Exploring Feature Collocation for Semantic Concept Identification

Meng Yu (myu3@uci.edu)
Department of Mathematics,
University of California, Irvine, CA, 92697, USA

Keywords: feature collocation, semantic concept identification, l_1 regularization.

Problem Description

Motivated by object representation in psychology, we present a binary feature classifier for the purpose of semantic concept category identification/classification by incorporating feature distribution. We propose the classification algorithm based on the variant of l_1 norm regularized sparse classifiers, where the features are weighted according to their distribution, which is estimated by "maximum collocation". This method achieves high accuracy in identifying semantic concepts, outperforming standard benchmark methods on a large database of animal and artifact features.

Suppose your friend tells you they are thinking of a particular mammal animal, asks you what type it is, and starts listing its features: it has a tail, has four legs, can't swim, and so on. You are now faced with a *category identification* problem, which requires you to infer the most likely category of an instance given knowledge of some of its features (Kemp, Chang, & Lombardi, 2010). Category identification offers an interesting window onto the structure of mental representations, since it involves the relationship between categories and features, and so requires the representation of both what makes instances different, and what makes them the same. One of the main shortcomings of existing classification work is that feature importance has not been well investigated (Zhang, Yu, Lee, & Xin, 2011). Features are often preselected from the beginning which actually do not equally or positively contribute to the performance of classification. However, not all of the features will be important to an object's representation. Thus, weighting features without adversely affecting the performance is an important task for classification.

Feature Distribution and Weighting

We propose to weight the features such that categories can be differentiated more efficiently according to the binary feature's distribution[1]. This is motivated by stimuli representation in psychology (Jones, 1983) since it was studied that people identify the semantic concepts by choosing features in a systematic way. One task is to choose important features by how useful they are in distinguishing categories. For example, in mammal domain, feature "is pregnant" is less important than "has long neck". This empirical motivation becomes the principle for feature importance measure.

Maximum collocation is described here for measuring the feature importance based on two heuristics (Zeigenfuse & Lee, 2010). The first of these is maximum cue validity, defined as the maximum over categories c_j ($j = 1, 2, ..., n_c{}^2$) of cue validity, the probability an instance belongs to c_j given that it has a feature f, $p(c_j|f)$. We also look at maximum category validity, defined similarly as the maximum over categories c_j of the category validity, the probability an instance has a feature f given that it belongs to c_j, $p(f|c_j)$. Finally, the maximum collocation is the maximum over categories c_j of the collocation, the product of a feature's cue and category validities, $p(c_j|f)p(f|c_j)$. Maximum collocation is a measure of how simultaneous concentrated in and diffuse across a category a feature is. Features with high maximum collocation are associated with most instances within a category and few outside it, as illustrated by Feature 1 in Table 1. Alternatively, Features 4 and 6 show why it is necessary for both of these to be true. Those features associated with only a small fraction of instances within a single category will have high maximum cue validity but low maximum category validity (Feature 4). Those features possessed by most instances in more than one category will have high maximum category validity but low maximum cue validity (Feature 6).

Table 1: Representative features illustrating behavior of the usefulness measures. Black dot means that the instance has the corresponding feature.

	C 1	C 2	C 3	Cue	Cat.	Colloc.
f 1	• • •			1	1	1
f 2	• • •	•		3/4	1	3/4
f 3	• •			1	2/3	2/3
f 4	•			1	1/3	1/3
f 5	• • •	• •	• •	3/7	1	3/7
f 6	• • •		• • •	1/2	1	1/2

Collocation Weighted Classifiers

The motivation for using sparse representation (SR) for category identification is that SR adaptively selects the relevant support data points from the training data, allowing us to identify the semantic concept using a few relevant examples from the training dataset, and alleviating adverse effects of instances variability in the training dataset. Mathematically, in a typical SR formulation, a *dictionary* D is constructed as $D = [d_1, d_2, ..., d_n]$, where each $d_i \in R^m$ is a feature vector of ith instance. To represent a test instance in terms of its feature vector y, SR solves the equation $y = D\theta$, where a regularization is enforced on θ, such that only a small number

[1] Feature value is "yes" or "no"

[2] n_c is the number of categories.

of instances from the dictionary D are selected to describe y. Sparsity regularization helps the representation to rule out irrelevant instances and be insensitive to within-category variability in the dictionary. The test instance is assigned to the category with the smallest residual in representing y as a linear combination using all instances from that category.

With the feature importance measure, features are weighted by maximum collocation. Denote U_{col} as the diagonal matrix with $u_{col}(f)$ as the diagonal entries, where $u_{col}(f)$ is the maximum collocation for feature f. The weighted dictionary and test instance become $U_{col}D$ and $U_{col}y$ respectively. We developed three SR variants for the classification. Then, as a result, all the features contribute unequally in the sparse representation. The conventional SR optimization is given by

$$\theta^* = \arg\min_{\theta} \frac{1}{2}||U_{col}y - U_{col}D\theta||_2^2 + \mu||\theta||_1, \qquad (1)$$

where μ is the trade-off parameter.

Due to the non-negativity of the features, the above l_1 regularized unconstrained convex optimization (1) becomes a non-negative penalized l_1 regularized constrained convex optimization (2) as below. The non-negative weights in θ indicate the importance of an instance with the natural interpretation that this constraint forces representations that include on instances that provide evidence for a category identification decision.

$$\theta^* = \arg\min_{\theta} \frac{1}{2}||U_{col}y - U_{col}D\theta||_2^2 + \mu||\theta||_1 \quad s.t.\ \theta \geq 0$$

Eventually, since we expect sparse representation errors, for which l_1 norm regularization seems to be more appropriate. The optimization (1) is re-formulated as

$$\theta^* = \arg\min_{\theta} ||e||_1 + \mu||\theta||_1 \quad s.t.\ U_{col}y = U_{col}D\theta + e. \qquad (2)$$

Dataset and Evaluations

Our data come from the Leuven Natural Concept Database (DeDeyne, et al, 2008), involving 295 words (i.e. categories), distributed over 11 semantic domains: five animal domains (30 mammals categories, 30 birds categories, 23 fish categories, 26 insects categories, and 20 amphibians&reptiles categories) with 764 animal features, and six artifact domains (31 kitchen utensils categories, 30 clothing categories, 27 musical instruments categories, 29 vehicles categories, 19 weapons categories and 30 tools categories) with 1295 artifact features. Features used to describe those words include perceptual, functional characteristics and any other background information that applies. Most importantly for our modeling, the words (i.e. semantic concept categories) and features were combined in a feature verification task, in which four participants judged whether or not each of the features belonged to each of the words. In the experimental evaluations, we split the data into training and test sets for a 4-fold cross-validation. In each validation, we train the classifier using data from three participants and test on the participant that is left out, i.e. a leave-one-out cross-validation. We use the set of features a participant assigned to a word — “can fly”, “is small”, and so on — as the input to a category identification problem, for which the task is to identify the category associated with that list of features.

The identification accuracy for the proposed variants of weighted sparse representation methods achieves 84 percent in average, outperforming typical classification methods, such as k nearest neighbor, logistic regression and decision tree. To examine the performance variation of the proposed feature collocation based classifiers on different categories and domains, we compute an overall rank of average residual for each category, shown in Fig. 1 as an example. Large magnitudes suggest ambiguities in semantic category identification.The identification errors for mammals is very small, but large for amphibians&reptiles domain. We believe the proposed approach constitutes a useful starting point for understanding how people do semantic concept category induction.

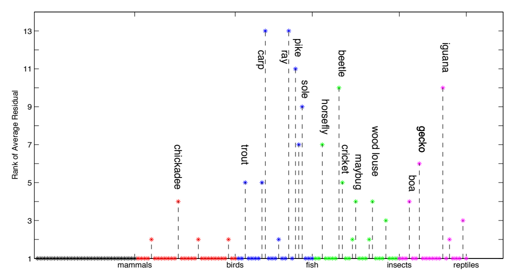

Figure 1: Performance variation for semantic concepts in animal domains.

References

DeDeyne, S., Verheyen, S., Ameel, E., Vanpaemel, M., Dry, M., & Voorspoels, W. (2008). Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior Research Methods*, *40*, 213–231.

Jones, G. V. (1983). Identifying basic categories. *Psychological Bulletin*, *94*, 423–428.

Kemp, C., Chang, K. M., & Lombardi, L. (2010). Category and feature identification. *Acta Psychologica*, *133*, 216–233.

Zeigenfuse, M. D., & Lee, M. D. (2010). Heuristics for choosing features to represent stimuli. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 1565–1570).

Zhang, S., Yu, M., Lee, M. D., & Xin, J. (2011). Modeling category identification using sparse instance representation. In *Proceedings of the 33rd annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.