Neurocomputing 146 (2014) 125–136

Contents lists available at ScienceDirect

Neurocomputing

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

An experimental study on evolutionary fuzzy classifiers designed for managing imbalanced datasets

Michela Antonelli, Pietro Ducange, Francesco Marcelloni *

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

ARTICLE INFO

Article history:
Received 16 October 2013
Received in revised form
7 February 2014
Accepted 3 April 2014
Available online 11 July 2014

Keywords:
Genetic and evolutionary fuzzy systems
Fuzzy rule-based classifiers
Imbalanced datasets

ABSTRACT

In this paper, we show an experimental study on a set of evolutionary fuzzy classifiers (EFCs) purposely designed to manage imbalanced datasets. Three of these EFCs represent the state-of-the-art of the main approaches to the evolutionary generation of fuzzy rule-based systems for imbalanced dataset classification. The fourth EFC is an extension of a multi-objective evolutionary learning (MOEL) scheme we have recently proposed for managing imbalanced datasets: the rule base and the membership function parameters of a set of FRBCs are concurrently learned by optimizing the sensitivity, the specificity and the complexity.

By using non-parametric tests, we first compare the results obtained by the four EFCs in terms of area under the ROC curve. We show that our MOEL scheme outperforms two of the comparison algorithms and results to be statistically equivalent to the third. Further, the classifiers generated by our MOEL scheme are characterized by a lower number of rules than the ones generated by the other approaches.

To validate the effectiveness of our MOEL scheme in dealing with imbalanced datasets, we also compare our results with the ones achieved, after rebalancing the datasets, by two state-of-the-art algorithms, namely FURIA and FARC-HD, proposed for generating fuzzy rule-based classifiers for balanced datasets. We show that our MOEL scheme is statistically equivalent to FURIA, which is associated with the highest accuracy rank in the statistical tests. However, the rule bases generated by FURIA are characterized by a low interpretability.

Finally, we show that the results achieved by our MOEL scheme are statistically equivalent to the ones achieved by four state-of-the-art approaches, based on ensembles of non-fuzzy classifiers, appropriately designed for dealing with imbalanced datasets.

1. Introduction

In the framework of classification systems, the automatic identification of patterns belonging to two classes or categories is denoted as "binary classification". In most of the cases, these two classes correspond, respectively, to a normal condition (*negative* or *majority class*) and to an anomalous or alert condition (*positive* or *minority class*) of a certain phenomenon or system. A number of real world applications can be found in the context of binary classification, such as intrusion detection [1], medical diagnostics [2,3] and fault detection [4].

In order to automatically design an accurate model for classification tasks, we need an effective set of training examples. Usually, in the case of the previously cited applications of binary classification, the available data samples associated with the negative classes are much more abundant than the ones corresponding to the positive classes, i.e. imbalanced datasets have to be handled [5]. Indeed, luckily, anomalous conditions are in general rare and it is often difficult to collect a large number of representative examples.

The issue of learning classifiers when managing imbalanced datasets is still a hot topic in the machine learning research community [5,6]. Indeed, classical machine learning algorithms generate models that aim to maximize the percentage of correct classification or to minimize the percentage of classification error. If we use these algorithms with imbalanced datasets, the generated models usually are characterized by a bias towards the recognition of the majority class, making the minority class poorly recognized.

Currently, a large number of contributions regarding classification with imbalanced datasets [7–13] are still being published in the specialized literature. As stated also in [7], the proposed approaches can be organized into four main categories, namely methods acting at data level [14,15], approaches acting on the learning algorithms [16–18], cost sensitive methodologies which

* Corresponding author. Tel.: +39 0502217678; fax: +39 0502217600.
E-mail address: francesco.marcelloni@iet.unipi.it (F. Marcelloni).

http://dx.doi.org/10.1016/j.neucom.2014.04.070
0925-2312/

combine both algorithmic and data level approaches [19–21], and techniques which consider ensembles of classifiers [7,8,22–24].

Since 2008, fuzzy rule-based classifiers (FRBCs) have attracted the attention of researchers also in the framework of binary classification with imbalanced datasets [15,17,25]. FRBCs have proved to be very suitable for classification tasks: on the one side they can achieve very high accuracy and on the other side they can explain through the rules how the classification is performed. In specific application domains, this feature is very appealing, because it allows understanding which features and which values of these features actually permit to discriminate a class among the other classes. With the aim of managing imbalanced datasets, however, some appropriate strategies have to be exploited for designing the FRBC structure, that is, designing the rule base (RB) and the data base (DB).

The first study on the use of FRBCs with imbalanced datasets has been discussed in [15]. Here, the authors have analyzed the synergy between preprocessing mechanisms for re-balancing the training instances and specific algorithms aimed at generating the RB. The results achieved by three state-of-the-art learning algorithms for FRBCs, using both the original imbalanced training sets and their rebalanced versions, have been compared. The Synthetic Minority Oversampling Technique (SMOTE) [14] algorithm resulted to be the most performing one among different preprocessing mechanisms. In [25], authors have also studied the positive influence of re-balancing techniques when performing the genetic rule selection for hierarchical FRBCs. A similar study has been also conducted in [26], where the influence of the Adaptive Inference System for FRBCs in the framework of imbalanced datasets has been analyzed. Also in this case, the authors have demonstrated that the use of SMOTE allows improving the accuracies of the generated FRBCs.

Recently, a number of contributions have exploited evolutionary optimization algorithms for generating FRBCs suitable for imbalanced datasets. These approaches are denoted as evolutionary fuzzy classifiers (EFCs). EFCs are a specific category of the genetic and evolutionary fuzzy systems [27].

In this paper, we aim to perform an experimental comparison among a set of EFCs suitable for managing imbalanced datasets. We have selected, to the best of our knowledge, the three most recent EFCs, discussed in [11,28,29], which represent the last advances in the framework of EFCs for imbalanced datasets. Further, we have compared the results achieved by these EFCs with the ones achieved by an extended version of the method we proposed in [17].

We carry out an extensive statistical analysis by using non-parametric statistical tests. First, we compare the four EFCs using twenty-two highly imbalanced datasets. We show that, even though we do not re-balance the training set, the set of FRBCs generated by our MOEL scheme outperforms two comparison EFCs and achieves statistically equivalent results as the remaining EFC. As regards the complexity of the generated FRBCs, the classifiers generated by our MOEL scheme are always characterized by a lower number of rules than the ones obtained by the three comparison EFCs.

Further, we statistically compare the results achieved by our MOEL scheme with the ones achieved by two of the most interesting state-of-the-art methods for generating FRBCs, namely FURIA [30] and FARC-HD [31]. In this analysis, we consider forty-four imbalanced datasets and we rebalance the training sets for FURIA and FARC-HD, since these algorithms were not proposed for imbalanced datasets. FURIA results to be the algorithm associated with the highest performance rank in the tests. On the other hand, we show that the solutions generated by our MOEL scheme are statistically equivalent to the ones generated by FURIA.

Finally, by using non-parametric statistical tests, we also compare the results achieved by the FRBCs generated by our MOEL scheme with the ones achieved by four state-of-the art approaches, based on boosting and bagging strategies, for generating ensembles of classifiers for imbalanced datasets. We show that the results are statistically equivalent, thus proving that our MOEL scheme produces classifiers which are not only interpretable, but also very efficient.

In a nutshell, the main contributions of this paper are:

- A new version of an MOEL scheme for generating highly interpretable FRBCs for imbalanced datasets.
- A statistical comparison, in terms of accuracy and interpretability, among the proposed MOEL scheme and three EFCs recently proposed for dealing with imbalanced datasets.
- A statistical comparison among the results achieved by the FRBCs generated by our MOEL and by two state-of-the-art learning algorithms.
- A statistical comparison among the results achieved by the FRBCs generated by our MOEL and the ones achieved by ensembles of non-fuzzy classifiers purposely designed for dealing with imbalanced datasets.

The paper is organized as follows: Section 2 introduces some preliminary concepts and notations regarding the FRBCs. In Section 3 we discuss the four EFCs considered in the experimental comparison. Section 4 shows the different statistical analyses and Section 5 draws some conclusion.

2. Fuzzy rule-based classifiers

Pattern classification consists of assigning a class C_k from a predefined set $C=\{C_1,\ldots,C_K\}$ of classes to an unlabeled pattern. We consider a pattern as an F-dimensional point in a feature space $\Re^F$. Let $\mathbf{X}=\{X_1,\ldots,X_F\}$ be the set of input variables and U_f, $f=1,\ldots,F$, be the universe of discourse of the fth variable. Let $P_f=\{A_{f,1},\ldots,A_{f,T_f}\}$ be a fuzzy partition of T_f fuzzy sets on variable X_f. T_f defines the granularity of the partition P_f. The DB of an FRBC is the set of parameters which describe the partitions P_f of each input variable. The RB contains a set of M rules usually expressed as

$$R_m: \textbf{IF } X_1 \textbf{ is } A_{1j_{m,1}} \textbf{ AND}\ldots\textbf{AND } X_F \textbf{ is } A_{Fj_{m,F}} \quad \textbf{THEN } Y \textbf{ is } C_{j_m} \text{ with } RW_m \tag{1}$$

where Y is the classifier output, C_{j_m} is the class label associated with the mth rule, $j_{m,f}\in[1,T_f]$, $f=1,\ldots,F$, identifies the index of the fuzzy set (among the T_f linguistic terms of partition P_f), which has been selected for X_f in rule R_m. RW_m is the rule weight, i.e. a certainty degree of the classification in the class C_{j_m} for a pattern belonging to the fuzzy subspace delimited by the antecedent of the rule R_m.

Let $(\mathbf{x}_t,y_t)$ be the tth input–output pair, with $\mathbf{x}_t=[x_{t,1}\ldots,x_{t,F}]\in\Re^F$ and $y_t\in C$. The strength of activation (*matching degree* of the rule with the input) of the rule R_m is calculated as

$$w_m(\mathbf{x}_t)=\prod_{f=1}^{F} A_{fj_{m,f}}(x_{t,f}), \tag{2}$$

where $A_{fj_{m,f}}(x)$ is the membership function (MF) associated with the fuzzy set $A_{fj_{m,f}}$.

The *association degree* with the class C_{j_m} is calculated as

$$h_m(\mathbf{x}_t)=w_m(\mathbf{x}_t)\cdot RW_m \tag{3}$$

Two different definitions of the rule weight RW_m can be commonly found in the literature [32,33]:

1. *The certainty factor*:

$$CF_m=\frac{\sum_{\mathbf{x}_t\in C_{j_m}} w_m(\mathbf{x}_t)}{\sum_{t=1}^{N} w_m(\mathbf{x}_t)}. \tag{4}$$