Classification from imbalanced datasets. A framework for improving the application of sampling strategies

Mohamed S. Kraiem

Supervisor: María N. Moreno García

Department of Computing and Automation, University of Salamanca
Plaza de los Caídos s/n 37008, Salamanca, Spain
ing_kriem@yahoo.com

Start date: 1st February 2016

1 Summary

This research work is focused on classification from imbalanced datasets, which represents an important obstacle in supervised learning. Datasets are imbalanced when at least one class, called the minority class, is represented by only a small number of training examples while the other class make up the majority. In these situation the precision for the minority class is usually significantly lower than the precision for the majority class, therefore predictive models are not valid even when they present an acceptable global accuracy. That is, the classifier can achieve a high percentage of correctly classified instances but the percentage of instances belonging to the minority class that are correctly classified can be very low. Additionally, the minority class is usually the most interesting one and misclassification of its instances is the least desirable. Most original classification algorithms only pursue to minimize the error rate without considering the difference between types of misclassification errors, which must be evaluated in imbalanced data contexts.

1.1 Starting hypothesis and previous researches

Imbalanced class distribution are frequently encountered in many real world classification applications such as medical diagnosis, survival studies, fraud detection, customer retention and many other domains. A number of solutions to the class imbalance problem were proposed both at the data and algorithmic levels. The approaches most commonly used are [Galar et al., 2012] [Bekkar and Alitouche, 2013]:

1. Basic sampling methods. There are three kinds of strategies:

   - Oversampling: Randomly replicates the number of minority class examples.
   - Undersampling: Removes random examples from the majority class.
   - Hybrid: Combination of oversampling and undersampling strategies.

Actas de la XVII Conferencia de la Asociación Española para la Inteligencia Artificial, pp. 947-951

2. Advanced sampling methods. These methods implement specific sampling strategies, such as generating synthetic examples, removing class label noise, taking into account borderline or redundant examples, etc. Some of these methods are: Tomek Links, SMOTE, SMOTE variants, One-Side Selection (OSS) and Neighborhood Cleaning Rule (NCL).
3. Classifier ensembles: Combination of ensembles with sampling processing techniques
4. Cost sensitive methods: They take into account the misclassification costs.
5. Algorithm modification: The aim is to provide adjustments on the learning algorithms in order to make them more appropriate to imbalanced data situations. This approach is used mainly with decision trees and SVM.

The two last approaches are less often employed owing to certain difficulties in their application. Adapting every algorithm to imbalanced data demands a great effort and sometimes provides poorer results than resampling techniques. On the other hand, cost sensitive learning usually requires domain experts to give values to the cost matrix containing the penalties for the different types of misclassification making it difficult to find the most suitable values.

In this research we will focus on basic and advanced sampling methods because they are commonly used in practice and the most broadly applicable to solve imbalanced datasets problem.

Basic undersampling and oversampling strategies present important shortcomings [Hulse et al., 2007]. Removing potentially valuable data is the main drawback of undersampling the majority class, while oversampling the minority class can cause overfitting problems as well as an increase in the computational cost of inducing the models. On the other hand, classification models induced from the oversampled datasets, are usually very large and complex. Advanced sampling methods have been proposed to avoid these drawbacks. One of the most widely used is SMOTE (Synthetic Minority Over Sampling Technique) [Chawla et al., 2002], which creates artificial instances of the minority class by introducing new computed examples along the line segments joining the k minority class nearest neighbors. When examples are not linearly separable, machine learning algorithms increase the accuracy by assigning the overlapped area to the majority class and treating the minority class as noise. To address this problem, some SMOTE modifications have been proposed. Some of them guide the creation of examples to specific parts of the input space inside of the limits of the positive class [Maciejewski et al., 2011] [Bunkhumpornpat et al., 2011] while others use noise filters after the application of SMOTE [Sáez et al., 2015]. There are other advanced sampling techniques as the one based on the Tomek-Links concept for removing the borderline majority samples, or the Neighbourhood Cleaning Rule [Laurikkala, 2001] based on the Wilson´s ENN method. One Sided Selection (OSS) [Kubat and Matwin, 1997] is a Tomek-Link based algorithm especially efficient in the case of high imbalanced data, but it requires significant execution time and processing resources.

There are many works in the literature addressing the topic of imbalanced data classification. Most of them consist of analysis and proposals to improve the existing algorithms, but they are focused on particular aspects. There is a lack of more general works

that can be taken as reference to choose the most appropriated approaches to each problem.

1.2 Main Objectives and Sub-objectives

The main objective of the doctoral thesis is to address the problem of imbalanced data classification, specifically for basic and advanced sampling approaches. As most of the methods, these techniques have advantages and drawbacks that it is necessary to analyze for different contexts and diverse datasets since their behavior depends on many factors. For instance, as the degree of data complexity increases, the class imbalance affects in a greater extend to the generalization ability of the classifiers. If there is a lack of data, the estimated decision boundary can be far from the true boundary.

We intend to conduct a deep comparative study about the behavior of machine learning algorithms in imbalanced data contexts and the effectiveness of the sampling strategies. The specific objectives to be achieved are the following:

- Study of the behavior of classifiers induced with several machine learning algorithms from imbalanced datasets
- Study and analysis the effect of imbalanced ratio parameter, this parameter should be taken in account because it's as a one factor that determine which resampling strategies will be used.
- Study of other factors influencing the utility of basic and advanced resampling techniques.
- Analysis of the improvements in the performance of the classifiers when original imbalanced datasets are preprocessed by mean of basic and advanced sampling strategies.
- Study and analysis of the results provided by classical and novel performance measures, such as optimized precision and generalized index of balanced accuracy [García et al., 2019, 2012].
- Examination of the size, complexity, sparsity, imbalance degree and other characteristics of the datasets that can influence the success of the sampling strategies.
- Inferring advantages and disadvantages of the sampling methods based on their sampling procedure, the analyzed characteristics of data and the applied machine learning algorithms.
- Proposal and validation of a framework that provides support for selecting the most suitable combination of machine learning algorithm and sampling technique for the problem to be solved at a given time.

1.3 Novel contribution

By means of the proposed study, we intent to get an insight into the behavior and applicability of sampling methods broader and deeper than other studies in the bibliog-

raphy. The research aims to contribute to finding theoretical and practical aspects influencing the effectiveness of the application of these strategies considering a great variety of datasets and machine learning algorithms.

The framework proposed as the final outcome of the research work can provide researchers from diverse areas with a useful tool that allows them to tailor the methods to be used to every particular problem without the need of spending time in checking their suitability.

2 Methodology and Work Plan

A comprehensive study of published work regarding the problem of classification from imbalanced data will be carried out. Special attention will be paid to the works where the problem is addressed by using sampling strategies. It is also important to include in the study papers analyzing different causes that lead to a loss of performance of classifiers induced from imbalanced data as well as the ways of dealing with these problem of the most common sampling methods.

Next step is data collection. A great variety of data sets will be used in this study in order to achieve the objectives described previously. At the moment we are planning to use datasets containing clinical information from patients hospitalized in ICU, datasets from the UCI Machine learning Repository as well as data obtained from other sources.

All datasets will be analyzed in order to know their characteristics, which could influence the effectiveness of the sampling strategies and the behavior of the machine learning algorithms.

The performance of the machine learning algorithms will be tested by using both the original datasets and these datasets preprocessed by several sampling techniques. Classifiers are typically evaluated by mean of their accuracy, however, this measure is not appropriate when there is imbalance in the data since in these scenarios, as commented before, machine learning algorithms can achieve a good accuracy but the precision for the minority class can be very low. Therefore, accuracy can be complemented with other metrics that provide additional error perspectives. Some of the metrics that will be used to evaluate the quality and performance of the models are area under the ROC Curve, precision, recall, F-Measure, optimized precision and generalized index of balanced accuracy. K-fold cross-validation will be used in the validation of all classifiers.

We will make use of Weka tool and R platform to support this research.

The research work will be divided in four stages. During the first stage, with an estimated duration of 4 months, we will conduct the bibliographic study. In the second stage we will design and implement the experiments required for the comparative study. It will take us about 18 months. The next 12 months (stage 3) will be spent on analyzing overall results in order to propose a general framework. In the fourth stage that comprises the last two months the final discussion and presentation will be prepared.

References

1. M. Bekkar, T.A. Alitouche, Imbalanced data learning approaches review. International Journal of Data Mining & Knowledge Management Process (IJDKP), 3(4), 2013, 15-33.
2. C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, Safe-level-SMOTE, safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, in: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD '09, Springer-Verlag, Berlin, Heidelberg, 2009, 475–482.
3. N. Chawla, K. Bowyer, L. Hall, W.P. Kehelmeyer. SMOTE: Sysnthetic Minority Over-sampling Technique. Journal of Artificial Intelligent Research, 2002; 16,321-357.
4. M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, IEEE Transactions on Syst.Man Cybernetics Part C: appl. Rev. 42(4), 2012, 463-484.
5. J. Hulse, T. Khoshgoftaar, Napolitano A. Experimental perspectives on learning from imbalanced data. Proceedings of the 24th International Conference on Machine learning. 2007, 935-942.
6. M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: One-sided selection. In Douglas H. Fisher, editor, ICML, 1997, 179–186. Morgan Kaufmann.
7. J. Laurikkala, Improving Identification of Difficult Small Classes by Balancing Class Distribution", AIME, LNAI 2101, 2001, 63–66, Springer-Verlag Berlin Heidelberg.
8. T. Maciejewski, J. Stefanowski, Local neighbourhood extension of SMOTE for mining imbalanced data, in: Proceedings of IEEE Symposium on Computational Intelligence and Data Mining, SSCI IEEE, IEEE Press, 2011, 104–111.
9. J.A. Sáez, J. Luengo, J. Stefanowski, F. Herrera. SMOTE-IPF: Addressing the noisy and bordeline examples problem in imbalanced classification by a resampling method with filtering. Information Sciences, 291(2015), 184-203.
10. V. García, J.S. Sanchez, R.A. Mollineda, On the effectiveness of preprocessing methods when dealing with different levels of class imbalance, ELSEVIER, 28(2012), 13-21
11. V. García, R.A Mollineda and J.S. Sanchez, Index of Balance Accuracy: A performance Measure for Skewed Class Distributions, H. Araujo et al (Eds): IbPRIA 2009, LNCS 5524, pp 441-448, 2009, ©Springer-Verlag Berlin Heidelberg 2009