A Pareto Based Ensemble with Feature and Instance Selection for Learning from Multi-Class Imbalanced Datasets

Alberto Fernandez
DaSCI Andalusian Institute of Data Science and Computational Intelligence
University of Granada
Granada, Spain
alberto@decsai.ugr.es

Cristobal José Carmona
Department of Computer Science
University of Jaen
Jaen, Spain
ccarmona@ujaen.es

Maria Jose del Jesus
Department of Computer Science
University of Jaen
Jaen, Spain
mjjesus@ujaen.es

Francisco Herrera
DaSCI Andalusian Institute of Data Science and Computational Intelligence
University of Granada
Granada, Spain
herrera@decsai.ugr.es

***Abstract*—This is a summary of our article published in International Journal of Neural Systems [1], to be considered as part of the Multi- Conference CAEPIA'18 - FINO'18 KeyWorks**

***Index Terms*—Imbalanced Classification, Multi-class, Overlapping, Feature Selection, Instance Selection, Multi-objective Evolutionary Algorithms, Ensembles**

I. Summary

In many classification tasks, we observe that some concepts are more difficult to be recognized than others [2]. This is related to the structure and inner characteristics of the data [3]. In particular, we may refer to the problem of imbalanced classification tasks as an important framework in Data Science problems [4]. It is found when some classes are underrepresented and the accuracy obtained for them is on average much lower than on the remaining classes [3].

As commented above, there are several reasons that, in conjunction with the imbalance, impose a hard restriction for the learning algorithms. The most significant one is the overlapping between classes [5]. This issue is strongly related with the attributes that represent the problem. Our hypothesis is that the use of feature selection will allow simplifying the boundaries of the problem by limiting the influence of those features that may create difficulties for the discrimination [6].

However, the imbalance class problem cannot be addressed by itself just by carrying out a feature selection. For this reason, it is also mandatory to perform a preprocessing of instances by resampling the training data distribution [7], avoiding a bias of the learning algorithm towards the majority classes. In addition, we may go one step further to integrate both approaches into an ensemble-type classifier [8].

Obtaining the optimal set of features and instances for a given problem is not a trivial task. For this reason, an optimization procedure is often required. Among different approaches, recent works have shown the goodness of Multi-Objective Evolutionary Optimization (MOEA) procedures [9] due to their ability to perform a good exploration and exploitation of the solution space.

In our research, we proposed EFIS-MOEA, which stands for "Ensemble classifier from a Feature and Instance Selection by means of Multi-Objective Evolutionary Algorithm." To do so, we embedded the C4.5 decision tree [10] in a wrapper procedure, applying the well-known NSGA-II multi-objective optimization algorithm [11].

The ultimate goal of our proposal was to provide a rule-based model that *maximizes the recognition of all individual classes*. This was achieved by focusing on the minority class clusters that were hard to identify. To do so, we focused on boosting the confidence of those rules associated with the former areas by means of the cleaning procedure, i.e. instance selection. In this way, the optimal criteria was minimizing the number of "bad" examples or, in other words, *maximizing the reduction of instances*. Additionally, and taking into account the findings made in [12], the coverage of the rules may imply capturing some of the non-related classes. We must point out that in order to obtain the quality of the recognition ability of the classifier, we computed MAUC metric as the macro-average of the pairwise AUC values of all pairs of classes:

$$\text{MAUC} = \frac{2}{m(m-1)} \sum_{i<j} AUC(C_i, C_j) \qquad (1)$$

This work was supported by the Spanish Ministry of Economy and Competitiveness under the projects TIN2015-68454-R and TIN2017-89517-P, including European Regional Development Funds.

The basis of our methodology involved several components:

1) First, feature selection was devoted to simplify the overlapping areas easing the generation of rules to distinguish between the classes.
2) Selection of instances from all classes addressed the imbalance itself by finding the most appropriate class distribution for the learning task, as well as possibly removing noise and difficult borderline examples.
3) Finally, the non-dominated solutions of the Pareto front from the MOEA could be directly combined into an ensemble of classifiers.

We set up a fair validation framework for the novel EFIS-MOEA proposal, considering two different case studies in binary and multi-class problems. Several approaches from the state-of-the-art were chosen in order to contrast the results. Particularly, the SMOTE+ENN preprocessing approach [13] for binary class problems and multi-class problems (using a binarization scheme [14]), and both Global-CS [15] and AdaBoost.NC [16] for the multi-class case study. Finally, we must recall that the behavior of EFIS-MOEA was contrasted versus 1-FIS-MOEA, i.e. a classifier obtained by selecting the best solution of the Pareto in terms of M-AUC.

In Table I we show the results for the binary case study. We observe that the synergy between feature selection and instance selection boosts the performance of our approach versus the oversampling and cleaning carried out by SMOTE+ENN, especially for highly overlapped problems in which the absolute differences are almost 4 points on average.

TABLE I
AVERAGE TRAINING-TEST RESULTS (*AUC*) AND STATISTICAL STUDY FOR BINARY IMBALANCED DATASETS.

Scenario	Method	AUC Train	AUC Test	Ranking	APV (Holm test)	W/T/L
Low overlap (F1 > 1.5) [30])	C4.5	.9510 ± .0253	.8892 ± .0661	94.00 (4)	.00000*	4/0/26
	C4.5-SMOTE+ENN	.9797 ± .0090	.9263 ± .0472	53.30 (2)	.00737*	8/0/22
	1-FIS-MOEA	**.9943 ± .0031**	.9195 ± .0514	65.47 (3)	.00005*	3/0/27
	EFIS-MOEA	.9906 ± .0072	**.9439 ± .0414**	29.23 (1)	*****	-/-/-
High overlap (F1 < 1.5) [36])	C4.5	.8437 ±.0454	.7352 ±.0726	113.78 (4)	.00000*	2/0/34
	C4.5-SMOTE+ENN	.9338 ±.0182	.7817 ±.0740	71.61 (2)	.00000*	3/0/33
	1-FIS-MOEA	**.9761 ±.0081**	.7749 ±.0757	79.22 (3)	.00000*	0/0/36
	EFIS-MOEA	.9717± .0100	**.8273± .0596**	25.39 (1)	*****	-/-/-

In Table II we show the results for the case study of multi-class problems. The findings extracted from this analysis is similar of that of the previous scenario. The goodness shown by our EFIS-MOEA approach is clear, as it is able to outperform all algorithms selected for comparison. The statistical results provide a strong support to the excellent capabilities for our approach. By taking advantage from all the solutions discovered in the optimization stage into an ensemble, results are significantly boosted with respect to the best classifier found in the MOEA search, i.e. 1-FIS-MOEA, which suffers from the curse of overfitting.

The results obtained by EFIS-MOEA were very competitive, especially for highly overlapped problems. The selection of instances allowed rebalancing the training set as well as to clean the low quality data, i.e. noisy and redundant examples. In addition, feature selection simplified the boundaries of the problem to manage the aforementioned overlapping issue.

TABLE II
AVERAGE TRAINING-TEST RESULTS (*M-AUC*) AND STATISTICAL STUDY FOR MULTI-CLASS IMBALANCED DATASETS.

Method	AUC Train	AUC Test	Ranking	APV (Holm test)	W/T/L
C4.5	.9006 ± .0141	.8157 ±.0297	102.54 (6)	.00000*	2/0/22
OVO-SMOTE+ENN	.9369 ± .0136	.8292 ± .0352	74.58 (5)	.00725*	6/0/19
Global-CS	**.9726 ± .0060**	.8324 ± .0346	72.48 (3)	.01206*	4/0/20
AdaBoost.NC	.9530 ± .0147	.8233 ± .0319	69.06 (2)	.02597*	8/0/16
1-FIS-MOEA	.9715 ±.0041	.8299 ±.0355	74.08 (4)	.00820*	5/0/19
EFIS-MOEA	.9691 ± .0058	**.8441 ± .0322**	42.25 (1)	*****	-/-/-

The behavior of EFIS-MOEA is excelled as it was shown to outperform the state-of-the-art algorithms, especially the AdaBoost.NC algorithm, a robust approach in this context.

References

[1] A. Fernandez, C. Carmona, M. del Jesus, and F. Herrera, *International Journal of Neural Systems*, vol. 27, pp. 1 750 028:1–1 750 028–21, 2017.

[2] M. Galar, A. Fernandez, E. Barrenechea, and F. Herrera, "Empowering difficult classes with a similarity-based aggregation in multi-class classification problems." *Inf. Sci.*, vol. 264, pp. 135–157, 2014.

[3] V. Lopez, A. Fernandez, S. Garcia, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, no. 20, pp. 113–141, 2013.

[4] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modelling under imbalanced distributions," *ACM Computing Surveys*, vol. 49, no. 2, pp. 31:1–31:50, 2016.

[5] V. Garcia, R. Mollineda, and J. S. Sanchez, "On the k–NN performance in a challenging scenario of imbalance and overlapping," *Pattern Analysis Applications*, vol. 11, no. 3–4, pp. 269–280, 2008.

[6] S. Alshomrani, A. Bawakid, S.-O. Shim, A. Fernandez, and F. Herrera, "A proposal for evolutionary fuzzy systems using feature weighting: Dealing with overlapping in imbalanced datasets." *Knowledge-Based Systems*, vol. 73, pp. 1–17, 2015.

[7] R. C. Prati, G. E. A. P. A. Batista, and D. F. Silva, "Class imbalance revisited: a new experimental setup to assess the performance of treatment methods," *Know. and Inform. Syst.*, vol. 45, no. 1, pp. 247–270, 2015.

[8] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for class imbalance problem: Bagging, boosting and hybrid based approaches," *IEEE Transactions on System, Man and Cybernetics Part C: Applications and Reviews*, vol. 42, no. 4, pp. 463–484, 2012.

[9] A. Zhou, B. Y. Qu, H. Li, S. Z. Zhao, P. N. Suganthan, and Q. Zhangd, "Multiobjective evolutionary algorithms: A survey of the state of the art," *Swarm and Evol. Comp.*, vol. 1, no. 1, pp. 32–49, 2011.

[10] J. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kauffman, 1993.

[11] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[12] A. Manukyan and E. Ceyhan, "Classification of imbalanced data with a geometric digraph family," *Journal of Machine Learning Research*, vol. 17, pp. 1–40, 2016.

[13] A. Fernandez, S. Garcia, F. Herrera, and N. Chawla, "Smote for learning from imbalanced data: Progress and challenges. marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.

[14] A. Fernandez, V. Lopez, M. Galar, M. J. del Jesus, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches." *Knowledge-Based Systems*, vol. 42, pp. 97–110, 2013.

[15] Z.-H. Zhou and X.-Y. Liu, "On multi-class cost-sensitive learning," *Computational Intelligence*, vol. 26, no. 3, pp. 232–257, 2010.

[16] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Transactions on System Man and Cybernetics: Part B*, vol. 42, no. 4, pp. 1119–1130, 2012.