Constructive Induction of Fuzzy Cartesian Granule Feature Models using Genetic Programming

James G. Shanahan
Xerox Research Centre Europe (XRCE)
6 chemin de Maupertuis
38240 Meylan, FRANCE
Email: Jimi.Shanahan@XRCE.Xerox.com

James F. Baldwin
Dept. of Engineering Mathematics
University of Bristol, Bristol,
BS8 1TR, ENGLAND
J.F.Baldwin@Bristol.ac.uk

Trevor P. Martin
Dept. of Engineering Mathematics
University of Bristol, Bristol,
BS8 1TR, ENGLAND
T.P.Martin@Bristol.ac.uk

Abstract

The G_DACG (Genetic Discovery of Additive Cartesian Granule feature models) constructive induction algorithm is presented as a means of automatically identifying rule-based Cartesian granule feature models from example data. G_DACG combines the powerful search capabilities of genetic programming with a rather novel and cheap fitness function based upon the semantic separation of learnt concepts expressed in terms of fuzzy sets extracted over Cartesian granule features. G_DACG is illustrated on a variety of artificial and real world problems.

1. INTRODUCTION AND APPROACH

Constructive induction algorithms refer to learning algorithms, which produce hypothesis that employ/derive features that are not present in the original training set (Michalski, Bratko et al. 1998). A Cartesian granule feature is an example of a derived multidimensional feature. Cartesian granules (characterised by fuzzy sets) provide an abstraction of the multidimensional universe by carving it into regions that are drawn together as result of indistinguishability, similarity, proximity or functionality. (Shanahan 1998) has shown that systems can be quite naturally described in terms of Cartesian granule features incorporated into rule-based models. Due to the constructive nature of the induction algorithm the identification of good, parsimonious Cartesian granule feature models, which can adequately model problems in a transparent fashion, turns out to be an exponential search problem. In this paper we propose a population-based search algorithm which relies upon genetic programming (Koza 1992) to iteratively hone in on good Cartesian granule features; G_DACG (Genetic Discovery of Additive Cartesian Granule feature models). G_DACG considers feature abstraction and feature selection simultaneously thereby avoiding local minima models that can result from treating these tasks independently as in the case of for example, decision tree approaches (Quinlan 1986). Most induction techniques that employ GP use a fitness function that is based upon evaluating the induced model on a test dataset (Koza 1992). This form of fitness evaluation may prove expensive with large datasets. Here we propose a novel and cheap fitness function which relies on the semantic separation of learnt concepts that are expressed in terms of Cartesian granule fuzzy sets, thus avoiding the mammoth task of evaluating the model on a test dataset. The key steps involved in the G_DACG algorithm are as follows :

- Generate a random set of individual Cartesian granule features
  - Assign a fitness value to each individual
  - REPEAT
    - Generate *n* new fitnessed children
    - Insert new children into population
    - Eliminate *n* individuals from the population
    - Select best *m* individuals and form a rule-based model
- UNTIL a good solution or the number of generations expires.
- Select best Cartesian granule feature rule-based model

2. RESULTS

G_DACG has been illustrated on variety of problems and the discovered models in general performed as well or outperformed other well-known machine learning algorithms e.g. on the Pima dataset an accuracy of 79.7% is achieved using G_DACG. See **Table 1** for details. NN refers to neural networks and MAID3 to mass assignment based decision trees (Baldwin, Lawry et al. 1997).

Table 1: Comparison of various approaches.

Problem	*Class Count*	*G_DACG*	*NN*	*MAID3*
Pima Diabetes (Smith, Evalhart et al. 1988)	2	79.7	78	79.7
Road Classification	2	97	97	-
Sin(x*y) using RMS Error	Continuous	2.6%	2.2	4.2

Acknowledgements

James Shanahan carried out this work while at the University of Bristol on a EU Marie Curie Fellowship.

References

Baldwin, J. F., J. Lawry, et al. (1997). Mass assignment fuzzy ID3. Fuzzy Logic Workshop, London, UK.

Koza, J. R. (1992). Genetic Programming. Massachusetts, MIT Press.

Michalski, R. S., I. Bratko, et al., Eds. (1998). Machine Learning and Data Mining. New York, Wiley.

Quinlan, J. R. (1986). "Induction of Decision Trees." Machine Learning **1**(1): 86-106.

Shanahan, J. G. (1998). Cartesian Granule Features: Knowledge Discovery of Additive Models for Classification and Prediction, PhD Thesis, Dept. of Engineering Maths, University of Bristol, Bristol, UK.

Smith, J. W., J. E. Evalhart, et al. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Computer Applications and Medical Care Symp.