Expert Systems with Applications 36 (2009) 9805–9812

Contents lists available at ScienceDirect

Expert Systems with Applications

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

On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets ☆

Alberto Fernández [a,*], María José del Jesus [b], Francisco Herrera [a]

[a] Department of Computer Science and A.I., University of Granada, Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain
[b] Department of Computer Science, University of Jaén, Spain

ARTICLE INFO

Keywords:
Fuzzy rule-based classification systems
Inference mechanism
Parametric conjunction operators
Genetic fuzzy systems
Imbalanced data-sets

ABSTRACT

Classification with imbalanced data-sets supposes a new challenge for researches in the framework of data mining. This problem appears when the number of examples that represents one of the classes of the data-set (usually the concept of interest) is much lower than that of the other classes. In this manner, the learning model must be adapted to this situation, which is very common in real applications.

In this paper, we will work with fuzzy rule based classification systems using a preprocessing step in order to deal with the class imbalance. Our aim is to analyze the behaviour of fuzzy rule based classification systems in the framework of imbalanced data-sets by means of the application of an adaptive inference system with parametric conjunction operators.

Our results shows empirically that the use of the this parametric conjunction operators implies a higher performance for all data-sets with different imbalanced ratios.

1. Introduction

Fuzzy rule based classification systems (FRBCSs) (Ishibuchi, Nakashima, & Nii, 2004) are a very useful tool in the ambit of machine learning, since they provide an interpretable model for the end user. There are many real applications in which the FRBCS have been employed, including anomaly intrusion detection (Tsang, Kwong, & Wang, 2007), cloud cover estimation from satellite imagery (Ghosh, Pal, & Das, 2006) and image processing (Nakashima, Schaefer, Yokota, & Ishibuchi, 2007). In most of these areas the data used is highly skewed, i.e. the number of instances of one class is much lower than the instances of the other classes. This situation is known as the imbalanced data-set problem, and it has been recently identified as one important problem in data mining (Chawla, Japkowicz, & Kolcz, 2004).

Most learning algorithms obtain a high predictive accuracy over the majority class, but predict poorly over the minority class (Weiss, 2004). Furthermore, the examples in the minority class can be treated as noise and they might are completely ignored by the classifier. In fact, there are studies that show that most classification methods lose their classification ability when dealing with imbalanced data (Japkowicz & Stephen, 2002; Phua, Alahakoon, & Lee, 2004). In this manner, many recent studies are focused on developing new approaches in this area (Hong, Chen, & Harris, 2007; Lee, Tsai, Wu, & Yang, 2008; Su, Chen, & Yih, 2006).

The use of the appropriate conjunction connectors in the Inference System can improve the fuzzy system behaviour by using parametrized expressions, while maintaining the original interpretability associated to fuzzy systems (Crockett, Bandar, Fowdar, & O'Shea, 2006; Crockett, Bandar, Mclean, & O'Shea, 2006; Wu & Mendel, 2004). This approach is usually called Adaptive Inference System (AIS) and it has shown very good results in fuzzy modelling (Alcalá-Fdez, Herrera, Márquez, & Peregrín, 2007; Márquez, Peregrín, & Herrera, 2007).

Our aim in this paper is to analyze the influence of the AIS for FRBCSs in the framework of imbalanced data-sets. We start from the analysis performed in Fernández, García, del Jesus, and Herrera (2008), where we studied different configurations for FRBCS in order to determine the most suitable model for imbalanced data-sets. Furthermore, we showed the necessity to apply a re-sampling procedure; specifically, we found a very good behaviour in the case of the "Synthetic Minority Over-Sampling Technique" (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

We will present a postprocessing study on the tuning of parameters with a previously established Rule Base (RB), using Genetic Algorithms (GAs) as a tool to evolve the connector parameters. We will develop an experimental study with 33 data-sets from UCI repository with different imbalance ratios. Data-sets with more than two classes have been modified by taking one against the others or by contrasting one class with another. To evaluate

☆ Supported by the Spanish Ministry of Science and Technology under Projects TIN-2005-08386-C05-01 and TIN-2005-08386-C05-03.

* Corresponding author. Tel.: +34 958 240598; fax: +34 958 243317.
E-mail addresses: alberto@decsai.ugr.es (A. Fernández), mjjesus@ujaen.es (M.J. del Jesus), herrera@decsai.ugr.es (F. Herrera).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.02.048

our results we have applied the geometric mean metric (Barandela, Sánchez, García, & Rangel, 2003; Kubat, Holte, & Matwin, 1998) which aims to maximize the accuracy of both classes. We have also made use of some non-parametric tests (Demšar, 2006; García, Fernández, Luengo, & Herrera, in press) with the aim to show the significance in the performance improvements obtained with the AIS model.

In order to do that, the paper is organized as follows: Section 2 presents an introduction on the class imbalance problem, including the description of the problem, proposed solutions, and proper measures for evaluating classification performance in the presence of the imbalance data-set problem. In Section 3, we describe the fuzzy rule learning methodology used in this study, the Chi et al. rule generation method (Chi, Yan, & Pham, 1996), and introduces the AIS with the parametric conjunction operators and the evolutionary algorithm that tunes these parameters. In Section 4, we include our experimental analysis in imbalanced data-sets with different degrees of imbalance. Finally, in Section 5 some concluding remarks are pointed out.

2. Imbalanced data-sets in classification

In this section, we will first introduce the problem of imbalanced data-sets. Then we will describe the preprocessing technique we have applied in order to deal with the imbalanced data-sets: the SMOTE algorithm. Finally, we will present the evaluation metrics for this kind of classification problem.

2.1. The problem of imbalanced data-sets

Learning from imbalanced data is an important topic that has recently appeared in the machine learning community. When treating with imbalanced data-sets, one or more classes might be represented by a large number of examples while the others are represented by only a few.

We focus on the two class imbalanced data-sets, where there is only one positive and one negative class. We consider the positive class as the one with the lowest number of examples and the negative class the one with the highest number of examples. Furthermore, in this work we use the imbalance ratio (IR) (Orriols-Puig & Bernadó-Mansilla, 2009), defined as the ratio of the number of instances of the majority class and the minority class, to organize the different data-sets according to their IR.

The problem of imbalanced data-sets is extremely significant because it is implicit in most real world applications, such as fraud detection (Fawcett & Provost, 1997), text classification (Tan, 2005), risk management (Huang, Hung, & Jiau, 2006), medical diagnosis (Mazurowski et al., 2008) and classification of weld flaws (Liao, 2008) among others.

In classification, this problem (also named the "class imbalance problem") will cause a bias on the training of classifiers and will result in the lower sensitivity of detecting the minority class examples. In fact, the main handicap on imbalanced data-sets is the overlapping between the examples of the positive and the negative class, because of the difficulty of most learning algorithms to detect those small disjuncts (Weiss & Provost, 2003). This fact is depicted in Fig. 1.

For this reason, a large number of approaches have been previously proposed to deal with the class imbalance problem. These approaches can be categorized into two groups: the internal approaches that create new algorithms or modify existing ones to take the class imbalance problem into consideration (Barandela et al., 2003; Hung & Huang, 2008; Xu, Chow, & Taylor, 2007) and external approaches that preprocess the data in order to diminish the effect cause by their class imbalance (Batista, Prati, & Monard, 2004; Estabrooks, Jo, & Japkowicz, 2004). Furthermore, cost-sensitive learning solutions incorporating both the data and algorithmic level approaches assume higher misclassification costs with samples in the minority class and seek to minimize the high cost errors (Domingos, 1999; Sun, Kamel, Wong, & Wang, 2007).

The internal approaches have the disadvantage of being algorithm specific, while external approaches are independent of the classifier used and are, for this reason, more versatile. Furthermore, in our previous work on this topic (Fernández et al., 2008) we analyzed the cooperation of some preprocessing methods with FRBCSs, showing a good behaviour for the oversampling methods, specially in the case of the SMOTE methodology (Chawla et al., 2002).

According to this, we will employ in this paper the SMOTE algorithm in order to deal with the problem of imbalanced data-sets. This method is detailed in the next subsection.

2.2. Preprocessing imbalanced data-sets. The SMOTE algorithm

As mentioned before, applying a preprocessing step in order to balance the class distribution is a positive solution to the imbalance data-set problem (Batista et al., 2004). Specifically, in this work we have chosen an oversampling method which is a reference in this area: the SMOTE algorithm (Chawla et al., 2002).

In this approach, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours. Depending upon the amount of over-sampling required, neighbours from the k nearest neighbours are randomly chosen. This process is illustrated in Fig. 2, where x_i is the selected

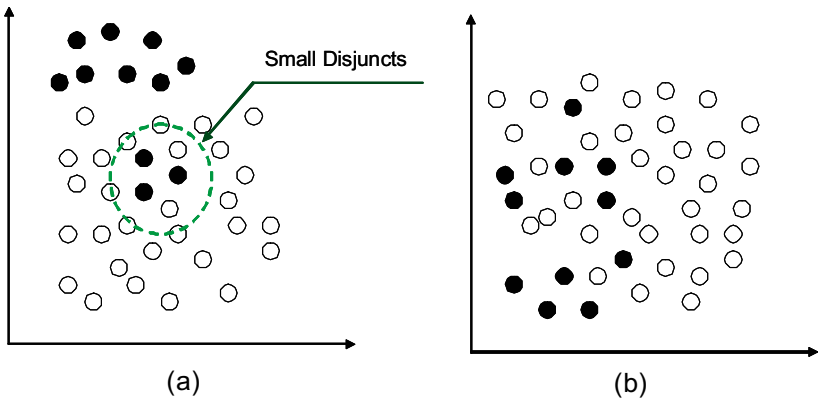

Fig. 1. Example of the imbalance between classes: (a) small disjuncts; (b) overlapping between classes.