MDL-based Fitness for Feature Construction

Leila S. Shafti
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, E-28049, Spain
Leila.Shafti@uam.es

Eduardo Pérez
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, E-28049, Spain
Eduardo.Perez@uam.es

Categories and Subject Descriptors: I.2.6 [Artificial Intelligence]: Learning–*Induction*; I.5.2 [Pattern Recognition]: Design Methodology–*Feature evaluation and selection*

General Terms: Algorithms, Design

Keywords: machine learning, attribute interaction, feature construction, fitness function, Entropy, MDL.

1. EXTENDED ABSTRACT

Primitive data representation of real-world data facilitates attribute interactions, which make information opaque to most learners [1]. Feature Construction (FC) aims to abstract and encapsulate interactions into new features and outline them to the learner. When a GA is applied to perform FC, the goal is to generate features that facilitate more accurate learning. Then the GA's fitness function should estimate the quality of the constructed features. We propose a new fitness function based on Minimum Description Length (MDL). This fitness is incorporated in MFE2/GA [4] to improve its accuracy. The new system is compared with other systems based on Entropy or error-rate fitness.

There are three common forms of evaluating features: i) MDL-based fitness function measures the inconsistency and complexity of constructed features based on MDL principles [3]. ii) Entropy-based fitness measures amount of uncertainty produced by features. iii) Classifier error rate measure redescribes data using constructed features and applies a learner to classify data and measure its error rate. This measure is computationally expensive and not the most appropriate for a GA. Therefore, we only focus on MDL and Entropy-based fitness.

The MDL principle establishes that the optimal solution is obtained by selecting a theory that minimizes the sum of the code lengths corresponding to theory and errors. This criterion is used to design the new fitness function. The new fitness includes two terms. The first one approximates the complexity of the collection of new features (theory). The complexity of each feature is determined by attributes participated in construction of the feature. The second term accounts for the misclassifications produced by new features (errors). It measures the inconsistency in data after adding new features. We normalized each measure by dividing it by its maximum value. GA aims to minimize the sum of the two normalized measures. Given two individuals equally consistent with the data, the fitness function prefers the one with several features defined over smaller attribute subsets, rather than one feature defined over the union of subsets.

To compare this fitness with Entropy-based fitness, we also modified MFE2/GA to apply an Entropy-based fitness function and called it MFE2/GA$_E$. For each individual, the fitness function measures the Entropy of the concept given the values of new features [2]. To reduce overfitting, part of training data are used for constructing functions and all training data are used for fitness evaluation.

We performed an empirical study using a benchmark of synthetic concepts designed to involve several complex attribute interactions. The study shows that the proposed MDL-based fitness function yields significantly better predictive learning accuracy than other fitness functions solely based on entropy or error reduction. MFE2/GA$_E$ in most cases achieves lower accuracy because Entropy does not consider the complexity of the theory proposed by the new features. It constructs large features that perfectly match training data and produce overfitting. Comparing the average number of GA's generations shows that MDL-based fitness function helps GA to converge to optimal solution faster than the Entropy-based fitness. In addition, our results show that even without the improvement of an MDL-based fitness, the MFE2/GA$_E$ approach with an Entropy-based fitness measure retains most of its accuracy advantage over two relevant learners: a standard learner as C4.5 (trees and rules) [2], and HINT [5], a non-GA FC method. Finally, similar empirical results were found using real-world data from the Braille Code domain.

2. ACKNOWLEDGMENTS

This work is supported by the Spanish Ministry of Science and Technology, Grant number TSI2005-08225-C07-06.

3. REFERENCES

[1] A. A. Freitas. Understanding the crucial role of attribute interaction in data mining. *AI Review*, 16(3):177–199, Nov. 2001.

[2] R. J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.

[3] J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, Jun. 1983.

[4] L. S. Shafti and E. Pérez. Reducing complex attribute interaction through non-algebraic feature construction. In *Proc. of the IASTED-AIA*, pages 359–365, Innsbruck, Austria, Feb. 2007. Acta Press.

[5] B. Zupan, M. Bohanec, I. Bratko, and J. Demsar. Learning by discovering concept hierarchies. *Artificial Intelligence*, 109(1-2):211–242, 1999.

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