On the Number of Subpopulations in Coevolutionary Computation: A Database Application

Rui Li, Bir Bhanu and Krzysztof Krawiec
Center for Research in Intelligent Systems
University of California
Riverside, CA, 92521
{rli, bhanu}@vislab.ucr.edu, krawiec@cs.put.poznan.pl

ABSTRACT

Among the existing feature selection/synthesis approaches, *Coevolutionary Feature Synthesis* (CFS) based on *Coevolutionary Genetic Programming* (CGP) has shown good performance on a variety of applications. In this paper, we propose an MDL-based fitness function to help pick a reasonable number of synthesized features which is equal to the number of subpopulations. It naturally balances the feature transformation complexity and classification performance. Experiments on a real image database show that the new fitness function solves the problem quite well.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology – *Classifier design and evaluation.*

General Terms: Algorithms.

Keywords

Co-evolution, pattern classification, image retrieval.

1. MDL-BASED FITNESS FUNCTION

We introduce an MDL-based fitness function in a coevolutionary genetic programming algorithm [1]. It is defined as fit = $\sum_{i=1}^{S} size(CO_i) + n_e(\log(n) + \log(C))$. It takes both the size of the composite operator vector and the training error into the fitness evaluation process. The first term is the total number of bits required to describe the best composite operator from each subpopulation. The second term is the number of bits to represent the indices of the misclassified training samples and their new labels (which class a sample is classified to). For the first term, the number of bits for the composite operator is defined as follows. In CGP algorithm, there are 12 possible primitive operators in the tree [1]. The leaf nodes are primitive features so *D* labels (the number of primitive features) are enough to label them. Basically $\log_2(12+D)$ bits are enough to encode a node. But an extra of 16 bits are required for four unary operators with a constant float type parameter. The tree is pre-traversed to get the total number of bits for the complete representation of the tree. In the second term, n_e is the number of misclassified training samples, *n* is the total number of training samples and *C* is the number of classes. Thus it gives the indices of the misclassified images and their new labels. With such information, the receiver can recover the classification on the training data and obtain the Bayesian classifier in the low dimension. The fitness function is integrated into the CGP iterations to help pick the number of subpopulation.

2. EXPERIMENTAL RESULTS

We apply the CGP algorithm with the proposed new MDL-based fitness function on a dataset *Corel-1200* from the Corel stock and evaluate the classification performance. The dataset has 12 classes with a total of 1200 images. Each image is represented by 40 features. Figure 1 shows the training accuracy, the testing accuracy and the MDL value with regard to the number of composite operators. From the figure, we can see that, when the composite feature number increases, the training classification accuracy increases to 100% then at a certain point (composite feature number = 35, named *crash point*) it drops to a very small value. The drastic drop means that when the composite feature dimensionality increases too high, there are not enough data to fit the Gaussian distribution in each class. It looks like a "crash. When the composite feature number increases, the testing classification accuracies increases at the beginning, reaches a maximum value at a certain point (composite feature number = 10, named *target point*), then starts dropping gradually, while the training accuracy is still increasing. The composite feature number at the target point (10) is the optimal number of subpopulations. The accuracy with 95% confidence interval is 0.8643±-0.027.The performance between the *target point* and the *crash point* describes the "overfitting" behavior. As the direct indicator of the classification performance, the MDL curve reaches its minimum value at the *target point* to point out it reaches the target – finding a best number of subpopulations.

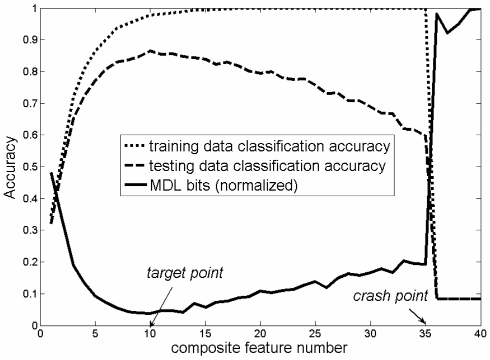

Figure 1. MDL value (normalized) and classification accuracies.

REFERENCES

[1] A. Dong, et al., "Evolutionary feature synthesis for image databases," in *IEEE WACV*, 2005, pp. 330-335.

GECCO'07, July 7–11, 2007, London, England, United Kingdom
ACM 978-1-59593-697-4/07/0007.