Response to discussion on “Improved overlap-based undersampling for imbalanced dataset classification with application to epilepsy and Parkinson’s disease”.

VUTTIPITTAYAMONGKOL, P. and ELYAN, E.

2020

Response to Discussion on

"Improved Overlap-Based Undersampling for Imbalanced Dataset Classification with Application to Epilepsy and Parkinson's Disease," *International Journal of Neural Systems*, 30:8, August 2020

Pattaramon Vuttipittayamongkol and Eyad Elyan*
School of Computing, Robert Gordon University, Garthdee Road
Aberdeen, AB10 7GJ, United Kingdom
*E-mail: p.vuttipittayamongkol@rgu.ac.uk; e.elyan@rgu.ac.uk**
https://www.rgu.ac.uk

In the paper "Improved Overlap-Based Undersampling for Imbalanced Dataset Classification with Application to Epilepsy and Parkinson's Disease", the authors introduced two new methods that address the class overlap problem in imbalanced datasets. The methods involve identification and removal of potentially overlapped majority class instances. Extensive evaluations were carried out using 136 datasets and compared against several state-of-the-art methods. Results showed competitive performance with those methods, and statistical tests proved significant improvement in classification results. The discussion on the paper related to the behavioral analysis of class overlap and method validation was raised by Fernández. In this article, the response to the discussion is delivered. Detailed clarification and supporting evidence to answer all the points raised are provided.

Keywords: class overlap; imbalanced data; undersampling; classification; medical, Fuzzy C-means.

1. Introduction

The authors appreciate the discusser's interest and comments on their paper. The discusser raised two main issues regarding the research presented in the paper. The first point was related to the behavioral analysis of class overlap. The second point was related to experiments and the validation of the results. A detailed response to the discusser's comments is provided in the following two sections.

2. Behavioral Analysis of Class Overlap

In the paper, the authors introduced two new methods that extend their previously published work, namely OBU (the Overlap-Based Undersampling method).[1] The two new methods are AdaOBU and BoostOBU. The objectives of this work were 1) to introduce an adaptive elimination threshold, which enables generalization across different datasets without the need for parameter tuning while ensuring that the undersampling rate is proportional to the class overlap, and 2) to enhance the performance of the approach by improving identification of potentially overlapped majority class instances.

The fuzzification of data, which by the definition given in Section 3.2.1 of the paper was highly related to the amount of class overlap, was addressed and incorporated as the main component in the proposed methods by means of Fuzzy C-means. The behavior of these overlap- based methods in relation to class overlap in the imbalanced context was extensively analyzed and evaluated (main concern raised by the discusser). An experiment using 66 simulated datasets representing different combinations of class overlap and class imbalance was carried out. Using these simulated datasets enabled us to measure the true class overlap degree. This analysis has provided us with further understanding of the impact of class overlap, and validates the main capability of our methods to adaptively remove problematic instances.

Moreover, as the full spectrum of class overlap degrees with a wide range of imbalanced scenarios were used in the experiment, it was shown that the datasets were not purposely selected to show only good results of the proposed methods, which was a concern raised by the discusser. This can also be supported by the other experiments provided in the paper, where the methods were further evaluated using 70 real-world public datasets.

More analysis of the methods in relation to class overlap was provided in Section 5.1.1 of the paper. This was on the adaptive threshold (μ_{th}) that was designed to control the undersampling based on fuzzification, i.e. class overlap, of the dataset. It was shown that μ_{th} was self- adaptive to the amount of class overlap in different imbalanced cases. Cooperating μ_{th} with the technique to identify instances potentially in the overlapping region facilitates adequate and accurate removal of overlapped instances.

With the above evidence, it can therefore be said that the performance of the proposed methods was evaluated and analyzed through the full scale of overlapping between classes in the context of various imbalance scenarios including extreme ones.

Another point was raised by the discusser regarding the use of Fuzzy C-means (FCM). Firstly, Euclidean distance was considered in this work because such distance is one of the most robust and widely-used metrics in FCM.[2] It was shown in Ref. 2 that FCM with Euclidean distance provided better results on most simulated and real-world datasets than other distance metrics. That said, certainly, it would be interesting to see how performance of the proposed methods can vary based on different other distance metrics. However, this is beyond the scope of this paper, but can be considered as one possible future work along with other potential directions to improve the methods as suggested in Section 7. The issue of dimensionality and possible solutions were also discussed in the same section.

3. Results Validation

It has to be admitted that it can be subjective when it comes to determining how much is enough to confirm the good performance of a newly-proposed algorithm. To demonstrate the value of the methods presented in the paper as well as to maintain high standards of the work, the proposed methods had been systematically and thoroughly evaluated and fairly compared with well- known and state-of-the-art approaches. Another key consideration was ensuring how well the proposed methods generalize across a wide range of simulated and real-world datasets. A total of 136 datasets were used in this paper for evaluation purposes.

Three main experiments were carried out. In Experiment I, 66 simulated datasets were used. These are covering 0% – 100% overlap degrees and imbalance degrees of 1.5 – 120, which imply 60% – 99.17% negative instances in the datasets. The data sizes ranged between 6,050 – 10,000 instances with various data densities. In Experiment II, 66 real-world imbalanced datasets from public repositories that are commonly used to evaluate methods that handle similar research problem were used. These datasets vary in terms of imbalance degrees, number of features, and sizes as shown in Section 4.3. Experiment III, the methods were evaluated on 2 large and high dimensional datasets. Furthermore, 2 additional datasets related to neurological diseases were used for further evaluation of the proposed methods in the medical context.

Detailed discussion and analysis of the results of all these experiments can be found Section 5. It showed clearly that the methods presented in this paper outperformed several well-established and state-of-the-art methods such as SMOTE, Borderline-SMOTE, k-means undersampling, OBU, SMOTE-ENN, SMOTEBagging and RUSBoost.

Another concern raised by the discusser was low specificity of the proposed methods. While the proposed methods provided significant improvement in sensitivity, lower specificity can be expected as a trade-off, which is inevitable when there exists overlapping between classes.[3] Since in this work imbalanced problems were considered, more focus was put on improving sensitivity, and more importantly, how to achieve that with a good trade-off, which can be reflected through good G-mean or F1-score. This is a feature of the BoostOBU method which was demonstrated through the results as can be seen in the paper. The discusser's claimed that the method performed statistically worse than SMOTE in terms of G-mean and F1-score. However, the detailed results and discussion of BoostOBU in Section 5 proves otherwise.

Statistical tests were also carried out in all experiments where an adequate amount of samples was available, i.e. in Experiment I and II. These were carefully done under the supervision of a professional statistician to ensure meaningful statistical practices and results.

The statistical test results proved that there was significance in the improvement provided by the presented methods in many cases as detailed in the paper. Such robust results with adequate statistical comparisons with state-of-the-art methods clearly emphasized persistent contributions of this research work.

References

1. P. Vuttipittayamongkol, E. Elyan, A. Petrovski and C. Jayne, Overlap-based undersampling for improving imbalanced data classification, *Int. Conf. Intell. Data Eng. Autom. Learn.*, (Springer, 2018), pp. 689-697.
2. J. Arora, K. Khatter and M. Tushir, Fuzzy c-means clustering strategies: A review of distance measures. *In Software Engineering*, (Springer, 2019), pp. 153-162.
3. B. R. Walker and N. R. Colledge. Davidson's Principles and Practice of Medicine E-Book. *Elsevier Health Sciences*, 2013.