Available online at www.sciencedirect.com
ScienceDirect
Data & Knowledge Engineering 60 (2007) 90–108
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

# Evolutionary stratified training set selection for extracting classification rules with trade off precision-interpretability ☆

José Ramón Cano [a,*], Francisco Herrera [b], Manuel Lozano [b]

[a] Department of Computer Science, University of Jaén, 23700 Linares, Jaén, Spain
[b] Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain

Available online 3 March 2006

**Abstract**

The generation of predictive models is a frequent task in data mining with the objective of generating highly precise and interpretable models. The data reduction is an interesting preprocessing approach that can allow us to obtain predictive models with these characteristics in large size data sets. In this paper, we analyze the rule classification model based on decision trees using a training selected set via evolutionary stratified instance selection. This method faces the scaling problem that appears in the evaluation of large size data sets, and the trade off interpretability-precision of the generated models.


*Keywords:* Training set selection; Interpretability; Precision; Evolutionary algorithms; Rule classification; Decision trees

## 1. Introduction

A basic process in data mining is the generation of representative models from data [1]. The models, depending on their domain of application, can be descriptive or predictive. The classical objective of predictive models is the accuracy or precision of the model. On the other hand, the interpretability of the model is an important aspect for the expert point of view, to understand the model behaviour [2]. In classical literature, we can find different proposals to measure the quality of the predictive models, as well as the precision, like simplicity, interpretability, etc. [3].

In this paper we are going to focus our attention on the predictive models based on classification rules for different size data sets, with the special interest in the trade off interpretability-precision [2]. Our models have been extracted from the data sets by means of *C4.5* algorithm [4].

☆ This work was supported by projects TIC2002-04036-C05-01 and TIN2005-08386-C05-01.
* Corresponding author.
*E-mail addresses:* jrcano@ujaen.es (J.R. Cano), herrera@decsai.ugr.es (F. Herrera), lozano@decsai.ugr.es (M. Lozano).

0169-023X/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2006.01.008

A possible way to improve the behaviour of predictive models, precision and interpretability, is to extract them from suitable reduced/selected training sets [5]. Training set selection can be developed using instance selection algorithms. The instance selection algorithms select representative instance subsets following a determined strategy, and they can improve the nearest neighbour rule prediction capabilities used in some cases as selection strategy objective [6,7]. In [5], Sebban et al. study the effect of the learning set size in decision trees performances. An important conclusion of this analysis is that the application of instance selection algorithms (and concretely, the *PSRCG* algorithm) can improve the generalization accuracy, reduce the decision tree size and tolerate the presence of noise, establishing a close link between instance selection and tree simplification.

Evolutionary algorithms (*EAs*) are adaptable methods based on natural evolution that can be applied to search and optimization problems [8–10]. The *EAs* offer interesting results when they are assessed on instance selection [11,12]. In this study, we use *CHC* algorithm as *EA* [13], considering its behaviour shown in [14]. The basic idea consists of combining in the fitness function both objectives, interpretability and precision [14,15].

The evaluation of instance selection algorithms over large size data sets makes them ineffective and inefficient. The effect produced by the size of data set in the algorithms is called scaling problem.

We focus our attention on evolutionary instance selection for large size data sets with the aim of extracting high precise-interpretable rules. To tackle the scaling problem we combine the stratification of the data sets with the instance selection over them [15]. The stratification reduces the original data set size, splitting it into strata where the selection will be applied. We analyze the selected training sets quality by means of the models (decision trees) extracted from them by means of *C4.5*, from the precision and interpretability perspectives. To compare the results we provide a statistical analysis using some statistical tests (ANOVA, Levene and Tamhane [16]).

The outline of the document is the following. In Section 2, we analyze the predictive models and their extraction using *C4.5*, presenting the measures considered to assess their behaviour. Section 3 describes the training set selection process and the drawbacks that the evaluation of very large data sets introduced in the instance selection algorithms. Section 4 presents the evolutionary stratified instance selection process applied to training set selection. Section 5 contains the experimental study developed, offering the methodology followed, the results and their analysis. Finally, in Section 6 we will point out some concluding results.

## 2. Predictive models: classification trees extraction with C4.5

The importance of decision trees and rules is that they are favoured techniques to build understandable models, a key point for the helpfulness of them and their application. A decision tree is a predictive model that can be viewed as a tree.

In this study we are going to extract the decision trees using the *C4.5* algorithm [4]. The models generated are complete and consistent, covering all the examples of the training set. The induction algorithm may over fit outliers, mislabelled, noisy data resulting in the inference of more structures than is justified by the training set. This situation is increased when the size of the learning set is large, so decision trees size is increased considerably [17–19]. The high size of the decision tree produces:

- Over fitting. In this case, the learned hypothesis is so closely related to the training examples that its generalization capabilities would be penalized [20].
- Low human interpretability. The highest size of the decision tree introduces the disadvantage of excessive complexity that can render it incomprehensible to experts [3,21].

To avoid this situation, there are several ways to simplify the decision tree, which were classified by Breslow and Aha in [22].

Among them, prune methods are more popular than the rest to be applied to the decision trees generated [23]. Prune methods can be classified in:

- Preprune methods. The prune process is developed during the tree generation. The prune determines the stopping condition for the branch specialization.
- Postprune methods. In this case, the prune process is applied after the tree construction. The prune removes nodes from bottom to top until a determined limit is reached.