MLSMOTE: Approaching Imbalanced Multilabel Learning Through Synthetic Instance Generation

Francisco Charte Ojeda[1], Antonio J. Rivera Rivas[2], María J. del Jesus Díaz[2], Francisco Herrera Triguero[1]

[1]Dept. of Computer Science and Artificial Intelligence, E.T.S.I.I.T., University of Granada, Granada 18071, Spain.
[2]Dept. of Computer Science, E.P.S., University of Jaén, Jaén 23071, Spain.
francisco@fcharte.com, arivera@ujaen.es, mjjesus@ujaen.es, herrera@decsai.ugr.es

Abstract. This is a summary of our article published in Knowledge-Based Systems [1] to be part of the MultiConference CAEPIA'16 Key-Works.

Keywords: Multilabel classification, Imbalanced learning, Resampling, Instance generation, SMOTE

1 Summary

The learning of classification models from imbalanced data is usually a challenging task, since most algorithms produce classifiers that tend to be biased towards the majority class. The higher is the imbalance level, the greater the likelihood of inducing this bias. Commonly, the classifier would achieve a good performance by simply predicting the majority class for all data patterns. As a consequence the minority class, which frequently is the most interesting to the researchers, suffers from miss-classification errors.

Resampling methods are among the most popular approaches to face imbalanced learning, since they provide a classifier independent mechanism to do so. The goal of these methods is to balance the distribution of classes, either by generating new samples associated to the minority class or by removing those linked to the majority class. Random oversampling and random undersampling are basic ways of performing this work. One the of the most successful ways to achieve this balanced distribution consists in generating synthetic instances, as proposed in the SMOTE [2] (*Synthetic Minority Over-sampling Technique*) algorithm.

Imbalance is a quite common problem in multilabel learning [3]. This is a non-standard classification task in which each data instance is associated to several class labels at once. Therefore, a multilabel classifier has to be able of predicting several outputs for each processed pattern. In this kind of datasets there are usually several minority labels and several majority ones. In addition the imbalance levels, the ratio between the most common labels and the rarest ones, are usually huge. As a result, balancing the labels distribution in this kind of datasets is a more complex job than in standard classification. In late years several resampling algorithms for multilabel data have been proposed, including random undersampling and oversampling, as well as heuristic undersampling. Some ensemble-based solutions have been presented as well.

The method we introduced in [1] is a multilabel version of the well-known SMOTE [2] algorithm. It aims to produce synthetic instances linked to minority labels, instead of mere clones as previous oversampling proposals did. The main characteristics of MLSMOTE are the following:

- It takes into account the presence of several minority labels, producing new data samples associated to each one of them. Both SMOTE and previous multilabel oversampling techniques consider one class only, so new samples are exclusively associated to the rarest label.
- New data samples produced by MLSMOTE are given a synthetic set of input attributes, instead of being clones of existing instances. This way new patterns are not located into the same position than those taken as reference. The new attributes values are obtained by means of interpolation techniques, as in SMOTE.
- A synthetic labelset is computed for each new data sample. For doing so, the labels in the reference instance and their neighbors are taken into account. Three different strategies to produce these synthetic labelsets were tested. All previous resampling multilabel methods cloned the labelset of the reference instance.

An extensive experimentation was conducted to assess the performance of MLSMOTE. In it, a dozen multilabel datasets were processed with five imbalance aware algorithms, including oversampling (LP-ROS, ML-ROS, SmoteUG), undersampling (BR-IRUS) and ensemble (EML) based approaches. Each configuration was then given to five different multilabel classifiers (BR, RAkEL, CLR, HOMER and IBLR-ML), whose results were evaluated using common multilabel performance metrics. MLSMOTE was the best performer in all cases, achieving statistically significant differences in some of them. These result proved that synthetic instance generation through MLSMOTE, which includes the creation of synthetic labelsets, can be a successful approach when it comes to tackle imbalanced multilabel learning.

References

1. F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera. MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385–397, 2015.
2. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *J. Artificial Intelligence Res.*, 16:321–357, 2002.
3. F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus. *Multilabel Classification. Problem analysis, metrics and techniques.* Springer, 2016.