Information Sciences 474 (2019) 187–204

Contents lists available at ScienceDirect

Information Sciences

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

Chain based sampling for monotonic imbalanced classification

Sergio González[a,*], Salvador García[a], Sheng-Tun Li[b], Francisco Herrera[a,c]

[a] Department of Computer Science and Artificial Intelligence, University of Granada, Granada 18071, Spain
[b] Department of Industrial and Information Management, National Cheng Kung University, Tainan 701, Taiwan ROC
[c] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 19 March 2018
Revised 18 July 2018
Accepted 26 September 2018
Available online 27 September 2018

Keywords:
Monotonic classification
Imbalanced classification
Sampling techniques
Preprocessing
Data mining

ABSTRACT

Classification with monotonic constraints arises from some ordinal real-life problems. In these real-life problems, it is common to find a big difference in the number of instances representing middle-ranked classes and the top classes, because the former usually represents the average or the normality, while the latter are the exceptional and uncommon. This is known as class imbalance problem, and it deteriorates the learning of those underrepresented classes. However, the traditional solutions cannot be applied to applications that require monotonic restrictions to be asserted. Since these were not designed to consider monotonic constraints, they compromise the monotonicity of the data-sets and the performance of the monotonic classifiers. In this paper, we propose a set of new sampling techniques to mitigate the imbalanced class distribution and, at the same time, maintain the monotonicity of the data-sets. These methods perform the sampling inside monotonic chains, sets of comparable instances, in order to preserve them and, as a result, the monotonicity. Five different approaches are redesigned based on famous under- and oversampling techniques and their standard and ordinal versions are compared with outstanding results.

1. Introduction

Ranking and evaluation of assets or even individuals are intrinsic characteristics of human nature. Hence, the presence of ordinal variables is common in tons of real-life data-sets. Credit rating [13,48], house ranking [50] and employee evaluation [6,37] are good examples of their presence in present-day applications.

These problems aim to determine the most valuable items according to their virtues, i. e. classification into ordinal labels according to ordinal attributes. Additionally, these applications usually require a monotonic restriction between the inputs and the class. That is, the class prediction of an individual should not decrease with a better value for a certain variable, fixing the remainder. Otherwise, an unfair evaluation of the individuals can be made. These classification problems with prior knowledge of the order relations between attributes and the class are known as classification with monotonicity constraints or monotonic classification [4]. Failure to respect these constraints are referred to as violations of monotonicity and must be avoided in the class decision of new samples.

When dealing with monotonic classification problems [4], we look for those examples that belong to the most remarkable class, with a higher value. It is reasonable to have fewer samples of really good and remarkable individuals than those

[*] Corresponding author.
E-mail addresses: sergiogvz@decsai.ugr.es (S. González), salvagl@decsai.ugr.es (S. García), stli@mail.ncku.edu.tw (S.-T. Li), herrera@decsai.ugr.es (F. Herrera).

https://doi.org/10.1016/j.ins.2018.09.062
0020-0255/© 2018 Elsevier Inc. All rights reserved.

considered normal or average. For example, in the evaluation of future employees of a company, there will probably be fewer “excellent candidates” than “average candidates.”

This difference in the number of representatives between classes has proven to cause a great loss of prediction accuracy in the minority classes [38,44]. This issue is known as a class imbalance problem or imbalanced classification. Multiple real-life applications present this problem, even those from non-standard classifications, such as Monotonic Classification [4,6,48] or Multi-task learning [36,37]. The majority of the monotonic problems considered in the literature suffers due to this issue. Therefore, the imbalance class distribution must be approached in the scope of monotonic classification.

Traditionally data level approaches [44] have been well accepted because they allow the use of a standard classifier after balancing the skewed training sets by under-/over- sampling. However, these techniques also have their own drawbacks. When using under-sampling, there is the risk of losing relevant information from the treated class. On the other hand, over-sampling can introduce noisy instances.

These approaches are not designed for monotonic classification [4] and do not take monotonic constraints into consideration. Due to this lack of awareness of monotonicity [4], these preprocessing techniques can severely deteriorate the monotonicity of the data-sets and reduce the performance of the classifiers. For example, the possible noisy instances generated by over-sampling could mean a greater damage in monotonic classification, because they may increase the number of monotonicity violations in the data-sets. The under-sampling techniques could remove important instances that determine the limit of the classes in term of monotonicity.

Therefore, new sampling approaches must be designed considering the monotonicity constraints. We propose new sampling techniques based on monotonic chains. In monotonic classification, a chain [35] is a set of comparable instances, that is, they can be sorted. These are very important assets of the classification carried out relevant methods such as KNN [16] and OSDL [34,35], because they determine the possible classes without monotonic violations for new instances. Our techniques perform the sampling using these chains and preserving, as much as possible, the monotonicity of the data-sets. Additionally, these methods take monotonic noise into consideration, in order to avoid instances that violate monotonicity during the sampling process. These differences with the traditional methods reduce the deterioration of monotonicity of sampled data-sets and maintain the improvement of the accuracy for minority classes.

To do so, we have put together a new scheme for applying both under- and over-sampling to monotonic imbalanced data-sets. This scheme consists of several good practices, related to the influence of monotonic violations and chains on sampling, that can be extended to the almost all the sampling techniques in the literature. This scheme has been implemented in five famous under- and over-sampling approaches of the State-of-the-Art of imbalanced classification: Random Under-Sampling (RUS), Random Over-Sampling (ROS), Synthetic Minority Oversampling TEchnique (SMOTE) [10], ADAptive SYNthetic sampling approach (ADASYN) [27] and Majority Weighted Minority Oversampling TEchnique (MWMOTE) [2].

Throughout this paper, two different empirical studies are carried out with exactly the same experimental framework. The first experiments test the selected sampling techniques in their standard and ordinal versions using 8 monotonic imbalanced sets which are very common in the literature. The majority are multi-class and can be considered highly imbalanced problems. The original and sampled data-sets are classified by five well-known classifiers. Two evaluation metrics are used: Macro Average Arithmetic (MAvA) [47] evaluates the prediction capability in multi-class imbalanced scenarios and Non-Monotonic Index (NMI) [4] determines the monotonicity of data-sets and predictions. The obtained results show empirically the deterioration of the monotonicity degree in data-sets caused by standard and ordinal sampling approaches.

Then, a second experimental study is performed following the same framework to analyze the behavior of new monotonic sampling techniques. The different predictions obtained are compared in terms of multi-class accuracy and monotonicity. The comparison shows the capacity of monotonicity preservation of the monotonic sampling techniques over the standard ones. The outcomes are corroborated by the use of non-parametric statistical tests: Friedman ranking test [20,25] and Bayesian Sign test [5].

This paper is organized as follows. In Section 2, we present the problems approached and their solutions: classification with monotonic constraints and class imbalance problem. Section 3 is devoted to setting up the bases to adapt sampling approaches to monotonic scenarios and explain in detail the chain based sampling techniques. In Section 4, the experimental framework used in the different empirical studies is presented. Section 5 recalls two experimental studies: an analysis on the impact of standard and ordinal sampling on monotonic classification and a comparison of the results achieved by monotonic sampling. Finally, in Section 6, the main conclusions of this study are given.

2. Background

In this section, we introduce the background knowledge of the problems addressed in this paper.

2.1. Monotonic classification

Monotonic classification, just as ordinal regression and/or classification, aims to predict an ordinal class label y for new sample x with ordinal attributes with the help of a labeled set, i.e. f: $x \rightarrow y$. In both problems, the classes $\mathcal{Y}$ are categories $\mathcal{Y} = \{\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_C\}$ with a problem imposed arrangement $\mathcal{L}_1 \prec \mathcal{L}_2 \prec \ldots \prec \mathcal{L}_C$. However, there is a big difference between both problems. Ordinal classification just focuses on minimizing the errors of predicted and real labels. Monotonic classification imposes monotonicity constraints between the input variables and predicted labels, that is, every instance x'