques is that they don’t take into account the internal structure of the data. Therefore, as the structure of the data becomes more and more complex, these techniques will start performing suboptimal [13]. There are basically two alternatives. One alternative is using classification methods (e.g condensned nearest neighbour rule [5] to capture the structure of the data. The second alternative are methods that work with the probability density function (PDF) such as: Reduced Parzen Classifier (RPC) [3], de Density-Based Multiscale Data Condensation (DBMSDC) [13] and the Reduced Set Density Estimator (RSDE) [4].

A. Human Categorization

One of our basic cognitive processes is the ability to order our environment into categories. An important question is how these categories are mentally represented. This question touched off a large debate in the perceptual categorization literature and many theories and (formal) models are developed over the years. According to prototype models, a category is represented by an abstract summary of the category. This abstract summary is called a prototype and reflects the central tendency of all the exemplars observed [16, 18]. However, these models seem to be too simple and inflexible and as a consequence prototype models fail to learn categories with a more rich internal structure [9].

In contrast with prototype models the main idea of the Exemplar view is that categories are represented by its collection of exemplars. An exemplar refers to a specific observed instance and no abstraction is made [11, 14]. Exemplar models are very successful in describing humans’ categorization performance and came to dominate the categorization field [12].

Despite the dominance of the exemplar models, there is still some ambiguity about the number of exemplars stored in memory. Some exemplar theorists insist on storing all observed exemplars that have been observed (see for example [16]). Also, in practice, the all-exemplars-are-stored principle is widely spread: exemplar models tend to include all exemplars when they are fit to empirical data [2]. However, it seems implausible that every exemplar we have encountered would be a separate part of the representation. To avoid this memory load problem, some authors suggest that only the best, most typical or most frequent instances are stored [8].

In [2] it is shown that a subset of exemplars may be sufficient for category representation. The authors used the Rex Leopold I model for testing explicitly the assumption that subjects only retain a subset of exemplars. The Rex LI model is the most basic model of the Reduced Exemplar models (see [17]) and is designed to be identical to the General Context Model [14], except that the full set of exemplars can be replace by a reduced set. A shortcoming of this model is that all possible subsets need to be tested. As a consequence, for larger categories, it becomes impractical to test all possible subsets. To limit the possible number of subsets that should be considered, the Rex Leopold II (Rex LII) was developed.
II. THE REX LEOPOLD II MODEL

The Rex LII model is designed to be identical to the Generalized Context Model [14] with the exception that the full set of exemplars can be replaced by a reduced set of exemplars. The model should fulfill only one condition: omitting exemplars may not affect categorization performance. Indeed, the exemplar-reduction algorithm applied by Rex LII is essential in finding those exemplars that can be omitted without any harm done.

A. The Generalized Context model

One of the most important exemplar models is the Generalized Context Model or GCM [14], a generalization of [11] context theory of classification. Let \( C_k = \{x_1, \ldots, x_n\} \) the set of stored category \( C_k \) exemplars. According to the GCM, the probability that stimulus \( i \) is classified in category \( C_k \) is given by

\[
P(C_k|S_i) = \frac{b_k \sum_{j \in C_k} s_{ij}}{\sum_{k=1} b_k \sum_{j \in C_k} s_{ij}}
\]

where \( b_k \) (0 \( b_k \leq 1\)) is the category \( C_k \) response bias and \( s_{ij} \) denotes the similarity between exemplars \( i \) and \( j \). In the equation it is shown that the classification of stimulus \( i \) is based on the summed similarity of this stimulus with all exemplars that form the category representation. The similarity measure is assumed to be related to the psychological distance \( d_{ij} \) by,

\[
s_{ij} = \exp(-d_{ij}^q),
\]

where \( q = 1 \) yields an exponential function and \( q = 2 \) yields a Gaussian function. The psychological distance between stimuli \( i \) and \( j \) is given by

\[
d_{ij} = c|\sum_{m=1}^m (w_m | x_{im} - x_{jm} |^r)^{1/r},
\]

where \( x_{im} \) is the psychological value of exemplar \( i \) on dimension \( m \). The parameter \( c \) (0 \( c < \infty\)) is a sensitivity parameter and the parameter \( w_m \) (0 \( w_m \leq 1\)) is the attention weight for dimension \( m \). The distance metric \( (r = 1\): city-block-metric; \( r = 2\): Euclidean metric) is defined by the exponent \( r \). In most application of the GCM, the model has been restricted to the versions whereby \( (1) \) the similarity function is exponential and the distance metric is the city-block metric \( (q = r = 1) \) and \( (2) \) the similarity function is Gaussian and the distance metric is the Euclidean distance metric \( (q = r = 2) \) [1, 7].

[?], showed that the Generalized Context Model with \( q = r \) is equivalent to probability matching with a kernel estimator using a multivariate Laplacian kernel function \( (q = r = 1) \) or a multivariate normal kernel function \( (q = r = 2) \) (see also [17]). For a model, whereby \( q = r = 2 \), one can rewrite the summed similarity as defined in (2) as follows:

\[
\sum_{j \in C_k} s_{ij} = (2\pi)^{m/2} |\Sigma|^{1/2} \sum_{j \in C_k} N(x, x_j, \Sigma)
\]

where \( \Sigma \) is a \( m \times m \) matrix with diagonal elements \( \frac{1}{\sum_{j \in C_k} s_{ij}} \) and is called the covariance or bandwidth matrix.

The category similarity or summed similarity \( \sum_{j \in C_k} s_{ij} \) is defined by two (free) parameters: \( (1) \) the attention weight \( w \) and \( (2) \) the sensitivity parameter \( c \). The attention weight reflects the proportion of attention subjects distribute among dimensions. Dimensions that have greater classificatory significance, will have a greater weight [15]. The second parameter is the sensitivity parameter \( c \) and acts as a scale factor, denoting the steepness of the exponential decay. The similarity between two instances will be greater for a small value of \( c \), while for a large value of \( c \) the similarity will be smaller for these two instances. Indeed, the sensitivity parameter reflects the discriminability between stimuli: if subjects have more experience with the stimuli, a larger value of \( c \) is expected. If \( c \to 0 \) no distinction is made between instances. On the other hand, if \( c \to \infty \) only the learned exemplars show similarity with the category and as a consequence, it’s impossible to generalize.

A relative low value for the sensitivity parameter \( c \) corresponds to a relative wide bandwidth of the kernel. A consequence is that the model needs fewer exemplars to estimate the density function. In other words, only exemplars that are very typical for the category will be retained. In contrast, a high value of \( c \) will correspond to using a lot of exemplars in order to represent the category.

B. The Reduced Set Density Estimator

The Rex Leopold II model uses the Reduced Set Density Estimator (RSDE) developed by [4], a kernel-based density estimator optimal in the \( L_2 \) sense. The RSDE allows for several optimization strategies such as Multiplicative Updating or Sequential Minimal Optimization (SMO) in order to estimate the kernel weighting coefficients. The reduced set of exemplars used by the RSDE results from the minimization of the integrated square error between the estimator and the true density. This algorithm is very interesting for the Rex LII. Firstly, it is a density-based approach which can be easily incorporated in the categorization model. Furthermore, the algorithm does not require the estimation of additional parameters (DBMSDC needs to compute a k-nearest neighbour). Finally, the algorithm uses the entire dataset (in contrast to RPC, that depends on the initially chosen random sample). For more details we refer the the article of [4].

III. EXPERIMENT

To illustrate the proposed exemplar-reduction algorithm, the Rex Leopold II model were fit to data collected from a human categorization experiment. We replicated Experiment
1 (Condition 1) in the [10] study. In this experiment, two large-size categories, generated from mixtures of bivariate normal distributions were constructed.

A. Method

1) Subjects: Four students from Ghent University participated in exchange for a small payment.

2) Stimuli and apparatus: Stimuli were circles with a radial line and were presented on a 17 in. monitor with 800 x 600 resolution. The circles varied in size and angle of orientation of the radial line. The category distribution parameters are given in Table 1. For both the training phase and the transfer phase, stimuli were generated from one of the four density functions of Category A or B (a density function was randomly selected with probability of 0.25 each).

Table I

<table>
<thead>
<tr>
<th>Category</th>
<th>den.f.</th>
<th>( \mu_r )</th>
<th>( \mu_a )</th>
<th>( \sigma_r )</th>
<th>( \sigma_a )</th>
<th>( \rho_{ra} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>50</td>
<td>33</td>
<td>14</td>
<td>14</td>
<td>-0.9</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>67</td>
<td>50</td>
<td>14</td>
<td>14</td>
<td>0.9</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>33</td>
<td>50</td>
<td>14</td>
<td>14</td>
<td>-0.9</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>50</td>
<td>67</td>
<td>14</td>
<td>14</td>
<td>-0.9</td>
</tr>
</tbody>
</table>

Note. den.f. = density function; \( \mu_r \) (\( \mu_a \)) = mean of the radius (angle); \( \rho_{ra} \) = correlation between radius and angle.

3) Procedure: Participants were tested individually in a dimly lit room in a total of five daily sessions. The first four days, 3200 experimental trials were presented in order to learn the stimuli assigned to category A and the stimuli assigned to category B. Each day, stimuli were presented in eight blocks of 100 trials. Stimuli were presented until the participant responded by pressing X (for category A) or N (for category B) on the keyboard. Feedback was given after each trial. During the last session, one block of 100 trials with feedback was presented. So, a total of 3300 stimuli were presented during the training phase. The last session ended by a transfer phase, consisting of three block of 100 stimuli. During this transfer phase, no feedback was given.

B. Results

When fitting the GCM (whereby \( q = r = 2 \) and \( b_A = 0.5 \)), the sensitivity parameter \( c \) and the attention weight parameter \( w_1 \) were freely estimated. A computer search was used to find the parameter values for each participant that maximized the log-likelihood function

\[
L_{GCM} = - \sum_i \sum_A \ln f_{iA} + \sum_i \sum_k f_{iA} \ln P(C_i|S_i) \tag{4}
\]

where \( f_{iA} \) and \( P(C_i|S_i) \) are respectively the observed frequency and predicted probability with which stimulus \( i \) is classified in category \( A \). The obtained values of the sensitivity parameter \( c \) and the attention weight parameter \( w_1 \) were needed in the exemplar-reduction algorithm in order to estimate the kernel weighting coefficients. Then, the full set of exemplars was replaced by the retained subset of exemplars to compute the log-likelihood \( L_{Rex} \). However, the same values for the free parameters as used to calculate the log-likelihood of the GCM \( L_{GCM} \) were applied, so no new parameter-values were estimated. Based on the log-likelihood measures the AIC score was calculated and should give an indication about the accuracy of the reduction. Results are reported in Table 2.

Table II

<table>
<thead>
<tr>
<th>P</th>
<th>AIC(_{GCM})</th>
<th>c-par</th>
<th>AIC(_{Rex})</th>
<th>% Rex</th>
<th>% red</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>132</td>
<td>91</td>
<td>145</td>
<td>91</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>230</td>
<td>85</td>
<td>223</td>
<td>86</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>142</td>
<td>94</td>
<td>130</td>
<td>94</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>164</td>
<td>92</td>
<td>161</td>
<td>93</td>
<td>9</td>
</tr>
</tbody>
</table>

Note. \( P \) = participant, \( AIC_{GCM} \) = AIC score for the GCM, \% GCM = Percentage correct prediction of the GCM, \% c-par = Estimated c-parameter, \( AIC_{Rex} \) = AIC score for the ReX LII, \% Rex = Percentage correct prediction of the ReX LII, \% red = Percentage reduction based on the RSDE.

Note that the amount of reduction strongly depend on the value of the c-parameter. The AIC-scores obtained by the ReX LII are acceptable in comparison to the GCM (The aim of ReX LII is not to outperform the GCM, but to fit at least as well using only a reduced set of exemplars). Finally, the percentage correct classification is almost similar for both models.

IV. Conclusion

In this paper we showed how machine learning methods can be used to give insight in human categorization strategies. In particular, we implemented the RSDE, a reduction technique developed by [4], in our formal model of categorization. This gave evidence for the view that not all categories can be used to give insight in human categorization strategies.

REFERENCES


