Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

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

Fuzzy rough classifiers for class imbalanced multi-instance data

Sarah Vluymans [a,b,*], Dánel SánchezTarragó [c], Yvan Saeys [b,e], Chris Cornelis [a,d], Francisco Herrera [d,f]

[a] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
[b] VIB Inflammation Research Center, Ghent, Belgium
[c] Department of Computer Science, Central University of Las Villas, Cuba
[d] Department of Computer Science and Artificial Intelligence, University of Granada, Spain
[e] Department of Respiratory Medicine, Ghent University, Belgium
[f] Faculty of Computing and Information Technology – North Jeddah, King Abdulaziz University, Saudi Arabia

ARTICLE INFO

Article history:
Received 29 April 2015
Received in revised form
22 October 2015
Accepted 3 December 2015

Keywords:
Multi-instance learning
Fuzzy rough set theory
Imbalanced data

ABSTRACT

In multi-instance learning, each learning object consists of many descriptive instances. In the corresponding classification problems, each training object is labeled, but its constituent instances are not. The classification objective is to predict the class label of unseen objects. As in traditional single-instance classification, when the class sizes of multi-instance data are imbalanced, classification is degraded. Many multi-instance classifiers have been proposed, but few take into account the possibility of class imbalance, which causes them to fail in this situation. In this paper, we propose a new type of classifier that embodies a solution to the multi-instance class imbalance problem. Our proposal relies on the use of fuzzy rough set theory. We present two families of classifiers respectively based on information extracted at bag-level and at instance-level. We experimentally show that our algorithms outperform state-of-the-art solutions to multi-instance imbalanced data classification, evaluated by the popular metrics AUC and geometric mean.

1. Introduction

In machine learning, multi-instance learning (MIL, [1]) is a generalization of the traditional single-instance attribute-value approach. While in the single-instance setting each learning object has a single descriptive vector, in MIL each learning object is composed of many vectors, although all vectors relate to the same set of descriptive attributes. In MIL jargon, a learning object is called a *bag* and every descriptive vector is an *instance*. As in traditional learning, classification is one of the most important tasks of MIL. Only the bags, and not their instances, have class labels. The objective of multi-instance classification is to predict the class label of unseen bags using a model built from a training set.

Many multi-instance classification algorithms have been proposed. However, most have been designed and evaluated considering data with balanced classes. The class imbalance problem affects both single and multi-instance learning. The problem occurs when at least one of the classes has a disproportionately small size compared to the other classes. In these cases, classifiers tend to make more errors on small classes and may even ignore them completely, although small classes are usually more of interest. This problem has received much attention in single-instance learning (e.g. [2–4]), but has barely been studied in MIL. To our knowledge, existing solutions in the MIL scenario are limited to the contributions of [5–7], that consider both preprocessing techniques to modify the class imbalance as well as a set of cost-sensitive boosting algorithms.

The *K*-nearest neighbor classifier (KNN, [8]) is one of the most popular learning algorithms [9]. It assigns an unseen object to the decision class most frequent among the *K* closest training objects to the unseen object. Over the years, several improvements and adaptations of KNN have been proposed. One successful modification is the fuzzy rough nearest neighbor classifier (FRNN, [10]), which introduces fuzzy rough set theory into KNN. Fuzzy rough set theory [11] is a framework to model vague (fuzzy) and incomplete (rough) information, by introducing fuzzy set theory [12] into rough set theory [13]. Rough sets approximate a concept by means of a lower and upper approximation. The former contains elements which *certainly* belong to the concept, while the latter consists of elements *possibly* belonging to it. The integration of fuzzy set theory in rough sets allows for a more flexible instance similarity measure and graded membership degrees of elements to the approximations. Concretely, similarity between instances is

* Corresponding author at: Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium. Tel.: +32 9 264 47 57; fax: +32 9 264 49 95.
E-mail address: Sarah.Vluymans@UGent.be (S. Vluymans).

http://dx.doi.org/10.1016/j.patcog.2015.12.002
0031-3203/

measured by a fuzzy relation and the constructed concept approximations are fuzzy sets. Fuzzy rough set theory has been used successfully in many single-instance machine learning applications, including classification (e.g. [14–16]). In the classification framework, it allows to model a degree of membership of elements to approximations of the decision classes. The FRNN method uses the unseen object's K nearest neighbors to construct the lower and upper approximation of each decision class and then computes the membership of the unseen object to this approximations. The hybridization between fuzzy rough sets and KNN results in a classifier more robust to vague and incomplete information [10]. Recently, a further improvement on FRNN was introduced in [17] by using ordered weighted average operators (OWA, [18]) and class dependent weight vectors. This method, called IFROWANN, was specifically designed to handle class imbalance and proved to be very effective in single-instance class imbalanced classification.

In this paper, we introduce a new type of multi-instance classifiers, based on the IFROWANN method, which inherently contain a solution to the class imbalance problem in multi-instance classification. The more complex nature of multi-instance data prompts us to propose two families of classifiers: (1) bag-based fuzzy rough classifiers which rely on relationships between bags, considering the bag as a whole and (2) instance-based fuzzy rough classifiers based on affinities that instances themselves have with classes. In our experimental study, we show that the proposed fuzzy rough nearest neighbors classifiers outperform state-of-the-art solutions to class imbalanced multi-instance classification.

The remainder of this paper is structured as follows. We set out in Section 2 with a specification of multi-instance classification and the class imbalance problem and review previous proposals. In Section 3, we recall the IFROWANN method from [17], which forms the inspiration for our proposal. Section 4 considers multi-instance classification and introduces our proposed method dealing with class imbalance in this situation. The experimental evaluation of our proposal is conducted in Section 5. We conclude the paper and lay out future research paths in Section 6.

2. The class imbalance problem in multi-instance classification

In this preliminary section, we specify the formal definition of multi-instance classifiers. We recall the class imbalance problem and how it has been dealt with in single-instance classification. Finally, we review the efforts made to handle class imbalance in multi-instance classification problems.

2.1. Multi-instance classification

MIL was introduced in [1] in a study of drug activity prediction based on multiple molecular conformations. Since then, it has attracted a considerable amount of attention due to its ability to model data ambiguity and the link it forms between classical attribute-value learning and relational learning [19]. MIL has mainly been used in applications related to image recognition (e.g. [20–22]). Other important application domains include bioinformatics (e.g. [23–25]), text classification and web mining (e.g. [26–29]) and computer-aided medical diagnosis and medical imaging (e.g. [30–32]).

Given the instance space $\mathcal{X}$ and the label set $\mathcal{Y}$, a bag X_i is a multiset of instances $\{x_{i1}, x_{i2}, \ldots, x_{in_i}\}$ with $x_{ij} \in \mathcal{X}$. The number n_i denotes the cardinality of X_i. Note that we use lowercase letters to denote instances and uppercase letters for bags. Each bag is paired with a label $y_i \in \mathcal{Y}$. Considering training data $T = \{(X_1, y_1), \ldots, (X_m, y_m)\}$, we formally define a multi-instance classifier $h(X)$ as an approximate model to the real function $f : \mathbb{N}^{\mathcal{X}} \rightarrow \mathcal{Y}$, where $\mathbb{N}^{\mathcal{X}}$ is the set of all multisets consisting of elements from $\mathcal{X}$, that is, the set of all possible bags.

Multi-instance datasets traditionally consist of two classes, one positive and one negative. Several hypotheses exist to decide when a bag of instances can be considered as positive [33]. The standard multi-instance hypothesis assumes that a bag is positive when at least one of its instances is positive. If not, the bag is negative. An alternative is the threshold based assumption, which states that the number of positive instances in a bag should exceed a given threshold before the bag can be considered positive. Overall, it implies that it is a too naive approach to assume that all instances in a positive bag can be labeled as positive and all instances in a negative bag as negative. This was for instance shown in the review of [34]. We will take this into account in the development of our classifiers.

2.2. The class imbalance problem

In single-instance classification, class imbalance occurs when the elements in the dataset are unequally distributed among the classes. The main focus has been on binary imbalanced problems, where elements of the *majority* class outnumber those of the *minority* class. The elements of the majority class are traditionally denoted as *negative* and those of the minority class as *positive*. This coincides with the fact that the minority or positive class is usually the class of interest (e.g. [35]).

There are three main types of solutions used in traditional classification to deal with class imbalance. Firstly, solutions at the data level (resampling methods) perform undersampling of the majority class, oversampling of the minority class or a combination of both in order to balance the number of examples in the two classes. Secondly, there exist solutions at the algorithmic level, in which heuristics are incorporated into classic algorithms to handle class imbalance, for example, by adjusting probabilities and weights to favor the positive class. Of particular interest in this type of solutions are cost-sensitive methods [36], which assign higher costs to the misclassification of positive examples, while aiming to minimize the overall classification cost. The third group consists of ensemble solutions, that introduce one of the above solutions (e.g. resampling or cost-sensitivity) in an ensemble algorithm to create a layer of abstraction effectively separating the method used to counteract the class imbalance from the base classifier used in the ensemble.

Although many solutions to class imbalance have been proposed in traditional classification, they are not directly applicable to multi-instance scenarios due to the structural differences in the datasets. In particular, multi-instance data consists of two levels: instances and bags. The grouping of instances in bags is essential additional information that should be taken into account. Furthermore, the actual labeled data samples (bags) in multi-instance data are far more complex than those in single-instance data (instances) and the single-instance solutions simply cannot process them. Class imbalance appears in multi-instance problems like text, web, and image applications [5–7], but it has been little addressed in the literature so far. In multi-instance classification, class imbalance presents itself as an unequal distribution of the bags among the classes, that is, we encounter a larger number of negative bags compared to positive ones. The imbalance ratio (IR) expresses the degree of class imbalance and is defined, for a two-class dataset, as the ratio of the number of negative over the number of positive bags, i.e., $\mathrm{IR} = |N|/|P|$, where P and N are the positive and negative classes respectively. While multi-instance classification is not limited to two-class problems, this setting has been the main focus of researchers in the field [34]. Moreover, class imbalance has also been mainly studied for binary problems in single-instance classification. All previous proposals dealing

Please cite this article as: S. Vluymans, et al., Fuzzy rough classifiers for class imbalanced multi-instance data, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.12.002