Information Sciences 180 (2010) 1183–1200

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

GP-COACH: Genetic Programming-based learning of COmpact and ACcurate fuzzy rule-based classification systems for High-dimensional problems ☆

F.J. Berlanga [a,*], A.J. Rivera [b], M.J. del Jesus [b], F. Herrera [c]

[a] University of Zaragoza, Dept. of Computer Science and Systems Engineering, E-50018 Zaragoza, Spain
[b] University of Jaén, Dept. of Computer Science, E-23071 Jaén, Spain
[c] University of Granada, Dept. of Computer Science and Artificial Intelligence, E-18071 Granada, Spain

ARTICLE INFO

Article history:
Received 7 October 2008
Received in revised form 25 November 2009
Accepted 22 December 2009

Keywords:
Classification
Genetic programming
Fuzzy rule-based systems
Genetic fuzzy systems
High-dimensional problems
Interpretability-accuracy trade-off

ABSTRACT

In this paper we propose GP-COACH, a Genetic Programming-based method for the learning of COmpact and ACcurate fuzzy rule-based classification systems for High-dimensional problems. GP-COACH learns disjunctive normal form rules (generated by means of a context-free grammar) coded as one rule per tree. The population constitutes the rule base, so it is a genetic cooperative-competitive learning approach. GP-COACH uses a token competition mechanism to maintain the diversity of the population and this obliges the rules to compete and cooperate among themselves and allows the obtaining of a compact set of fuzzy rules. The results obtained have been validated by the use of non-parametric statistical tests, showing a good performance in terms of accuracy and interpretability.

1. Introduction

In the design of any fuzzy rule-based system (FRBS) learning method, there are two main (and contrary) goals to be maximized: the accuracy and the interpretability of the knowledge extracted. In the 1990s, more attention was given to accuracy maximization, and different approaches were developed to improve the accuracy of FRBSs, although this improvement was usually at the cost of their interpretability. However, more recent studies [17–19,24,41,46] have indicated the necessity of an interpretability-accuracy trade-off in the design of FRBSs.

Such a trade-off is more difficult to achieve when the problem to be solved has high dimensionality, that is, a high number of input features or a high number of examples. In this paper we consider the high dimensionality problem with regard to the number of features. In this kind of problems, a linear increase in the number of input features causes an exponential growth of the fuzzy rule search space, what is popularly known as the combinatorial rule explosion problem [23]. This growth makes the learning process more difficult and, in most cases, leads to an FRBS with a high level of complexity (with respect to the number of rules, features and conditions included in each rule).

An analysis of the specialized literature shows that there exist two main solutions for tackling this problem of high dimensionality in the learning of compact, interpretable and accurate FRBSs:

☆ Supported by the Spanish Ministry of Science and Technology under the Projects TIN2008-06681-C06-02 and TIN2008-06681-C06-01 and the Andalusian Research Plan TIC-3928.
* Corresponding author.
E-mail addresses: berlanga@unizar.es (F.J. Berlanga), arivera@ujaen.es (A.J. Rivera), mjjesus@ujaen.es (M.J. del Jesus), herrera@decsai.ugr.es (F. Herrera).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2009.12.020

1. *Carrying out a feature selection process* [40,50,54], which determines the most relevant variables before (*a priori feature selection*) or during (*embedded feature selection*) the FRBS inductive learning process. This process reduces the fuzzy rule search space and increases the efficiency and accuracy of the learning stage.
2. *Compacting and reducing a previously learned rule set in a postprocessing stage* [16,48,65]. The methods which employ this strategy operate by combining rules and/or selecting a subset of them from a given rule set in order to achieve the goal of minimizing the number of rules used while maintaining (or even improving) the FRBS performance.

Evolutionary algorithms (EAs), and particularly genetic algorithms (GAs) [36], have been successfully applied to FRBS learning, giving way to the appearance of the so-called genetic fuzzy systems (GFSs) [27,41,46]. Many different GFSs have been formulated for the learning of fuzzy rule sets. Although most of these methods are GA-based, it is also possible to find proposals using other different types of EAs such as genetic programming (GP) [53], a type of evolutionary algorithm that uses variable-length trees to represent the different individuals in the population, instead of fixed-sized vectors with binary, integer or real codification [35,56,62,68].

In this paper we propose GP-COACH (Genetic Programming-based learning of COmpact and ACcurate fuzzy rule-based classification systems for High-dimensional problems), a method for dealing with problems having a high dimensionality with regard to the number of input features considered. It is a GP-based method that allows the absence of some of the input features in each rule. GP-COACH learns disjunctive normal form (DNF) fuzzy rules (generated by means of a context-free grammar), coded as one rule per chromosome and with the population forming the rule set, thus following a genetic cooperative-competitive learning approach [41]. It uses a competition mechanism between rules (token competition) which simultaneously maintains the diversity in the population and deletes irrelevant rules during the learning process, allowing us to obtain compact FRBCSs (with few rules, variables, and conditions per rule) with a high generalization capability.

An experimental study involving 24 data sets and five well-known FRBCS learning algorithms has been carried out. Non-parametric statistical methods have been used to compare and analyze the compactness and accuracy of the experimental results. They show the good performance (in terms of accuracy and interpretability) of our approach. Moreover, the suitability of some GP-COACH components such as the token competition diversity mechanism, the use of specific genetic operators and the advantages of using a GP algorithm instead of a traditional GA have been also analyzed.

This paper is organized in the following way. In Section 2, some preliminaries are described. The GP-COACH algorithm is comprehensively described in Section 3. The experimental framework is presented in Section 4. In Sections 5 and 6 we have included the experimental results and their analysis. Finally, in Section 7 some concluding remarks are pointed out.

2. Preliminaries

In this section, we introduce the notation that has been used in this paper. Then we describe the structure of an FRBCS, a brief introduction to GFSs and finally a short review of the main approaches we find in the specialized literature on the use of GP for learning FRBSs.

2.1. Notation

A classification problem is considered with:

- A set of input variables $X = \{X_i / i = 1, \ldots, n_v\}$, where n_v is the number of features of the problem.
- A set of values for the target variable (class) $C = \{C^j / j = 1, \ldots, n_c\}$, where n_c is the number of different values for the class variable.
- A set of examples $E = \{e^h = (e^h_1, \ldots, e^h_{n_v}, C^h) / h = 1, \ldots, n_e\}$, where C^h is the class label for the sample e^h, and n_e is the number of examples.

2.2. Fuzzy rule-based classification systems

FRBCSs have been successfully applied to pattern classification problems [9,46], and the interest in their use arises from the fact that they provide a good platform for managing noisy, imprecise or incomplete information, which is often encountered in any human-cognition system.

An FRBCS is composed of a knowledge base and a fuzzy reasoning method. Both components are described in the next subsections.

2.2.1. The knowledge base

Composed of a rule base (RB) and a data base (DB):

- *Rule base*. Our method learns RBs containing the following type of rules: