Information Sciences 329 (2016) 649–669

Contents lists available at ScienceDirect

Information Sciences

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

On the influence of feature selection in fuzzy rule-based regression model generation

Michela Antonelli, Pietro Ducange, Francesco Marcelloni*, Armando Segatori

Dipartimento di Ingegneria dell'Informazione, University of Pisa, Pisa 56122, Italy

ARTICLE INFO

Article history:
Received 23 December 2014
Revised 18 September 2015
Accepted 26 September 2015
Available online 9 October 2015

Keywords:
Fuzzy rule-based systems
Feature selection
Multi-objective evolutionary fuzzy rule-based systems
Fuzzy mutual information
Regression problems
High dimensional datasets

ABSTRACT

Fuzzy rule-based models have been extensively used in regression problems. Besides high accuracy, one of the most appreciated characteristics of these models is their interpretability, which is generally measured in terms of complexity. Complexity is affected by the number of features used for generating the model: the lower the number of features, the lower the complexity. Feature selection can therefore considerably contribute not only to speed up the learning process, but also to improve the interpretability of the final model. Nevertheless, a very few methods for selecting features before learning regression models have been proposed in the literature. In this paper, we focus on these methods, which perform feature selection as pre-processing step. In particular, we have adapted two state-of-the-art feature selection algorithms, namely NMIFS and CFS, originally proposed for classification, to deal with regression. Further, we have proposed FMIFS, a novel forward sequential feature selection approach, based on the minimal-redundancy-maximal-relevance criterion, which can manage directly fuzzy partitions. The relevance and the redundancy of a feature are measured in terms of, respectively, the fuzzy mutual information between the feature and the output variable, and the average fuzzy mutual information between the feature and the just selected features. The stopping criterion for the sequential selection is based on the average values of relevance and redundancy of the just selected features.

We have performed two experiments on twenty regression datasets. In the first experiment, we aimed to show the effectiveness of feature selection in fuzzy rule-based regression model generation by comparing the mean square errors achieved by the fuzzy rule-based models generated using all the features, and the features selected by FMIFS, NMIFS and CFS. In order to avoid possible biases related to the specific algorithm, we adopted the well-known Wang and Mendel algorithm for generating the fuzzy rule-based models. We present that the mean square errors obtained by models generated by using the features selected by FMIFS are on average similar to the values achieved by using all the features and lower than the ones obtained by employing the subset of features selected by NMIFS and CFS. In the second experiment, we intended to evaluate how feature selection can reduce the convergence time of the evolutionary fuzzy systems, which are probably the most effective fuzzy techniques for tackling regression problems. By using a state-of-the-art multi-objective evolutionary fuzzy system based on rule learning and membership function tuning, we show that the number

* Corresponding author. Tel.: +39 0502217678; fax: +39 0502217600.
E-mail addresses: michela.antonelli@iet.unipi.it (M. Antonelli), pietro.ducange@iet.unipi.it (P. Ducange), francesco.marcelloni@iet.unipi.it (F. Marcelloni), armando.segatori@for.unipi.it (A. Segatori).

http://dx.doi.org/10.1016/j.ins.2015.09.045

of evaluations can be considerably reduced when pre-processing the dataset by feature selection.

1. Introduction

A large number of real-world applications require to determine regression models from input–output pairs of observed samples. In this context, during the last decades, fuzzy rule-based systems and in particular Mamdani-type fuzzy rule-based systems (MFRBSs) [37] have been largely employed [1,2,6,14,16,30,39,41,47]. Indeed, they are recognized as universal approximators [13,34] and allow achieving accuracies comparable to other approaches. Furthermore, they have the capability of explaining how their outputs are generated from the input values. An MFRBS consists of a linguistic rule base (RB), a database (DB) containing the fuzzy sets associated with the linguistic terms used in the RB and a fuzzy logic inference engine. RB and DB compose the knowledge base (KB) of the MFRBS. Formally, an MFRBS is a mathematical model that, given an input vector, computes an output value, exploiting the knowledge coded in the RB and in the DB, and an inference process based on fuzzy logic.

The inputs of observed samples are typically described by a large number of features. Often, some of these features are irrelevant or redundant, thus making the most popular algorithms for learning regression models, including the ones for identifying the structure of MFRBSs, inefficient and inaccurate. For this reason, a lot of research activities have been devoted to design techniques for reducing dimensionality.

Dimensionality reduction is usually performed by two main approaches, namely feature extraction and feature selection [50]. Feature extraction [49] is a process that extracts a set of new features from the set of original features by means of a mapping function, with the aim of representing the original data more concisely. The main drawbacks of this process are the computational time needed to search for a suitable mapping function and the loss of interpretability of the final results. Indeed, typically no explicit and intuitive relation exists between the original and the new features and only the original features have a physical explanation.

On the contrary, feature selection [10] generates no new feature but selects an optimal set of the original features according to a certain criterion. The main aim of this selection process is to speed up the learning algorithms by reducing the dimensionality of the feature space. Typically, both the accuracy and the complexity of the learned models are also improved.

In general, feature selection algorithms can be classified into wrapper and filter methods [38]. In wrapper methods, the feature selector behaves as a wrapper around a specific learning algorithm that is used to evaluate the goodness of the feature subset [32]. In filter methods, the feature selection algorithm is employed to remove irrelevant and/or redundant features in a pre-processing phase, independently of any specific learning algorithm [44]. The filter approaches are in general computationally more efficient, while wrapper methods usually yield to better results. In this paper, we focus on a novel filter method for selecting features partitioned by fuzzy sets when generating MFRBSs for regression problems.

To evaluate the optimal subset of features, both wrapper and filter methods should test exhaustively all the possible combinations of the features. Since the number of these combinations increases factorially with the number of features, this approach becomes unfeasible in high dimensional problems. Thus, heuristic approaches are generally adopted. Sequential search algorithms are the most popular among the heuristic approaches: they add or subtract features at each iteration in order to find the optimal subset.

The most common sequential search schemes are the forward sequential selection (FSS) and the backward sequential selection (BSS) [35]. FSS starts from an empty set and, at each step, adds to this set the best feature among the unselected ones on the basis of an evaluation criterion. Steps are repeated until either all the original features are included in the set or a stopping criterion is reached. On the contrary, BSS starts with a set containing all the features and, at each step, removes from the set the feature that produces the maximal performance degradation. Steps are repeated until either a stopping condition is reached or only one feature remains in the set.

Both FSS and BSS need an evaluation criterion to assess the relevance of each feature to be added to or removed from the set. To this aim, several measures have been proposed in the literature: these measures can be classified into distance, information and dependency measures [7,17]. One of the most used information measures is the *Mutual Information* (MI). MI aims to quantify the mutual dependence between two variables and is defined as the difference between the sum of the entropy values of the two variables and their joint entropy value: MI is equal to zero when the variables are independent and increases with the increase of the dependence of one variable on the other.

A recent review on the use of MI for feature selection [45] states that MI has two main properties. First, it can measure any kind of relationship between random variables, including non-linear relationships. Second, it is invariant under transformations in the feature space that are invertible and differentiable, e.g., translations, rotations, and any transformation preserving the order of the original elements of the feature vectors. For these reasons, MI has been extensively used as evaluation measure for feature selection, even if mainly for classification problems [8,11,27,28,40].

For example, in [8,40] MI is employed to measure both the relevance and the redundancy of a feature in the framework of the minimal-redundancy-maximal-relevance criterion (mRMR) adopted in the feature selection process. In particular, the relevance is measured as the MI between the feature and the target class, and the redundancy is computed as the average MI between the