A preprocessing approach for class-imbalanced data using SMOTE and belief function theory

Fares Grina[1(✉)], Zied Elouedi[1], and Eric Lefevre[2]

[1] Institut Supérieur de Gestion de Tunis, LARODEC, Université de Tunis, Tunisia
grina.fares2@gmail.com, zied.elouedi@gmx.fr
[2] Univ. Artois, UR 3926, Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A), F-62400 Béthune, France
eric.lefevre@univ-artois.fr

Abstract. Dealing with imbalanced datasets at the preprocessing level is an efficient strategy used by many methods to re-balance the data and improve classification performance. Specifically, SMOTE is a popular oversampling technique which modifies the training data by adding artificial minority samples. However, SMOTE may create instances in noisy and overlapping areas, far from safe regions. To tackle this issue, we propose SMOTE-BFT, in which we use the belief function theory to remove generated minority instances that are not in safe regions. After applying SMOTE, each generated minority instance is represented by an evidential membership structure, which provides detailed information about class memberships. Rules based on the belief function theory are then enforced to detect and remove generated instances that are in noisy and overlapping regions. Experiments on noisy artificial datasets show that our proposal significantly outperforms other popular oversampling methods.

Keywords: Imbalanced data · Supervised learning · belief function

1 Introduction

Learning from imbalanced data is one of the main challenges in machine learning. In a binary classification problem, the imbalanced data issue occurs when one class (the minority or positive class) is represented by a much smaller number of instances than the other class (the majority or negative class). This problem presents a crucial difficulty to many classifier learning algorithms that assume a fairly balanced distribution of the classes [11]. The imbalanced data problem can be translated to numerous real-world classification problems [13]. From an application point of view, the correct classification of minority instances has a greater importance than the reverse [4]. For example, in a medical diagnosis problem, the cases affected by the disease are usually relatively rare as compared with the normal population. Assuming we have 1% of disease-affected cases, a classifier which scores correctly all negative samples (cases not affected) will get a classification accuracy of 99% even though all positive cases remain undetected.

Instead of accuracy, one may use the Area under the Curve (AUC-ROC) [2], which reflects how good a model is at distinguishing between the minority and majority class.

In coping with this issue, resampling methods have been proposed to rebalance the data-set by adding new samples to the minority class (Oversampling), removing samples from the majority class (Undersampling), or both (Hybrid) [9]. Traditional replication strategies (e.g. random oversampling) [9] can cause overfitting by simply adding replicated samples to the dataset. To avoid this issue, Chawla et al. [1] suggested the Synthetic Minority Oversampling Technique (SMOTE), which adds new synthetic minority instances by interpolating among several minority examples that are close to each other. However, SMOTE produces synthetic samples without being aware of its surroundings which may potentially amplify the noise and overlap present in the data as illustrated in Figure 1, whereas, studies have shown that generated instances in safe positions improve classifier's performance [3].

Several extensions have been proposed to deal with those problems. To name a few, BorderlineSMOTE [10] was introduced to strengthen the decision boundaries of classifiers, by replacing SMOTE's random selection of minority samples with a directed selection of examples that are close to the class border. ADASYN [12], another oversampling method which decides the number of examples to generate based on density distribution. Other hybrid methods like SMOTE-RSB* [15] (based on rough set theory) and SMOTE-IPF [16] (based on iterative partitioning filter) uses SMOTE and undersampling to clean generated minority examples and noise already present in the dataset.

In our work, we present another improvement of SMOTE based on the belief function theory. After applying SMOTE, a cleaning procedure is executed to improve SMOTE's oversampling, that is, we use the belief function theory as a way to extract information on the surroundings of each synthetic minority instance, by assigning a soft label regarding class memberships. Three rules based on belief function theory are then imposed to identify and eliminate synthetic minority instances which are not in safe regions. It is important to note that our proposal is a purely oversampling method, meaning that only generated minority instances can be removed.

The remaining of the paper is structured as follows. In Section 2, we introduce the belief function theory. Section 3 presents our proposal in details. In section 4, we define the experimental framework and we analyze the results.

2 Belief Function theory

Belief function theory [6,17,18], also known as evidence theory or the Dempster-Shafer theory is a well-founded and efficient framework for the representation and combination of a range of uncertain information. Let $\Omega = \{w_1, w_2, ..., w_M\}$ be a frame of discernment representing a finite set of M events. A basic belief assignment (BBA) represents the belief committed to the elements of 2^Ω by a

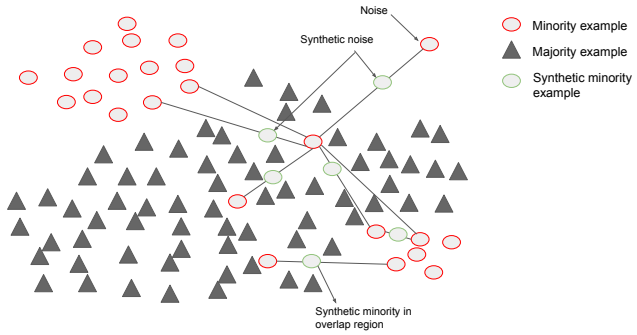

Fig. 1. SMOTE limitations (noise generation and introduction of more overlapping)

given source of evidence is a mapping function $m : 2^{\Omega} \rightarrow [0, 1]$ such that:

$$\sum_{A \in 2^{\Omega}} m(A) = 1 \tag{1}$$

Belief and *plausibility* functions are defined by *Shafer* [17] as follows:

$$Bel(A) = \sum_{B \subseteq A} m(B) \quad and \quad Pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \quad \forall \ A \in 2^{\Omega} \tag{2}$$

$Bel(A)$ represents the precise support for A and its subsets, whereas $Pl(A)$ represents the total possible support for A and its subsets. The interval $[Bel(A), Pl(A)]$ reflects the lower and upper bounds of support to A.

To combine several BBAs, *Dempster*'s rule [6] is a popular choice. Let m_1 and m_2 two BBAs defined on the same frame of discernment Ω, their combination based on *Dempster*'s rule gives the following BBA:

$$m_1 \oplus m_2(A) = \begin{cases} \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)} & for \ A \neq \emptyset \ and \ A \in 2^{\Omega}. \\ 0 & for \ A = \emptyset. \end{cases} \tag{3}$$

3 Combining SMOTE and belief function theory

We focus our method on binary classification. For multi-class cases, we can apply it by decomposing the multi-class problem into two-class sub-problems [5]. SMOTE-BFT consists of 3 main steps: First, we apply SMOTE to generate synthetic minority examples. Second, we compute for each generated instance an evidential soft label (BBA) using its nearest neighbors. Finally, 3 rules based on *plausibility* and *belief* functions are enforced in order to identify synthetic minority instances that are generated in noisy regions, overlapping regions, or majority class regions. This elimination procedure is repeated until a user-set minimum balance ratio Br_{min} is reached. Each step will be detailed in the remaining of this section.

3.1 Step 1: Applying SMOTE

Synthetic minority instances are firstly created by randomly selecting minority sample $\overrightarrow{a}$. Second, it searches for its k nearest minority neighbors $\overrightarrow{b}$. Finally, the algorithm chooses one of the neighbors and creates a synthetic point $\overrightarrow{s}$ anywhere on the line joining the selected sample and its chosen neighbor:

$$\overrightarrow{s} = \overrightarrow{a} + w * (\overrightarrow{b} - \overrightarrow{a}) \ , \ \ w \in [0, 1] \tag{4}$$

For this paper, we use the original version of SMOTE [1] to simplify the comparisons. However, one can use any other variant of SMOTE, since our proposal has a modular structure.

3.2 Step 2: Creating BBAs

To assign BBA's, we use the same evidential modeling defined by Denoeux in the Evidential KNN [8]. Let x_i be a synthetic minority instance obtained at Step 1. Each of its k neighbors (k is defined by the user) represents a piece of evidence to the evidential membership of the instance. For each neighbor x_j, a mass function m_i^j is calculated regarding the class membership of x_i as:

$$\begin{cases} m_i^j(\{w_q\}\,|x_j) = \alpha\phi_q(d_{ij}) \\ m_i^j(\Omega|x_j) = 1 - \alpha\phi_q(d_{ij}) \\ m_i^j(A|x_j) = 0 \qquad\qquad \forall A \in 2^\Omega \setminus \{\{w_q\}, \Omega\} \end{cases} \tag{5}$$

where d_{ij} denotes the euclidean distance between x_i and x_j, w_q is the class label of x_j, and α is a parameter such that $0 < \alpha < 1$. A recommended value of $\alpha = 0.95$ can be used to obtain good results on average, and a good choice for ϕ_q is $\phi_q(d) = exp(-\gamma_q d^2)$ where γ_q can be set to the inverse of the mean squared distance between training samples belonging to class w_q heuristically.

The BBAs for each neighbor x_j are then combined using the Dempster's rule defined in eq (3). As a result, each synthetic minority example x_i has three masses namely: $m_i(\{\omega_1\})$ degree of membership for the majority class, $m_i(\{\omega_2\})$ for the minority class and $m_i(\Omega)$ regarding both classes. Using these masses, it is now possible to compute *plausibility* and *belief* functions defined in eq (2).

3.3 Step 3: Eliminating synthetic examples

In order to perform the cleaning, we use overlap and noise rules that were introduced in [8] in addition to a misclassification rule. Each rule targets a specific type of synthetic points that are problematic to the classification task.

Overlap threshold rule: This situation typically arises when the synthetic minority sample is situated in a region where there is strong overlap between classes. In the belief function framework, this case is characterized by a BBA that is uniformly distributed between the two classes. As a result, the maximum

plausibility $Pl_{max} = max_{\omega \in \Omega} Pl_i(\{\omega\})$ will have a relatively low value. Thus, imposing a threshold to the maximum *plausibility* will reject synthetic instances in strong overlap regions. The sample will be rejected if:

$$Pl_{max} < \beta_{Pl}, \quad \beta_{Pl} \in [0, 1] \tag{6}$$

where β_{Pl} a threshold that is set by the user. The higher this parameter is, the more synthetic instances are removed.

Noise threshold rule: This situation represents synthetic points which are suspected of belonging to a class which is not represented in the training set. In the belief function framework, most of the mass values will be concentrated on the whole frame of discernment Ω. As a consequence, the maximum credibility $Bel_{max} = max_{\omega \in \Omega} Bel_i(\{\omega\})$ will take on a small value. As the distance between the synthetic point and its closest neighbors goes to infinity, Bel_{max} goes to zero. Thus, the generated sample will be rejected if:

$$Bel_{max} < \beta_{Bel}, \quad \beta_{Bel} \in [0, 1] \tag{7}$$

where β_{Bel} a threshold which is reasonably fixed to the minimum value of Bel_{max} across original minority samples.

Misclassifcation rule: This situation represents synthetic samples which are more likely to be misclassified after oversampling, meaning that they are located in the majority class region. Using BBAs, one way to make a decision about what class a sample belongs to is the maximum *plausibility*. In our situation, we want generated examples to be belonging to the minority class. Let w_1 be the majority class and w_2 be the minority one. The synthetic minority x_i example will be rejected if:

$$Pl_i(\{\omega_1\}) > Pl_i(\{\omega_2\}) \tag{8}$$

The cleaning phase is iterated over synthetic examples until the data reach a minimum balance ratio Br_{min} set by the user.

4 Experimental study

In this section, we discuss the evaluation of our proposal in details. Section 4.1 presents the experimental setup. Section 4.2 shows the results and analysis.

4.1 Experimental setup

To evaluate our contribution, we use synthetic imbalanced datasets with noisy and borderline examples selected from the KEEL repository [14]. All datasets are binary classification problems with combinations of two imbalance ratios (IR), 5 levels of disturbance ratios DR (reflects the amout of overlap) and three types of

non-linear shapes of minority examples: *Clover*, which shapes a flower with five elliptic petals, *paw*, which shapes three elliptic subareas of minority examples, and *subclus*, which has 3 rectangles of minority instances.

Results are averaged through a ten-fold stratified cross validation to avoid inconsistencies. To compare different oversamplers, we use CART for classification and the AUC-ROC as an evaluation measure. Statistical comparisons are also performed using the Wilcoxon's signed ranks test [7] to compare the results.

We compare our proposal, SMOTE-BFT, with popular oversampling techniques: the original SMOTE [1], Borderline-SMOTE [10] and ADASYN [12]. In addition to the evaluation against no oversampling performed. For the experiments, we consider the following parameters for SMOTE-BFT: number of nearest neighbors for SMOTE and the creation of BBAs fixed to 5, minimum plausibility threshold β_{Pl} is set to 0.7, and minimum imbalance ratio Br_{min} is fixed to 0.8.

4.2 Results analysis

Table 1 presents AUC obtained by CART on each dataset after oversampling. The best scores are marked in bold. We can observe that oversampling techniques improve the performance of the CART classifier in almost all cases.

The proposed SMOTE-BFT achieves the best AUC scores in 13 out of 30 datasets and obtains close to the best scores in the other cases. Even though CART performance worsens with higher DR, we can notice that our method performed relatively better especially when the dataset presents high percentage of DR. This shows that our method successfully identified and eliminated the noise and overlap generated by SMOTE.

In order to compare AUC results, statistical analysis was performed using the Wilcoxon's signed ranks test, which is a non-parametric pairwise test used to identify significant differences in performance of two methods [7]. In our study, we use this test to compare the performance obtained by SMOTE-BFT against other oversamplers. $R+$ represents the sum of ranks in favor of SMOTE-BFT, $R-$, the sum of ranks in favor of the other compared methods, and p-values are obtained for each comparison. As seen in Table 2, all p-values for Wilcoxon's test are lower than 0.05, which shows that SMOTE-BFT outperforms all compared methods at a significance level of $\alpha = 0.05$.

5 Conclusions

In this paper, we have proposed a new extension to SMOTE aiming at improving the quality of oversampling. Rules have been developed using the belief function theory to remove synthetic minority examples which are added in noisy, overlapping or majority class regions. Results from the experimental analysis show that SMOTE-BFT obtains significantly better results than compared methods, specifically on datasets with high disturbance ratios. Future work can include developing heuristic methods in order to automatically determine parameters such as the plausibility threshold β_{Pl} and the minimum balance ratio Br_{min}.

Table 1. AUC results of CART on synthetic datasets oversampled by different methods.

Datasets	NONE	SMOTE	BorderSM	ADASYN	SMOTE-BFT
$paw_{(S=600,IR=5,DR=0\%)}$	0.962	0.964	0.952	**0.972**	0.97
$paw_{(S=600,IR=5,DR=30\%)}$	0.79	0.817	**0.844**	**0.844**	0.835
$paw_{(S=600,IR=5,DR=50\%)}$	0.747	0.77	0.796	0.794	**0.798**
$paw_{(S=600,IR=5,DR=60\%)}$	0.679	0.731	0.714	0.742	**0.766**
$paw_{(S=600,IR=5,DR=70\%)}$	0.69	0.721	0.771	0.754	**0.778**
$paw_{(S=800,IR=7,DR=0\%)}$	0.94	0.952	**0.957**	0.948	0.943
$paw_{(S=800,IR=7,DR=30\%)}$	0.805	0.819	0.836	0.830	**0.837**
$paw_{(S=800,IR=7,DR=50\%)}$	0.749	0.778	0.752	0.757	**0.795**
$paw_{(S=800,IR=7,DR=60\%)}$	0.693	**0.762**	0.733	0.731	0.739
$paw_{(S=800,IR=7,DR=70\%)}$	0.634	0.721	**0.767**	0.718	0.745
$clover_{(S=600,IR=5,DR=0\%)}$	0.859	0.909	0.889	**0.91**	0.862
$clover_{(S=600,IR=5,DR=30\%)}$	0.755	0.814	**0.824**	0.817	0.823
$clover_{(S=600,IR=5,DR=50\%)}$	0.69	0.744	0.738	0.787	**0.79**
$clover_{(S=600,IR=5,DR=60\%)}$	0.696	0.728	0.705	**0.744**	**0.744**
$clover_{(S=600,IR=5,DR=70\%)}$	0.646	0.739	0.72	0.734	**0.753**
$clover_{(S=800,IR=7,DR=0\%)}$	0.845	0.905	0.869	**0.923**	0.92
$clover_{(S=800,IR=7,DR=30\%)}$	0.766	**0.836**	0.825	0.835	0.832
$clover_{(S=800,IR=7,DR=50\%)}$	0.704	0.749	0.757	**0.765**	0.763
$clover_{(S=800,IR=7,DR=60\%)}$	0.668	0.712	0.704	**0.743**	0.737
$clover_{(S=800,IR=7,DR=70\%)}$	0.659	0.727	0.725	0.724	**0.747**
$subcl_{(S=600,IR=5,DR=0\%)}$	**0.976**	0.942	0.955	0.952	0.954
$subcl_{(S=600,IR=5,DR=30\%)}$	0.822	0.811	0.82	0.801	**0.826**
$subcl_{(S=600,IR=5,DR=50\%)}$	0.715	0.728	0.734	**0.741**	0.738
$subcl_{(S=600,IR=5,DR=60\%)}$	0.671	0.728	**0.736**	0.729	0.731
$subcl_{(S=600,IR=5,DR=70\%)}$	0.688	0.746	0.718	0.71	**0.757**
$subcl_{(S=800,IR=7,DR=0\%)}$	**0.978**	0.955	0.963	0.972	0.968
$subcl_{(S=800,IR=7,DR=30\%)}$	0.793	**0.842**	0.797	0.811	0.810
$subcl_{(S=800,IR=7,DR=50\%)}$	0.702	0.753	**0.756**	0.722	0.739
$subcl_{(S=800,IR=7,DR=60\%)}$	0.702	0.724	0.742	0.739	**0.763**
$subcl_{(S=800,IR=7,DR=70\%)}$	0.605	0.716	0.699	0.704	**0.72**

Table 2. Wilcoxon's signed rank test results comparing SMOTE-BFT (R+) with other oversampling methods (R-).

Comparisons	R+	R-	P-value
SMOTE-BFT vs NONE	453.0	12.0	< 0.00001
SMOTE-BFT vs SMOTE	370.0	95	0.00467
SMOTE-BFT vs BorderSM	368.0	97	0.00530
SMOTE-BFT vs ADASYN	352.5	112.5	0.02307

References

1. Bowyer, K.W., Chawla, N.V., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. CoRR **abs/1106.1813** (2011), http://arxiv.org/abs/1106.1813
2. Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition **30**(7), 1145 – 1159 (1997)
3. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Advances in Knowledge Discovery and Data Mining. PAKDD 2009. Lecture Notes in Computer Science, vol 5476. pp. 475-482
4. Chawla, N., Japkowicz, N., Kołcz, A.: Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explorations **6**, 1–6 (06 2004)
5. Chen, K., Lu, B.L., Kwok, J.T.: Efficient classification of multi-label and imbalanced data using min-max modular classifiers. In: The 2006 IEEE International Joint Conference on Neural Network Proceedings. pp. 1770–1775. IEEE (2006)
6. Dempster, A.P.: A generalization of bayesian inference. Journal of the Royal Statistical Society: Series B (Methodological) **30**(2), 205–232 (1968)
7. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research **7**(Jan), 1–30 (2006)
8. Denoeux, T.: A k-nearest neighbor classification rule based on dempster-shafer theory. Systems, Man and Cybernetics, IEEE Transactions on **219**, 804 – 813 (06 1995)
9. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications **73**, 220–239 (2017)
10. Han, H., Wang, W.Y., Mao, B.H.: Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Lecture Notes in Computer Science **3644**(PART I), 878–887 (2005)
11. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263–1284 (2009)
12. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks. pp. 1322–1328. IEEE (2008)
13. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence **5**(4), 221–232 (2016)
14. Napierała, K., Stefanowski, J., Wilk, S.: Learning from imbalanced data in presence of noisy and borderline examples. In: International Conference on Rough Sets and Current Trends in Computing. pp. 158–167. Springer (2010)
15. Ramentol, E., Caballero, Y., Bello, R., Herrera, F.: SMOTE-RSB *: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowledge and Information Systems **33**(2), 245–265 (2012)
16. Sáez, J.A., 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. Information Sciences **291**(C), 184–203 (2015)
17. Shafer, G.: A mathematical theory of evidence. Princeton university press (1976)
18. Smets, P.: The transferable belief model for quantified belief representation. In: Quantified Representation of Uncertainty and Imprecision, pp. 267–301 (1998)