Evaluation of Data Imbalance Solutions and Generalization

Student: Eduardo Antônio de Lucena Lisboa
Date: 11/04/2024
Advisor: Profa. Dra. Fabiane da Silva Queiroz

1. Context

Data imbalance in classification problems is a complicated subject to deal with, as it hinders classifiers lowering their accuracy and ability to predict on new data.

Several methods have been developed to reduce this problem. Some deal with the algorithms that will handle the data and others deal with the data itself.

Some of these methods have proven to be quite successful in enhancing results with imbalanced data, but most of them only work in very specific datasets. A method or combination of methods that improves classification in a generalized scenario is essential.

2. Motivation

This problem first appeared to me in a PPSUS research during graduation where we dealt with human visceral leishmaniasis. In it, the dataset had an imbalance ratio of approximately 1 to 100. The whole point of the research was to try an improvement method and this project ended up becoming my undergraduate thesis.

As the research aimed to create an application that would be used by the brazilian public health care system, it would greatly reduce the workload on the professionals that had to manually analyze and classify each sample and also reduce the time spent identifying the parasite.

3. Objective

- Analyze data augmentation methods and classification algorithms to find the best combination that yields the best results in a generalized scenario, using multiple datasets of different areas.
- Goals:
  - Find a combination of algorithms that can enhance classification results in imbalanced datasets and possibly create a framework for the best combination found

4. Relevance

Currently there are many papers with new methods, but most of them only use specific datasets of the same type, or even a single dataset.

This research aims to find good methods that can work well with different types of datasets, being able to enhance results where the current state-of-the-art methods aren’t good enough to aid professionals in their tasks.

It can also potentially be of use for those with data imbalance problems who don’t know where to start to solve their problem.

5. Related Works

- (ABDULLAH MARAŞ; EROL, 2023) propose a method that uses *CMeans clustering* along with data augmentation techniques to try to achieve better results, comparing it to classic data augmentation methods such as SMOTE.
- (HASSANAT et al., 2022) propose a method of data partitioning with voting rule for classification, where it evenly divides the majority class into the size of the minority class, trains classifiers and the majority of predictions will be the final prediction for new data.
- These studies suffer from the same problem exposed before, not generalizing enough.

6. Preliminary Results

I have gathered six different imbalanced datasets of all sorts of areas and selected a few methods and algorithms, through a simplified systematic review, to try all sorts of combinations and evaluate their metrics.

At this moment, there are no results available.

7. Conclusion

This research aims to give a better understanding on methods for dealing with imbalanced datasets and present a good, effective, generalization method that can be used in many scenarios without much fiddling.

Next steps are to implement the algorithms and methods selected and start the experiments phase.

I would like to acknowledge my advisor for her endless patience and support and CAPES for the master’s scholarship.

Thank You!

Eduardo Antônio de Lucena Lisboa
eall@ic.ufal.br

References

ABDULLAH MARAŞ; EROL, Ç. **FuzzyCSampling: A Hybrid fuzzy c-means clustering sampling strategy for imbalanced datasets**. Turkish journal of electrical engineering and computer sciences/Elektrik, v. 31, n. 7, p. 1223–1236, 30 nov. 2023.

HASSANAT, A. B. et al. **RDPVR: Random Data Partitioning with Voting Rule for Machine Learning from Class-Imbalanced Datasets**. Electronics, v. 11, n. 2, p. 228, 12 jan. 2022.