Knowledge-Based Systems 54 (2013) 32–41

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

# A study on the application of instance selection techniques in genetic fuzzy rule-based classification systems: Accuracy-complexity trade-off

Michela Fazzolari [a,*], Bruno Giglio [b], Rafael Alcalá [a], Francesco Marcelloni [b], Francisco Herrera [a]

[a] Dept. of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain
[b] Dipartimento di Ingegneria dell'Informazione, University of Pisa, 56122 Pisa, Italy

ARTICLE INFO

*Article history:*
Available online 23 July 2013

*Keywords:*
Instance selection
Training set selection
Fuzzy rule-based classifiers
Evolutionary algorithms
Genetic fuzzy systems
Complexity-accuracy trade-off

ABSTRACT

In the framework of genetic fuzzy systems, the computational time required by genetic algorithms for generating fuzzy rule-based models from data increases considerably with the increase of the number of instances in the training set, mainly due to the fitness evaluation. Also, the amount of data typically affects the complexity of the resulting model: a higher number of instances generally induces the generation of models with a higher number of rules. Since the number of rules is considered one of the factors which affect the interpretability of the fuzzy rule-based models, large datasets generally bring to less interpretable models. Both these problems can be tackled and partially solved by reducing the number of instances before applying the evolutionary process. In the literature several algorithms of instance selection have been proposed for selecting instances without deteriorating the accuracy of the generated models.

The aim of this paper is to analyze the effectiveness of 36 training set selection methods when combined with genetic fuzzy rule-based classification systems. Using 37 datasets of different sizes we show that some of these methods can considerably help to reduce the computational time of the evolutionary process and to decrease the complexity of the fuzzy rule-based models with a very limited decrease of their accuracy with respect to the models generated by using the overall training set.



## 1. Introduction

Several real-world problems deal with classification tasks and several approaches have been proposed to manage them. Fuzzy Rule-Based Systems (FRBSs) have proved to be very effective as classifiers, especially when interpretable models are needed [1].

The components of an FRBS can be automatically derived from a set of data, for example by using Genetic Algorithms (GAs) [2,3]. In the last decades, GAs have been so extensively used to learn FRBSs that the term Genetic Fuzzy Systems (GFSs) has been coined to identify the hybridization between GAs and FRBSs [4–6].

The learning process of a GFS is strongly affected by the amount of instances used to generate the FRBS. A first problem is related to the computational time required by the fitness evaluation during the evolutionary process, since it is directly proportional to the number of instances. A second problem regards the complexity of the obtained models: in order to cover as much as possible instances of the dataset, the learning process tends to generate a high number of rules.

To reduce the amount of instances would speed up the learning process and possibly would lessen the complexity of the generated FRBSs. To this aim several approaches have been proposed in the literature. In particular, when considering medium and large datasets, the reduction can be obtained by applying techniques of Instance Selection (IS) [7–10], which aim to extract a small representative subset of instances from the initial set, by removing superfluous instances. The subset should maintain all the information of the original set, so that it can be used to generate classification models with the same accuracy as models generated by using the original set.

IS techniques can be grouped into two categories, depending on the aim pursued after obtaining the reduced set:

- Prototype Selection (PS) methods [11]: the reduced set is used by an instance-based classifier (for example K-NN) to classify new instances. Instance-based classifiers assume that unlabeled instances can be classified by relating them to the labeled instances, according to a certain similarity or distance function. The selected instances should provide the best trade-off between classification accuracy and reduction of the number of instances.

* Corresponding author. Tel.: +34 685643937.
*E-mail addresses:* fazzolari@decsai.ugr.es (M. Fazzolari), br1giglio@gmail.com (B. Giglio), alcala@decsai.ugr.es (R. Alcalá), f.marcelloni@iet.unipi.it (F. Marcelloni), herrera@decsai.ugr.es (F. Herrera).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.07.011

- Training Set Selection (TSS) methods [12,13]: the subset of instances is employed by a machine learning algorithm to build a predictive model (e.g. neural networks, FRBSs, decision trees, etc).

Several studies can be found in the literature for both PS and TSS. For example in [14], the authors perform IS by means of an evolutionary process. The quality of the reduced set is assessed by using the 1-NN classifier and a classification model constructed by the C4.5 algorithm. A comparison is carried out among evolutionary and non-evolutionary IS techniques, with respect to the classification accuracy and the instance reduction rate. This study has been subsequently extended in [15] and [12], where the concept of data stratification has been integrated in the framework with the aim of handling the scaling problem that appears when evaluating medium-large size datasets, and generating classification models with a good accuracy-interpretability trade-off.

A particular application of TSS is presented in [16]. Here, the authors focus on classification problems in presence of imbalanced dataset. Data are re-balanced by undersampling the instances belonging to the majority class thorough a TSS method. TSS is integrated in an evolutionary algorithm and the quality of the reduced set is evaluated by generating a classification model with the well-known C4.5 algorithm.

A recent approach can be found in [13], where the authors investigate the use of TSS to reduce the set of instances required by a Multi-Objective Evolutionary Algorithm (MOEA) to generate FRBSs for regression problems. The TSS is integrated in a co-evolutionary framework: cyclically, a single-objective GA selects a subset of instances which are used by the MOEA for generating the FRBSs. The GA maximizes an index that measures the quality of the reduced set of instances.

In this paper, we focus on the use of TSS techniques as pre-processing methods before applying a GA for generating Fuzzy Rule-Based Classification Systems (FRBCSs). We aim to investigate if TSS techniques can help to reduce the complexity of the generated FRBCSs, preserving or hopefully increasing their accuracy. A preliminary study discussed in [17], where a set of 20 small size datasets have been considered, has highlighted that a specific family of TSS methods is effective in achieving this objective.

We extend the study in [17] by considering also medium-large size datasets, which frequently appear in real-world problems. For these datasets, the number of rules of the generated FRBCSs can be quite large and therefore their interpretability can be quite low, blurred by the complexity. We have considered 36 TSS techniques and 17 additional medium-large size datasets. The TSS techniques have been applied to each dataset and reduced datasets have been obtained. Then, the reduced datasets have been used to generate FRBCSs by exploiting a recently developed GFS, named Fuzzy Association Rule-based Classification model for High-Dimensional problems (FARC-HD) [18], which has been demonstrated to be efficient when working with high-dimensional datasets, i.e. datasets with a high number of variables. Since medium-large size datasets usually involve also a high number variables, FARC-HD results to be particularly suitable for these datasets. The goal is to understand if TSS techniques are able to decrease the number of instances in a dataset without losing the information needed for allowing FARC-HD to generate FRBCSs that achieve high classification rate despite a low complexity and a low computational time.

A further study has been performed by considering the combination of small and medium-large size datasets. The aim is to obtain more reliable results when applying statistical tests and to investigate if there exist TSS techniques that can be effectively used with datasets of any size.

Finally, an analysis of the computational time required by the application of TSS techniques and by the execution of the GFS on the reduced datasets is reported, in order to evaluate if the selected subsets lead to a reduction in the time required by the GFS to generate classification models.

This contribution is organized as follows: Section 2 contains a brief description of the IS process and IS methods in general. The methods used in this study are listed, according to the taxonomy proposed in [10]. Section 3 describes the methodology used to carry out the experiments. Section 4 includes a brief overview of FARC-HD. In Section 5 the experimental framework is presented and the obtained results are examined and discussed.

## 2. Instance selection methods

Given a training set *TR*, the aim of an IS algorithm is to find a representative subset $S \subseteq TR$ of meaningful instances by removing superfluous instances (Fig. 1). The resulting subset will be used to build a classifier.

During the last years, more than fifty IS methods have been proposed in the literature and some reviews can be found in [7,8,10,19–22].

A comprehensive description of IS methods has been presented in a recent survey [10]. Here, a taxonomy based on the main characteristics of the methods has been proposed, analyzing advantages and drawbacks of each of them. Hereinafter, the categories of IS methods are briefly described and then the methods selected for the current study are listed.

### 2.1. Classification of instance selection methods

IS methods have been grouped in different categories according to their common properties and to the description presented in [10]:

- *Type of selection*: this characteristic is mainly influenced by the type of search strategy carried out by the IS algorithms, depending on the position of the instances to be retained with respect to the decision boundaries (border instances, central instances or some other set of instances). The techniques are condensation, edition and hybrid. Condensation methods try to compute a consistent subset *S* by removing unnecessary instances that will not affect the classification accuracy on the training set. Edition methods aim to remove noisy instances, allowing the classifier to increase its accuracy. Hybrid methods search for a subset in which both noisy and unnecessary instances are concurrently eliminated.
- *Direction of search*: when searching for a subset *S* of instances from the training set *TR*, there are a variety of directions in which the search can proceed: incremental, decremental, batch, mixed and fixed. Incremental methods start with an empty subset *S* and add instances during the selection process according to some criterion; on the contrary decremental methods start with the whole training set ($S = TR$) and remove instances

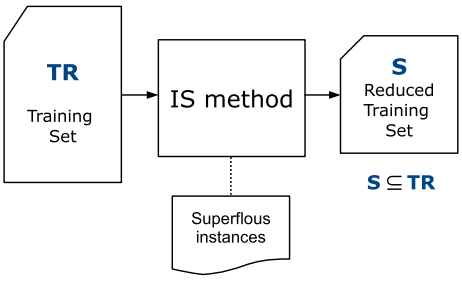


**Fig. 1.** Instance selection algorithm.