Addressing Overlapping in Classification with Imbalanced Datasets: A First Multi-objective Approach for Feature and Instance Selection

Alberto Fernández[1(✉)], Maria Jose del Jesus[1], and Francisco Herrera[2]

[1] Department of Computer Science, University of Jaén, Jaén, Spain
{alberto.fernandez,mjjesus}@ujaen.es
[2] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
herrera@decsai.ugr.es

Abstract. In classification tasks with imbalanced datasets the distribution of examples between the classes is uneven. However, it is not the imbalance itself which hinders the performance, but there are other related intrinsic data characteristics which have a significance in the final accuracy. Among all, the overlapping between the classes is possibly the most significant one for a correct discrimination between the classes.

In this contribution we develop a novel proposal to deal with the former problem developing a multi-objective evolutionary algorithm that optimizes both the number of variables and instances of the problem. Feature selection will allow to simplify the overlapping areas easing the generation of rules to distinguish between the classes, whereas instance selection of samples from both classes will address the imbalance itself by finding the most appropriate class distribution for the learning task, as well as removing noise and difficult borderline examples.

Our experimental results, carried out using C4.5 decision tree as baseline classifier, show that this approach is very promising. Our proposal outperforms, with statistical differences, the results obtained with the SMOTE + ENN oversampling technique, which was shown to be a baseline methodology for classification with imbalanced datasets.

Keywords: Imbalanced classification · Overlapping · Feature selection · Instance selection · Multiobjective evolutionary algorithms

1 Introduction

The imbalanced class problem is one of the new challenges that arose when Machine Learning reached its maturity [6], being widely present in the fields of businesses, industry and scientific research. This issue grew up in importance at the same time that researchers realize that the datasets they analyzed hold more instances or examples from one class than that of the remaining ones, and they

K. Jackowski et al. (Eds.): IDEAL 2015, LNCS 9375, pp. 36–44, 2015.
DOI: 10.1007/978-3-319-24834-9_5

standard classification algorithms achieved a model below a desired accuracy threshold for the underrepresented class.

One of the main drawbacks for the correct identification of the minority or positive class of the problem, is related to overlapping between classes [8]. Rules with a low confidence and/or coverage, because they are associated with an overlapped boundary area, will be discarded.

The former fact is related with the attributes that represent the problem. It is well known that a large number of features can degrade the discovery of the borderline areas of the problem, either because some of these variables might be redundant or because they do not show a good synergy among them. Therefore, the use of feature selection can ease to diminish the effect of overlapping [4].

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, avoiding a bias of the learning algorithm does towards the majority class.

In accordance with the above, in this work contribution we aim at improving current classification models in the framework of imbalanced datasets by developing both a feature and instance selection. This process will be carried out means of a multi-objective evolutionary algorithm (MOEA) optimization procedure. The multi-objective methodology will allow us to perform an exhaustive search by means of the optimization of several measures which, on a whole, are expected to be capable of giving a quality answer to the learnt system. In this sense, this wrapper approach will be designed to take advantage of the exploration of the full search space, as well as providing a set of different solutions for selecting the best suited for the final user/task.

Specifically, we will make use of the well known NSGA2 approach [3] as the optimization procedure, and the C4.5 decision tree [10] as baseline classifier. We must stress that, although the C4.5 algorithm carries out itself an inner feature selection process, our aim is to 'ease' the classifier by carrying out a pre-selection of the variables with respect to the intrinsic characteristics of the problem, mainly referring the overlapping between the classes.

This contribution is arranged as follows. Section 2 introduces the problem of classification with imbalanced datasets and overlapping. Section 3 describes our MOEA approach for addressing this problem. Next, Sect. 4 contains the experimental results and the analysis. Finally, Sect. 5 will conclude the paper.

2 Imbalanced Datasets in Classification

In this section, we will first introduce the problem of imbalanced datasets. Then, we will focus on the presence of overlapping between the classes.

2.1 Basic Concepts

Most of the standard learning algorithms consider a balanced training set for the learning stage. Therefore, addressing problems with imbalanced data may

cause obtaining of suboptimal classification models, i.e. a good coverage of the majority examples whereas the minority ones are misclassified frequently [8]. There are several reasons behind this behaviour which are enumerated below:

- The use of global performance measures for guiding the search process, such as standard accuracy rate, may benefit the covering of the majority examples.
- Classification rules that predict the positive class are often highly specialized, and they are discarded in favour of more general rules.
- Very small clusters of minority class examples can be identified as noise, and therefore they could be wrongly discarded by the classifier.

In order to overcome the class imbalance problem, we may find a large number of proposed approaches, which can be categorized in three groups [8]:

1. Data level solutions: the objective consists of rebalancing the class distribution via preprocessing of instances [2].
2. Algorithmic level solutions: these solutions try to adapt several classification algorithms to reinforce the learning towards the positive class [1].
3. Cost-sensitive solutions: they consider higher costs for the misclassification of examples of the positive class with respect to the negative class [5].

2.2 Overlapping or Class Separability

The problem of overlapping between classes appears when a region of the data space contains a similar quantity of training data from each class, imposing a hard restriction to finding discrimination functions.

In previous studies on the topic [9], authors depicted the performance of the different datasets ordered according to different data complexity measures (including IR) in order to search for some regions of interesting good or bad behaviour. They could not characterize any interesting behaviour according IR, but they do for example according the so called metric $F1$ or *maximum Fisher's discriminant ratio* [7], which measures the overlap of individual feature values.

This metric for one feature dimension is defined as: $f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$ where μ_1, μ_2, σ_1^2, σ_2^2 are the means and variances of the two classes respectively, in that feature dimension. We compute f for each feature and take the maximum as measure F1. For a multidimensional problem, not all features have to contribute to class discrimination. Therefore, we can just take the maximum f over all feature dimensions when discussing class separability. Datasets with a small value for the F1 metric will have a high degree of overlapping.

Finally, a closely related issue is the impact of noisy and borderline examples on the classifier performance in imbalanced classification [11]. Regarding this fact, a preprocessing cleaning procedure can help the learning algorithm to better discriminate the classes, especially in the overlapped areas.

3 Addressing Overlapping in Imbalanced Domains by a Multi-objective Feature and Instance Selection

In this work, our contribution is to introduce a new methodology that makes use of a MOEA to determine the best subset of attributes and instances in imbalanced classification. Instance selection aims at both balancing the data distribution between the positive and negative classes, and removing noisy and borderline examples that hinder the classification ability of the learning algorithm. Feature selection will simplify the boundaries of the problem by limiting the influence of those features create difficulties for the discrimination process.

However, the estimation of the best suited subset of instances and features is not trivial. In accordance with the former, an optimization search procedure must be carried out in order to determine the former values. Among the different techniques that can be used for this task, genetic algorithms excel due to their ability to perform a good exploration and exploitation of the solution space. Our ultimate goal is to build the simplest classifier with the highest accuracy in the context of imbalanced classification. Regarding this issue, the first objective can be overcome by *maximizing the reduction of instances*, whereas the second one is achieved by *maximizing the recognition of both the positive and negative classes*. In accordance with the former, we propose the use of the "Area Under the ROC Curve" (AUC), as it provides a good trade-off between the individual performance for each individual class (Eq. 1).

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{1}$$

Taking into account the objectives previously outlined, we propose the design of a work methodology using as basis a MOEA. This way, we can take advantage of both the exploration capabilities of this type of technique, as well as allowing the selection among a set of different solutions, depending on the user's requirements. We will name this approach as *IS + FS-MOEA*.

Specifically, we will make use of the NSGA-II algorithm [3] for implementing our model, as it is widely known for being a high-performance MOEA. Its two main features are first the fitness evaluation of each solution based on both the Pareto ranking and a crowding measure, and the other is an elitist generation update procedure.

In order to codify the solutions, we will make use of a chromosome with two well differentiate parts: one (FS) for the feature selection and another one (IS) for the instance selection. Both parts will have a binary codification, in such a way that a 0 means that the variable (or instance) will not take part for generating the classification model, whereas a 1 value stands for the opposite case. Chromosomes will be evaluated jointly with aims at obtaining the best synergy between both characteristics, instead of optimizing them separately. This issue is based on the fact that it is not clearly defined which the best order for carrying our both processes is. An initial chromosome will be built with all genes equal to '1' in order to implement the standard case study, i.e. the full training set, whereas the remaining individuals will be generated at random.

As baseline classifier, we will make use of the C4.5 decision tree [10] for several reasons. The first one is its wide use in classification with imbalanced data, so that we may carry out a fair comparative versus the state-of-the-art. The second one is its efficiency; since we need to perform a large number of evaluations throughout the search process, it is important the base model to be particularly quick for not biasing the global complexity of the methodology.

We must stress that, the C4.5 algorithm carries out itself an inner feature selection process. However, our aim to "ease" the classifier by carrying out a pre-selection of the variables with respect to the intrinsic characteristics of the problem, mainly referring the overlapping between the classes, so that we can improve the classification of both classes together.

For the evaluation of the chromosomes, we carry out the preprocessing of the training set codified in the phenotype, and then the C4.5 classifier is executed with the modified dataset. Then, the objective functions to be maximized are computed as stated in Eq. 2, being N the number of initial training instances, and IS_i the value of the chromosome for the instance selection part.

$$\begin{aligned} &OBJ_1 : AUC \\ &OBJ_2 : RED = N - \textstyle\sum_{i=0}^{N-1} IS_i; \end{aligned} \tag{2}$$

4 Experimental Study

This section includes the experimental analysis of the proposed approach. With this aim, we first present the experimental framework including the datasets selected for the study, as well as the parameters of the algorithms, and the use of statistical test. Then, we show the complete results and the comparison with the state-of-the-art to determine the goodness of our proposal.

4.1 Experimental Framework

Table 1 shows the benchmark problems selected for our study, in which the name, number of examples, number of attributes, and IR (ratio between the majority and minority class instances) are shown. Datasets are ordered with respect to their degree of overlapping. A wider description for these problems can be found at http://www.keel.es/datasets.php. The estimates of AUC measure are obtained by means of a 5 fold Cross-Validation, aiming to include enough positive class instances in the different folds.

The parameters of the NSGA-II MOEA have been set up as follows: 60 individuals as population size, with 100 generations. The crossover and the mutation (per gen) probabilities are 0.8 and 0.025 respectively. For the C4.5 decision tree we use a confidence level at 0.25, with 2 as the minimum number of item-sets per leaf, and the application of pruning will be used to obtain the final tree. As state-of-the-art approach for the sake of a fair comparison we have selected the SMOTE + ENN preprocessing technique [2], which has shown a good synergy with the C4.5 algorithm [8]. This approach creates synthetic examples of the

Table 1. Summary of imbalanced datasets used

Name	#Ex.	#Atts.	IR	F1	Name	#Ex.	#Atts.	IR	F1
glass4	214	9	15.47	1.4690	pimaImb	768	8	1.90	0.5760
ecoli01vs5	240	6	11.00	1.3900	abalone19	4174	8	128.87	0.5295
cleveland0vs4	177	113	12.62	1.3500	ecoli0147vs2356	336	7	10.59	0.5275
ecoli0146vs5	280	6	13.00	1.3400	pageblocks0	5472	10	8.77	0.5087
yeast2vs8	482	8	23.10	1.1420	glass2	214	9	10.39	0.3952
ecoli0347vs56	257	7	9.28	1.1300	vehicle2	846	18	2.52	0.3805
vehicle0	846	18	3.23	1.1240	yeast1289vs7	947	8	30.56	0.3660
ecoli01vs235	244	7	9.17	1.1030	yeast1vs7	459	8	13.87	0.3534
yeast05679vs4	528	8	9.35	1.0510	glass0146vs2	205	9	11.06	0.3487
glass06vs5	108	9	11.00	1.0490	yeast0359vs78	506	8	9.12	0.3113
glass5	214	9	22.81	1.0190	glass016vs2	192	9	10.29	0.2692
ecoli067vs35	222	7	9.09	0.9205	yeast1	1484	8	2.46	0.2422
ecoli0267vs35	244	7	9.18	0.9129	glass1	214	9	1.82	0.1897
ecoli0147vs56	332	6	12.28	0.9124	vehicle3	846	18	2.52	0.1855
yeast4	1484	8	28.41	0.7412	habermanImb	306	3	2.68	0.1850
yeast0256vs3789	1004	8	9.14	0.6939	yeast1458vs7	693	8	22.10	0.1757
glass0	214	9	2.06	0.6492	vehicle1	846	18	2.52	0.1691
abalone918	731	8	16.68	0.6320	glass015vs2	172	9	9.12	0.1375

minority class by means of interpolation to balance the data distribution, and then it removes noise by means of the ENN cleaning procedure. Its configuration will be the standard with a 50 % class distribution, 5 neighbors for generating the synthetic samples and 3 for the ENN cleaning procedure, and Euclidean Metric for computing the distance among the examples.

Finally, we will make use of Wilcoxon signed-rank test [12] to find out whether significant differences exist between a pair of algorithms, thus providing statistical support for the analysis of the results.

4.2 Analysis of the Results

In this case study, the final aim is to obtain the highest precision for both classes of the problem in the test set. In this way, we will always select the one solution of the Pareto with the best performance with respect to the AUC metric. In this case, a comparison with the optimal Pareto front is not possible since for classification functions this is often unavailable.

Average values for the experimental results are shown in Table 2, where datasets are ordered from low to high overlapping. From these results we may highlight the goodness of our approach, as it achieves the highest average value among all problems. Additionally, we must stress that in the case study of the higher overlapped problems, i.e. from "ecoli0147vs2356", that our proposed approach outperforms the baseline SMOTE + ENN technique in 12 out of 16 datasets. Finally, it is worth to point out that our IS + FS-MOEA does not show

Table 2. Experimental results for C4.5 with SMOTE + ENN (C4.5 + S_ENN) and our C4.5 with IS + FS-MOEA approach (C4.5 + MOEA) in training and test with AUC metric.

Dataset	IR	F1	C4.5+S_ENN		C4.5+MOEA		Dataset	IR	F1	C4.5+S_ENN		C4.5+MOEA	
glass4	15.47	1.4690	.9813	**.8292**	.9838	.8225	pima	1.90	0.5760	.7976	**.7311**	.8293	.7084
ecoli01vs5	11.00	1.3900	.9676	**.8477**	.9795	.8455	abalone19	128.87	0.5295	.9009	**.5185**	.5000	.5000
cleveland0vs4	12.62	1.3500	.9922	.7179	.9828	**.8582**	ecoli0147vs2356	10.59	0.5275	.9561	.8529	.9571	**.8755**
ecoli0146vs5	13.00	1.3400	.9861	.8923	.9856	**.8981**	page-blocks0	8.77	0.5087	.9792	.9437	.9798	**.9442**
yeast2vs8	23.10	1.1420	.9115	**.8012**	.8359	.7664	glass2	10.39	0.3952	.9402	.6819	.9364	**.7797**
ecoli0347vs56	9.28	1.1300	.9540	.8502	.9718	**.8541**	vehicle2	2.52	0.3805	.9846	.9396	.9842	**.9512**
vehicle0	3.23	1.1240	.9716	.9160	.9761	**.9448**	yeast1289vs7	30.56	0.3660	.9359	.6397	.7931	**.6733**
ecoli01vs235	9.17	1.1030	.9720	.8218	.9527	**.8873**	yeast1vs7	13.87	0.3534	.9107	.6968	.8890	**.7759**
yeast05679vs4	9.35	1.0510	.9276	**.7725**	.9207	.7674	glass0146vs2	11.06	0.3487	.9157	**.7344**	.9553	.7274
glass06vs5	11.00	1.0490	.9912	**.9647**	.9975	.9350	yeast0359vs78	9.12	0.3113	.9214	**.7078**	.8628	.6978
glass5	22.81	1.0190	.9480	.8232	.9988	**.9951**	glass016vs2	10.29	0.2692	.9237	.6667	.9947	**.9572**
ecoli067vs35	9.09	0.9205	.9700	.7875	.9632	**.8450**	yeast1	2.46	0.2422	.7781	**.6957**	.7857	.6677
ecoli0267vs35	9.18	0.9129	.9851	**.7854**	.9642	.7827	glass1	1.82	0.1897	.8601	.6668	.8912	**.7420**
ecoli0147vs56	12.28	0.9124	.9598	.8457	.9738	**.8538**	vehicle3	2.52	0.1855	.8892	**.7675**	.8894	.7206
yeast4	28.41	0.7412	.9113	**.7157**	.8648	.7089	haberman	2.68	0.1850	.7428	.6076	.7326	**.6178**
yeast0256vs3789	9.14	0.6939	.9121	**.7649**	.8140	.7581	yeast1458vs7	22.10	0.1757	.8719	.5192	.7996	**.5745**
glass0	2.06	0.6492	.8862	.7748	.8917	**.8103**	vehicle1	2.52	0.1691	.8881	.7170	.8960	**.7340**
abalone9-18	16.68	0.6320	.9302	**.7332**	.8425	.7122	glass015vs2	9.12	0.1375	.9342	.7226	.9429	**.7433**
			C4.5-SMOTE+ENN							C4.5-MOEA			
Average			.9247		.7626		Average			.9033		**.7899**	

Table 3. Wilcoxon test for the comparison between C4.5 + MOEA [R^+] and C4.5 + S_ENN [R^-].

Comparison	R^+	R^-	p-value	W/T/L
C4.5 + MOEA vs C4.5 + S_ENN	460.0	206.0	0.044745	21/0/15

the curse of over-fitting, as the training performance is even lower than that of the standard preprocessing approach.

In order to determine statistically the best suited metric, we carry out a Wilcoxon pairwise test in Table 3. Results of this test agree with our previous remarks, since significant differences are found in favour of our IS + FS-MOEA approach with a confidence degree above the 95 %.

Finally, we must remark that the IS-FS-MOEA approach has a greater computational cost in terms of both memory and CPU time than the C4.5 + S_ENN algorithm, as it carries out an evolutionary process. However, its advantage over the former is twofold: (1) it has been shown to clearly outperform the former in terms of precision; and (2) it allows the final user to apply several solutions in order to select the one that better suites to the problem that is being addressed.

5 Concluding Remarks

In this work we have proposed a novel MOEA in the framework of classification with imbalanced datasets. This approach has been designed under a double perspective: (1) to carry out an instance selection for compensating the example distribution between the classes, as well as removing those examples which

include noise, or which difficult the discrimination of the classes; and (2) to perform a feature selection to remove those attributes that may imply a high degree of overlapping in the borderline areas.

The goodness in the use of the MOEA is related to its high exploration abilities, the capability of using several metrics to guide the search, and the availability of several solutions so that they any of them can be selected depending on the problem requirements.

Our experimental results have shown the robustness of our novel proposal in contrast with the state-of-the-art, and confirms the significance of this topic for future research. Among others, we plan to study the use of different objectives to guide the search, the use of the solutions of the MOEA as an ensemble approach, or even to develop a heuristic rule to select the best suited solution overall. Finally, we will test the behaviour of our model with problems with a higher complexity, including both a wider number of instances and/or features.

Acknowledgments. This work was supported by the Spanish Ministry of Science and Technology under projects TIN-2011-28488, TIN-2012-33856; the Andalusian Research Plans P11-TIC-7765 and P10-TIC-6858; and both the University of Jaén and Caja Rural Provincial de Jaén under project UJA2014/06/15.

References

1. Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recogn. **36**(3), 849–851 (2003)
2. Batista, G., Prati, R.C., Monard, M.C.: A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explor. **6**(1), 20–29 (2004)
3. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: Nsga-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002)
4. Denil, M., Trappenberg, T.: Overlap versus imbalance. In: Farzindar, A., Kešelj, V. (eds.) Canadian AI 2010. LNCS, vol. 6085, pp. 220–231. Springer, Heidelberg (2010)
5. Domingos, P.: Metacost: A general method for making classifiers cost-sensitive. In: Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD 1999), pp. 155–164 (1999)
6. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
7. Ho, T., Basu, M.: Complexity measures of supervised classification problems. IEEE Trans. Pattern Anal. Mach. Intell. **24**(3), 289–300 (2002)
8. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf. Sci. **250**(20), 113–141 (2013)
9. Luengo, J., Fernández, A., García, S., Herrera, F.: Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling. Soft Comput. **15**(10), 1909–1936 (2011)
10. Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kauffmann, San Francisco (1993)

11. Sáez, J., Luengo, J., Stefanowski, J., Herrera, F.: Smote-IPF: addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf. Sci. **291**, 184–203 (2015)
12. Sheskin, D.: Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall/CRC, Boca Raton (2006)