Managing Borderline and Noisy Examples in Imbalanced Classification by Combining SMOTE with Ensemble Filtering

José A. Sáez[1], Julián Luengo[2], Jerzy Stefanowski[3], and Francisco Herrera[1]

[1] Department of Computer Science and Artificial Intelligence, University of Granada, CITIC-UGR, 18071, Granada, Spain
{smja,herrera}@decsai.ugr.es
[2] Department of Civil Engineering, LSI, University of Burgos, 09006, Burgos, Spain
jluengo@ubu.es
[3] Institute of Computing Science, Poznań, University of Technology, ul. Piotrowo 2, 60-965 Poznań, Poland
Jerzy.Stefanowski@cs.put.poznan.pl

Abstract. Imbalance data constitutes a great difficulty for most algorithms learning classifiers. However, as recent works claim, class imbalance is not a problem in itself and performance degradation is also associated with other factors related to the distribution of the data as the presence of noisy and borderline examples in the areas surrounding class boundaries.

This contribution proposes to extend SMOTE with a noise filter called Iterative-Partitioning Filter (IPF), which can overcome these problems. The properties of this proposal are discussed in a controlled experimental study against SMOTE and its most well-known generalizations. The results show that the new proposal performs better than exiting SMOTE generalizations for all these different scenarios.

Keywords: Classification, imbalanced data, SMOTE, class noise, noise filters.

1 Introduction

Real-world classification problems from many fields present a highly imbalanced distribution of examples among the classes. In imbalance data one class is represented by a much smaller number of examples than the other classes. The minority class is usually the most interesting one [2] and thus class imbalance becomes a source of difficulty for most learning algorithms which assume an approximately balanced class distribution. As a result, minority class examples usually tend to be misclassified. Re-sampling methods modify the balance between classes by taking into account local properties of examples. Among these methods, the *Synthetic Minority Over-sampling Technique* (SMOTE) [5] is one of the most well-known.

E. Corchado et al. (Eds.): IDEAL 2014, LNCS 8669, pp. 61–68, 2014.

Even though SMOTE achieves a better distribution of the number of examples in each class it presents several drawbacks related to its *blind* oversampling. These drawbacks may aggravate even more the difficulties produced for noisy and borderline examples as some researchers have shown that the class imbalance ratio (IR) is not a problem itself. For instance in [14] the influence of noisy and borderline examples on classification performance in imbalanced datasets is experimentally studied. Borderline examples are those located either very close to the decision boundary between minority and majority classes or located in the area surrounding class boundaries where classes overlap. In [14] noisy examples are referred as those located deep inside the region of the other class.

The main aim of this contribution is to examine a new extension of SMOTE, where the IPF noise filter is combined with it resulting in SMOTE-IPF, compared to other re-sampling methods also based on generalizations of SMOTE. Its suitability for handling noisy and borderline examples in imbalanced data will be particularly evaluated as they are one of the main sources of difficulties for learning algorithms. In order to control the noise scenario, the experimental study will be carried out with special synthetic datasets containing different shapes of the minority class example boundaries and levels of borderline examples as considered in related studies [14]. After preprocessing these datasets, the performances of the classifiers built with C4.5 will be evaluated and they will be also contrasted using the proper statistical tests.

The rest of this contribution is organized as follows. Section 2 presents the imbalanced dataset problem and the motivations of our extension of SMOTE. Section 3 describes the experimental framework and includes the analysis of the experimental results. Finally Section 4 presents some concluding remarks.

2 Borderline and Noisy Examples in Imbalanced Datasets

In this section, first the problem of imbalanced datasets related to borderline and noisy examples is described in Section 2.1. Next the details of the proposed extension of SMOTE to solve these issues are given in Section 2.2.

2.1 Imbalanced Classification with Borderline and Noisy Examples

The main difficulty of imbalanced datasets is that standard classifiers tend to misclassify examples belonging to the minority class. This is because *accuracy* does not distinguish between the number of correct labels of different classes, and thus measures without these drawbacks have been proposed in the literature [10]. This paper considers the usage of the *Area Under the ROC Curve* (AUC) measure, which provides a single-number summary for the performance of learning algorithms and it is recommended in many other works on imbalanced data.

Imbalance ratio is not the only source of difficulty for classifiers. Recent works have indicated other relevant issues related to the degradation of performance. Closely related to the overlapping between classes, in [14] another interesting problem in imbalanced domains is pointed out: the higher or lower presence

of examples located in the area surrounding class boundaries, which are called borderline examples. Researchers have found that misclassification often occurs near class boundaries where overlapping usually occurs as well and it is hard to find a feasible solution for it [9].

The authors in [14] showed that classifier performance degradation was strongly affected by the quantity of borderline examples and that the presence of other noisy examples located farther outside the overlapping region was also very difficult for re-sampling methods. To clarify terminology, one must distinguish (inspired by [14,13]) between safe, borderline and noisy examples:

- *Safe examples* are placed in relatively homogeneous areas with respect to the class label.
- *Borderline examples* are located in the area surrounding class boundaries, where either the minority and majority classes overlap or these examples are very close to the difficult shape of the boundary - in this case, these examples are also difficult as a small amount of the attribute noise can move them to the wrong side of the decision boundary [13].
- *Noisy examples* are individuals from one class occurring in the safe areas of the other class. According to [13] they could be treated as examples affected by class label noise.

This contribution focuses on studying the influence of noisy and borderline examples on generalizations of SMOTE considering the synthetic datasets used in [14] where safe, borderline and noisy examples are distinguished. The examples belonging to the two last groups often do not contribute to correct class prediction [11]. Therefore removing them partially or completely should improve classification performance and we propose to use noise filters to achieve this goal. We are particularly interested in ensemble filters as they are the most careful while deciding whether an example should be viewed as noise and removed.

2.2 Combining SMOTE and IPF

SMOTE is one of the most well-known and used re-sampling techniques. It generates new synthetic examples of the minority class by along the linear space between every minority example and some of its randomly selected k-nearest neighbors. One of the main shortcomings of SMOTE is overgeneralization as SMOTE blindly generalizes regions of the minority class without checking positions of the nearest examples from the majority classes. These problems may be aggravated with some distributions of data when imbalanced datasets are suffering from noisy and borderline examples. As result SMOTE is usually combined with an additional cleaning to remove noisy and borderline examples [11].

Combining SMOTE with an additional step of under-sampling aims to remove mislabeled data from the training data after the usage of SMOTE. However, they do not perform this task as well as they should in all cases. Specific and more powerful methods designed to eliminate mislabeled examples are thus required to successfully deal with noise data in imbalanced domains. Noise filters are

preprocessing mechanisms designed to detect and eliminate noisy examples in the training set [3,12,15]. The result of noise elimination in preprocessing is a reduced training set which is then used as an input to a machine learning algorithm.

In addition, there are many other noise filters based on the usage of ensembles [7]. Similar techniques have been widely developed considering the building of several classifiers with the same learning algorithm [8,17]. Instead of using multiple classifiers learned from the same training set, in [8] a *Classification Filter* approach is suggested, in which the training set is partitioned into n subsets, then a set of classifiers is trained from the union of any $n-1$ subsets; those classifiers are used to classify the examples in the excluded subset, eliminating the examples that are incorrectly classified. This paper proposes to extend SMOTE with one of this ensemble filters that has proven to work specially well: the IPF filter [12].

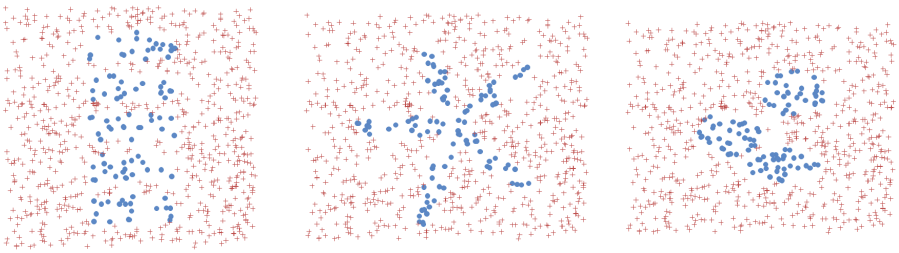
IPF removes noisy examples in multiple iterations until a stopping criterion is reached. The iterative process stops when, for a number of consecutive iterations k, the number of identified noisy examples in each of these iterations is less than a percentage p of the size of the original training dataset. Initially, the method starts with a set of noisy examples $A = \emptyset$. The basic steps of each iteration are the following:

1. Split the current training dataset E into n equal sized subsets.
2. Build a classifier with the C4.5 algorithm over each of these n subsets and use them to evaluate the whole current training dataset E.
3. Add to A the noisy examples identified in E using a voting scheme (consensus or majority).
4. Remove the noisy examples: $E \leftarrow E \setminus A$.

Two voting schemes can be used to identify noisy examples: consensus and majority. The former removes an example if it is misclassified by all the classifiers, whereas the latter removes an example if it is misclassified by more than half of the classifiers.

The parameter setup for the implementation of IPF used in this work has been determined experimentally in order to better fit it to the characteristics of imbalanced datasets with noisy and borderline examples once they have been preprocessed with SMOTE. More precisely, the majority scheme is used to identify the noisy examples, $n = 9$ partitions with random examples in each one are created and $k = 3$ iterations for the stop criterion and $p = 1\%$ for the percentage of removed examples are considered.

In short, the SMOTE algorithm balances the class distribution and it helps to fill in the interior of sub-parts of the minority class whereas IPF removes the noisy examples originally present in the dataset and also those created by SMOTE cleaning up the boundaries of the classes making them more regular.



(a) Sub-cluster (b) Clover (c) Paw

Fig. 1. Shapes of the minority class

3 Experimental Analysis

In this section we present the details of the experimental study developed. Section 3.1 shows how the synthetic imbalanced datasets with borderline examples were built. Then Section 3.2 presents the results and their analysis.

3.1 Datasets and Re-sampling Techniques for Comparison

This paper uses the family of synthetic datasets earlier used in research on the role of borderline examples [14] and generated by special software and evaluated [16] on the role of borderline examples for different basic classifiers, such as C4.5, and re-sampling methods.

The synthetic datasets consist of two classes (the minority versus the majority class) with examples randomly and uniformly distributed in the two-dimensional real-value space. Datasets with 600 examples and $IR = 5$ and datasets with 800 examples and $IR = 7$ are considered.

Three different shapes of the minority class examples are used. *Sub-cluster* (Figure 1a) where the examples from the minority class are located inside rectangles following related works on small disjuncts. *Clover* (Figure 1b) represents a more difficult, non-linear setting, where the minority class resembles a flower with elliptic petals. In *Paw* (Figure 1c) the minority class is decomposed into 3 elliptic subregions of varying cardinalities, where two subregions are located close to each other, and the smaller sub-region is separated.

By increasing the ratio of borderline examples, i.e., the disturbance ratio, from the minority class subregions we consider 5 levels of DR: 0%, 30%, 50%, 60% and 70%. The width of the borderline overlapping areas is comparable to the width of the safe parts of sub-regions.

The proposal SMOTE-IPF is compared using these datasets against well-known re-sampling techniques to adjust the class distribution: SMOTE-ENN [1] is *filtering-based* approach based on extending SMOTE with an additional filtering, whereas SL-SMOTE [4] is based on directing the creation of the positive examples (*change-direction* methods). The aforementioned preprocessing techniques will be analyzed comparing the AUC results obtained by C4.5 for our

Table 1. AUC results obtained by C4.5 on synthetic datasets

Dataset	SMOTE	SMOTE-ENN	SL-SMOTE	SMOTE-IPF
sub-cluster $_{IR=5,\ DR=0}$	**93.00**	90.00	91.50	92.70
sub-cluster $_{IR=5,\ DR=30}$	81.90	81.80	82.20	**83.40**
sub-cluster $_{IR=5,\ DR=50}$	**80.80**	77.80	80.40	77.80
sub-cluster $_{IR=5,\ DR=60}$	78.60	77.00	78.10	**79.70**
sub-cluster $_{IR=5,\ DR=70}$	77.50	77.00	**82.30**	80.10
sub-cluster $_{IR=7,\ DR=0}$	94.79	93.07	90.79	**95.21**
sub-cluster $_{IR=7,\ DR=30}$	81.57	79.64	81.71	**83.00**
sub-cluster $_{IR=7,\ DR=50}$	78.29	75.29	**81.29**	78.57
sub-cluster $_{IR=7,\ DR=60}$	80.21	75.71	**81.36**	79.79
sub-cluster $_{IR=7,\ DR=70}$	**82.50**	78.21	81.21	80.93
clover $_{IR=5,\ DR=0}$	83.50	**86.80**	85.70	85.50
clover $_{IR=5,\ DR=30}$	83.80	83.40	81.10	**85.30**
clover $_{IR=5,\ DR=50}$	**83.40**	81.00	83.00	82.40
clover $_{IR=5,\ DR=60}$	80.80	80.50	78.70	**82.20**
clover $_{IR=5,\ DR=70}$	77.20	76.10	77.10	**79.00**
clover $_{IR=7,\ DR=0}$	87.93	87.79	88.21	**91.64**
clover $_{IR=7,\ DR=30}$	83.64	83.14	84.36	**85.71**
clover $_{IR=7,\ DR=50}$	81.21	**82.86**	78.71	81.93
clover $_{IR=7,\ DR=60}$	78.79	77.71	77.21	**82.14**
clover $_{IR=7,\ DR=70}$	78.29	76.07	78.29	**80.21**
paw $_{IR=5,\ DR=0}$	**96.30**	94.90	92.80	94.50
paw $_{IR=5,\ DR=30}$	84.40	**86.40**	84.50	84.90
paw $_{IR=5,\ DR=50}$	84.60	83.30	**85.00**	84.90
paw $_{IR=5,\ DR=60}$	81.00	81.30	82.30	**83.30**
paw $_{IR=5,\ DR=70}$	82.80	**83.50**	82.70	83.40
paw $_{IR=7,\ DR=0}$	93.43	93.93	92.93	**94.29**
paw $_{IR=7,\ DR=30}$	84.29	84.93	83.50	**85.29**
paw $_{IR=7,\ DR=50}$	85.00	84.79	84.29	**85.79**
paw $_{IR=7,\ DR=60}$	**83.36**	80.71	83.00	82.14
paw $_{IR=7,\ DR=70}$	82.57	80.57	84.14	**85.71**

approach against applying SMOTE alone, SMOTE-ENN and SL-SMOTE. Additionally, statistical comparisons in each of these cases will be also performed using Wilcoxon's signed ranks statistical test [6].

3.2 Results on Synthetic Datasets

Table 1 presents the AUC results obtained by C4.5 on each synthetic dataset when preprocessing with each re-sampling approach considered in this paper. The best case for each dataset is remarked in bold. From these results, we can observe that increasing DR, fixing a shape of the minority class and an IR, strongly deteriorates the performance of C4.5 in all cases. SMOTE-IPF obtains better results than the rest of the re-sampling methods in 16 of the 30 datasets considered and obtains results close to the best performances in the rest of the cases. Highest improvements of SMOTE-IPF are obtained in non-linear datasets, since 12 of the 16 overall best performance results are obtained in them.

Table 2 collects the results of applying Wilcoxon's signed ranks statistical test between SMOTE-IPF versus the re-sampling techniques. As the p-values and the sums of ranks show, SMOTE-IPF produces a statistically significant improvement in the results obtained. From these results we can conclude that SMOTE-IPF performs better than other SMOTE versions when dealing with

Table 2. Wilcoxon's test results for the comparison of SMOTE-IPF (R^+) versus SMOTE variants (R^-) from AUC results obtained by C4.5

Methods	R^+	R^-	$p_{Wilcoxon}$
SMOTE-IPF vs. SMOTE	366.0	99.0	0.0050
SMOTE-IPF vs. SMOTE-ENN	408.0	27.0	< 0.0001
SMOTE-IPF vs. SL-SMOTE	362.5	102.5	0.0070

the synthetic imbalanced datasets built with borderline examples, particularly in those with non-linear shapes of the minority class.

4 Concluding Remarks

This work proposes to extend SMOTE by a new element, the IPF noise filter, to control the presence of noisy and borderline examples and the noise introduced by the balancing between classes produced by SMOTE and to make the class boundaries more regular. Synthetic imbalanced datasets with different shapes of the minority class, imbalance ratios and levels of borderline examples have been used to analyze the suitability of this approach. All these datasets have been preprocessed with SMOTE-IPF and several well-known re-sampling techniques.

AUC values using C4.5 over the preprocessed datasets and the supporting statistical test have shown that our proposal has a notably better performance when dealing with imbalanced datasets with noisy and borderline examples, especially in the non-linear synthetic datasets. This work opens future efforts on analyzing the proposal on real datasets with class and attribute noise, as well as using more and different classifiers to check the suitability of the proposed method.

Acknowledgment. Supported by the National Project TIN2011-28488 and Regional Projects P10-TIC-06858, P11-TIC-9704, P12-TIC-2958 and NCN-2013 /11/B/ST6/00963. José A. Sáez holds an FPU scholarship.

References

1. Batista, G., Prati, R., Monard, M.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter 6(1), 20–29 (2004)
2. Bhowan, U., Johnston, M., Zhang, M.: Developing new fitness functions in genetic programming for classification with unbalanced data. IEEE T. Syst. Man Cy. B 42(2), 406–421 (2012)
3. Brodley, C.E., Friedl, M.A.: Identifying Mislabeled Training Data. Journal of Artificial Intelligence Research 11, 131–167 (1999)
4. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 475–482. Springer, Heidelberg (2009)

5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357 (2002)
6. Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. J. Mach. Learn. Res. 7, 1–30 (2006)
7. Gamberger, D., Lavrac, N., Dzeroski, S.: Noise Detection and Elimination in Data Preprocessing: experiments in medical domains. Appl. Artif. Intell. 14, 205–223 (2000)
8. Gamberger, D., Boskovic, R., Lavrac, N., Groselj, C.: Experiments With Noise Filtering in a Medical Domain. In: Proceedings of the Sixteenth International Conference on Machine Learning, pp. 143–151. Morgan Kaufmann Publishers (1999)
9. García, V., Alejo, R., Sánchez, J.S., Sotoca, J.M., Mollineda, R.A.: Combined effects of class imbalance and class overlap on instance-based classification. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 371–378. Springer, Heidelberg (2006)
10. He, H., Garcia, E.: Learning from imbalanced data. IEEE T. Knowl. Data En. 21(9), 1263–1284 (2009)
11. Kermanidis, K.L.: The effect of borderline examples on language learning. J. Exp. Theor. Artif. In. 21, 19–42 (2009)
12. Khoshgoftaar, T.M., Rebours, P.: Improving software quality prediction by noise filtering techniques. J. Comput. Sci. Technol. 22, 387–396 (2007)
13. Kubat, M., Matwin, S.: Addresing the curse of imbalanced training sets: one-side selection. In: Proc. of the 14th Int. Conf. on Machine Learning, pp. 179–186 (1997)
14. Napierała, K., Stefanowski, J., Wilk, S.: Learning from imbalanced data in presence of noisy and borderline examples. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS, vol. 6086, pp. 158–167. Springer, Heidelberg (2010)
15. Sáez, J.A., Luengo, J., Herrera, F.: Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification. Pattern Recogn. 46(1), 355–364 (2013)
16. Stefanowski, J.: Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. In: Ramanna, S., Howlett, R.J. (eds.) Emerging Paradigms in ML and Applications. SIST, vol. 13, pp. 277–306. Springer, Heidelberg (2013)
17. Verbaeten, S., Van Assche, A.: Ensemble methods for noise elimination in classification problems. In: Windeatt, T., Roli, F. (eds.) MCS 2003. LNCS, vol. 2709, pp. 317–325. Springer, Heidelberg (2003)