NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

Undersampling with support vectors for multi-class imbalanced data classification

Krawczyk, Bartosz; Bellinger, Colin; Corizzo, Roberto; Japkowicz, Nathalie

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.
For the publisher's version, please access the DOI link below./ Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

Publisher's version / Version de l'éditeur:
https://doi.org/10.1109/IJCNN52387.2021.9533379
Proceedings of IJCNN 2021, 2021-09-20

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=bd3b4916-7e94-4b0a-9901-3ce3ef0461eb
https://publications-cnrc.canada.ca/fra/voir/objet/?id=bd3b4916-7e94-4b0a-9901-3ce3ef0461eb

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada | Conseil national de recherches Canada

Canada

Undersampling with Support Vectors for Multi-Class Imbalanced Data Classification

Bartosz Krawczyk
Department of Computer Science
Virginia Commonwealth University
Richmond, VA, USA
bkrawczyk@vcu.edu

Colin Bellinger
National Research Council
Ottawa, Canada
colin.bellinger@nrc-cnrc.gc.ca

Roberto Corizzo, Nathalie Japkowicz
Department of Computer Science
American University
Washington, DC, USA
{rcorizzo,japkowic}@american.edu

***Abstract*—Learning from imbalanced data poses significant challenges for the classifier. This becomes even more difficult, when dealing with multi-class problems. Here relationships among classes are no longer well-defined and it is easy to loose performance on one of the classes while gaining on other. In last years this topic has gained increased interest from the machine learning community - however, still there is a need for developing new and efficient algorithms to handle this challenge. In this paper we propose a new approach for balancing multi-class imbalanced problems. It is based on a two-step undersampling methodology. In the first step, a one-class classifier is being trained on each of the classes, achieving skew-insensitive data description. Support vectors for each class are extracted and used as new class representatives, thus achieving significant reduction in the terms of used instances. In the second step, an evolutionary undersampling approach is being used on these support vectors in order to further balance the training set. By applying this technique on a set of support vectors and not on a full dataset, we achieve a significant reduction of the computational time and increased accuracy. Finally, a standard multi-class classifier is being trained on the balanced data set. A thorough experimental study proves the usefulness of the proposed approach in comparison with state-of-the-art approaches for handling multi-class imbalanced data.**

***Index Terms*—Machine learning, Imbalanced data classification, Multi-class imbalance, Undersampling.**

I. Introduction

In machine learning and data mining, while one or more classes are underrepresented in the data set, it is called as class imbalance classification. Many real-world classification tasks suffer from the class imbalance problem, which is considered as one of the important challenges for the data mining community [1]. The main difficulty of these problems is that the skewed distribution makes conventional classification algorithms less effective, since standard learning algorithms consider a balanced training data set, which result in making it harder to predict minority class examples [2].

In recent years, many efforts have been focused on the binary class imbalance problems [3], which only contain two classes. However, multi-class imbalance classification, is widely applied in many areas, such as text categorization [4], human activity recognition [5] and medical diagnosis [6]. Unfortunately, it may be invalid to directly apply the solutions proposed for the two-class problems to the multi-class imbalance problems, and some algorithms cannot be used to solve the multi-class imbalance problems directly [1].

Fortunately, in the research community, decomposition strategies turn up to deal with multi-class classification problem. In this solution framework, the multi-class classification problems are transformed into binary class sub-problems, which are much easier to discriminate [7]. Such well-known approaches are the one versus one (OVO) [8] and one versus all (OVA) [9]. As OVA introduces an artificial class imbalance (e.g., for 10 class problem with roughly equally represented classes, the binary sub-problem will have an imbalance ratio 1:9), it is not advisable to use it for handling problems with initially skewed distributions [1]. Recent works show that using a multi-class decomposition with one-class classifiers offers significant benefits when dealing with a high number of classes and data-level difficulties [10]. Those methods offer interesting applications for imbalanced datasets.

Research goal. To propose an efficient undersampling algorithm for multi-class imbalanced data that can offer at least comparable performance to existing oversampling methods dedicated to this problem.

Motivation. Multi-class imbalanced data classification has gained a significant attention from the research community in recent years [1]. However, most of existing algorithms for handling skewed classes focus on oversampling approaches [11], [12]. Undersampling has proven itself to be very useful in binary imbalanced problems [13], [14], where it can alleviate many limitations of oversampling (e.g., increasing class overlapping, enhancing noise presence, or shifting class distributions) [15]. Currently there is a lack of dedicated undersampling algorithms that take into account presence of multiple classes and can assume dependencies among them.

Summary. We introduce a novel approach for handling mutli-class imbalanced data using a two-step undersampling. In the first phase a given M-class imbalanced dataset is decomposed by training a One-Class Support Vector Machine (OCSVM) [16] on each of classes, thus resulting in a pool of M one-class classifiers. They are used not for the classification purposes, but as data description tools. Then, for each classifier we extract support vectors that are used to represent the target concept. This allows us to maintain the most vital

samples from the original training set, at the same time leading to a significant decrease in the number of instances being considered. However, the number of support vectors for each class may still be skewed and thus a second phase is needed. Here, we apply an evolutionary undersampling (EUS) [17] approach, where a genetic algorithm is used to further reduce the number of support vectors. It works in a wrapper mode with given classifier, using its predictive power as an optimization criterion. This allows for further balancing of the feature set. By taking advantage of the one-class reduction phase, EUS computational complexity is greatly reduced, as it needs only to deal with smaller number of instances.

Main contributions. This paper offers following developments in the field of learning from multi-class imbalanced data:

- **Support Vector Undersampling.** We introduce OCSV-US, a novel two-phase undersampling for multi-class imbalanced data. It uses a one-class decomposition for extraction of core support vectors for each class and uses them as input prototypes for evolutionary undersampling.
- **Robust bias elimination for multi-class imbalanced data.** The proposed two-step undersampling is capable of selecting core instances for each class, especially focusing on class boundaries and borderline instances. At the same time, it displays robustness to outliers and redundant examples.
- **Extensive experimental study.** We prepared a thorough experimental study, where we compare the proposed OCSV-US with state-of-the-art oversampling algorithms. This allows us to empirically prove that undersampling for multi-class imbalanced data is capable of performing as well, and frequently significantly better, than oversampling - especially when facing various data-level difficulties.

II. Learning from imbalanced data

Algorithms for imbalanced classification problems. Cannonical classification algorithms are designed under an assumption of a balanced training set [18]. With such a precondition, it is challenging to deal with class imbalance problems, especially for identifying the most important minority class instances [19]. Most of the proposed solutions are designed only to address the binary class imbalance problems [20]. To overcome the dilemma of skewed class distribution, a large amount of techniques have been developed to deal with such problem. They can be categorized into three groups:

- Data-level solutions: the origin of the problem is the class distribution in the data sets, therefore, it is natural to consider of rebalancing by sampling the data space to reduce the impact of class imbalance [21], leading to a balanced and well-represented training set [22]. One of the advantages of such solution is independent from the classifier used, so they are also considered as pre-processing methods [18].
- Algorithm-level solutions: these solutions try to understand what mechanisms within the training procedure lead to bias towards majority observations. By understanding what is the cause of such an unwanted behavior, one may attempt to make this specific mechanism skew-insensitive. Algoirthm-level solutions are considered as internal modifications [23], since the effect depends on the problems and the classifier [1]. One of the most well-known solutions is the direct modification of the learning procedure for a selected algorithm [24]. Cost-sensitive solutions are another family of popular algorithm-level approaches that assign a higher costs for misclassifying the minority class instances, thus penalizing errors on smaller distributions [25]. The learning process aims at minimizing the classification cost, resulting in a more balanced decision boundary.
- Ensemble solutions: tackling class imbalance by combining ensemble learning [26] with one of the two previously mentioned strategies. This leads to creating a balanced training sets for base classifiers and managing the diversity among ensemble components [27], which can be realized as classifier combination [28] or selection [29].

Multi-class imbalanced data. Multi-class imbalance problems are significantly more challenging, since large number of classes needs to be analyzed and the relationships among these classes are complicated. Conventional solutions designed for two-class problems may be no longer feasible or underperform, being unable to model this much more challenging problem. There is still not a lot of dedicated multi-class approaches and more research in this area is needed. Static-SMOTE [30] applied resampling procedure in M steps, where M is the number of classes. In each iteration, the resampling procedure selects the minimum size class, and duplicates the number of instances of the class in the original data-set. Important trend in this domain points to a high importance of considering the individual types of minority classes examples and their learning difficulty when performing oversampling for multi-class imbalanced data and proposes a data-driven universal strategy that can be embedded in any data-level multi-class solution [31]. Recently proposed oversampling algorithms focus on utilizing information coming from multiple classes at once [11] and reducing the impact of overlapping and noisy instances [32].

III. Support Vector Underdsampling for multi-class imbalanced data

In this section the details of the proposed two-phase undersampling for mining multi-class imbalanced data will be presented.

Phase 1: Extracting class-specific support vectors. In the first phase, we propose to handle a M-class multi-class imbalanced dataset by training M one-class Support Vector Machines (OCSVM) on each class respectively. Each base classifier aims at adjusting itself to the given target class.

In one-class classification (OCC) the classifier is fit in such a way, that will allow for a best possible separation form potential outliers. On the other hand, OCC algorithms tend to avoid overfitting by not having to tight description around the data. This is an especially important feature in multi-class decomposition, as it allows to discriminate between the classes and at the same time preserves the generalization abilities of the base classifiers. Additionally, by using instances coming from a single class, one-class classifiers are naturally skew-insensitive.

However, in this work we propose not to use OCSVMs directly for a multi-class classification, but as a data pre-processing tool. Each OCSVM will return a set of support vectors - selected class instances that have the biggest influence on the shape of boundary enclosing this given concept. Therefore, one may assume that these instances carry the highest information value for the considered class and thus should be preserved during the data sampling phase. This allows us to carry a data-driven undersampling, replacing original instances with support vectors.

Firstly, let us present the OCSVM model which will be used for support vector extraction. The idea behind OCSVM is to find a hyperplane $\langle \mathbf{w}, x \rangle + \rho$ that separates the training data from the origin with the maximal margin. We can formulate this problem as a convex optimization task:

$$Q(\mathbf{w}, \xi_1, \cdots, \xi_\iota, \rho) = \min \left(\frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{\nu\iota} \sum_{i=1}^{\iota} \xi_i - \rho \right) \quad (1)$$

subject to:

$$\forall i \in [1, \iota] \quad \langle \mathbf{w}, x \rangle \geq \rho - \xi_i, \quad (2)$$

where $\| \cdot \|$ denotes the Euclidean norm, $[\xi_1, \cdots, \xi_\iota]$ are the slack variables that must satisfy the condition $\xi_i \geq 0$ and ν is the penalization factor incurred by these slack variables.

The corresponding Wolfe dual is subject to optimization:

$$Q(\alpha_1, \cdots, \alpha_\iota) = min \left(\sum_{i=1}^{\iota} \sum_{j=1}^{\iota} \alpha_i \alpha_j \langle x_i, x_j \rangle \right), \quad (3)$$

subject to:

$$\forall i \in [1, \iota] \quad 0 \leq \alpha_i \leq 1/(\nu\iota), \quad (4)$$

and

$$\sum_{i=1}^{\iota} \alpha_i = 1, \quad (5)$$

where $\langle \cdot, \cdot \rangle$ is the inner product and $[\alpha_1, \cdots, \alpha_\iota]$ are Lagrange multipliers.

One may denote the solution of the problem from Eq. 3 as $[\alpha_1^*, \cdots, \alpha_\iota^*]$. Then according to $\mathbf{w} = \sum_i \alpha_i^* x_i$, we may compute the distance between a new point x and the separating hyperplane using:

$$dst(x, \omega_T) = \sum_{i=1}^{\iota} \alpha_i^* \langle x_i, x \rangle - \rho. \quad (6)$$

Parameter ρ can be calculated with the usage of property that for any α_i^* satisfying $0 < \alpha_i^* < 1/(\nu\iota)$, the corresponding example x_i satisfies:

$$\rho = \sum_{j=1}^{\iota} \alpha_i^* \langle x_i, x_j \rangle. \quad (7)$$

OCSVM makes its decision about a new object x on the basis of distance $dst(x, \omega_T)$. This can be seen as a measure of similarity between the object and the target class. This formulation uses only inner product. It allows for an easy kernelization, by replacing each inner product $\langle x_i, x_j \rangle$ by the selected kernel function $K(x_i, x_j)$.

After M such models are being trained, we extract support vector returned by them, leading to an undersampling of the original multi-class dataset. Let us now discuss the advantages of such an approach.

Phase 2: Evolutionary selection of support vectors. EUS is an evolutionary prototype selection algorithm adapted to work in imbalanced domains (it uses an appropriate fitness function). Prototype selection is a sampling process aiming at reducing the reference set for the classifier in order to improve its accuracy and reduce the storage necessity. However, in an imbalanced scenario the objective differs, since the balance of the data distribution gains importance. On this account, EUS tries to obtain a useful undersampled data-set whose search is guided by a genetic algorithm. Initially, several randomly undersampled data subsets are created, which are then evolved until the currently best undersampled data-set cannot be further improved in terms of the fitness function. Likewise in every evolutionary method, the way in which the solutions are represented by means of chromosomes is an important issue. In EUS, a binary vector is used to represent each solution, where each gene (binary value) represents the presence or absence of the corresponding instance in the data set.

In our approach we propose to use a multi-class modification of EUS to further undersample the set of support vectors obtained from phase 1 of our approach. Using such an initially reduced set as an input, we simplify the search space for the evolutionary approach. Therefore, a chromosome is represented as follows:

$$V = (SV_1, SV_2, \cdots, SV_M) \quad (8)$$

where

$$SV_n = (sv_{x_1}, sv_{x_2}, sv_{x_3}, sv_{x_4}, \ldots, sv_{x_{n^-}}), \quad (9)$$

where sv_{x_i} takes the values 0 or 1, indicating whether support vector using instance x_i is included or not in the data-set (n^- stands for the number of support vectors in the n-th class).

In the evolutionary process, chromosomes are ranked using a fitness function, which in the case of EUS takes into account

the balanced accuracy of a given wrapper classifier on a multi-class imbalanced problem. It is given as follows:

$$\text{fitness}_{EUS} = \frac{\sum_{n=1}^{M} TPR_n}{M}, \tag{10}$$

where M is the number of classes of the dataset and TPR_n is the True Positive Rate of the n-th class (noted in percentage).

In order to perform the search, the well-known CHC algorithm is used due to its good balance between exploration and exploitation. CHC is an elitist genetic algorithm using the heterogeneous uniform cross-over (HUX) to combine two chromosomes (exactly half of the different genes are interchanged). An incest prevention mechanism is also considered where two parents are only recombined if their Hamming distance is greater than the threshold (initially $L/4$, being L the length of the chromosome); the threshold is reduced by one when no parents are recombined. In this genetic algorithm no mutation is applied, but when the recombined chromosomes are not able to improve their parents and the threshold reaches zero, the whole population (except for the best chromosome) are reinitialized. Reinitialization consists of using the best chromosome as a template, randomly changing 35% of its genes.

Advantages of the proposed Support Vector Undersampling. Let us discuss the advantages of the proposed Support Vector Undersampling approach for the task of multi-class imbalanced data classification:

- **Class boundary preservation.** Support vectors extracted from OCSVMs represent class boundaries, thus making this undersampling approach robust to removing valuable instances from the training set.
- **Reduced prototype redundancy.** OCSVMs will select support vectors that create an enclosing boundary around the target class. From the point of view of multi-class classification we are just interested in instances lying closest to boundaries among different classes. Therefore, redundant support vectors may be excluded.
- **Speeding-up the evolutionary selection.** Using support vectors as an input for EUS instead of the entire training set holds two advantages. Firstly, it reduced the computational complexity of the evolutionary search procedure. Secondly, it improves the efficiency of the search, as it is initialized by already pre-selected instances.
- **Robustness to extreme class imbalance.** As each OCSVM is trained independently then there will be no influence of the imbalance factor on the location of support vectors - minority classes will have assigned the same importance as majority ones.
- **Robustness to outliers and noise.** OCSVMs can exclude outliers from becoming support vectors, thus increasing the robustness of the method to noisy multi-class imbalanced datasets.

IV. Experimental study

The experimental study was designed to answer the following research questions:

RQ1: Can undersampling offer comparable, or better performance as oversampling for multi-class imbalanced data?

RQ2: Does the proposed extraction of support vectors for initialization of evolutionary undersampling leads to improved robustness to skewed distributions and decreased classifier training time?

A. Setup

Data benchmarks. In this study, twenty multi-class imbalanced data sets from the UCI repository were selected to test the methodology. The properties of the data sets were showed in Table I.

Reference methods. Throughout the conducted experiments the proposed method was compared with a selection of state-of-the-art multi-class data resampling algorithms. Reference methods include STATIC-SMOTE (S-SMOTE) [30], Mahalanobis Distance Oversampling (MBO) [33], and (k-NN)-based synthetic minority oversampling algorithm (SMOM) [34]. Additionally, we present results for multi-class EUS initialized with the entire dataset (MC-EUS), without support vector pre-selection. Parameters of the reference methods used throughout the experimental study were presented in Table II.

Base classifiers. All examined resampling algorithms for multi-class imbalance data use C5.0 as a base classifier.

Classifier validation procedure. For the purpose of training and testing of examined classifiers, we have employed a stratified 10-fold cross validation. All parameters for each tested algorithm were established for each dataset independently using an internal stratified 3-fold cross validation on the training set.

Performance metrics. We evaluate the performance of classifiers on multi-class imbalanced data using four dedicated skew-insensitive metrics [35]: Average Accuracy (AvAcc), Class Balanced Accuracy (CBA), multi-class G-measure (mGM), and Confusion Entropy (CEN). They are expressed as follows:

$$AvAcc = \frac{\sum_{i=1}^{M} TPR_i}{M} \tag{11}$$

$$CBA = \frac{\sum_{i=1}^{M} \frac{mat_{i,i}}{max\left(\sum_{i=1}^{M} mat_{i,j}, \sum_{i=1}^{M} mat_{j,i}\right)}}{M} \tag{12}$$

$$mGM = \sqrt[M]{\prod_{i=1}^{M} precision_i \cdot recall_i} \tag{13}$$

$$CEN = \sum_{i=1}^{M} P_i \cdot CEN_i, \tag{14}$$

$$\tag{15}$$

TABLE I: Summary description of the data sets used in the experimental study.

id	Dataset	#Instances	#Features	#Classes	Per-class distribution	IR
Bal	Balance	625	4	3	288/49/288	5.88
Car	Car	1728	6	4	384/69/1210/65	18.62
Cle	Cleveland	297	13	5	160/54/35/35/13	12.31
Con	Contraceptive	1473	9	3	629/333/511	1.89
Der	Dermatology	358	34	6	111/60/71/48/48/20	5.55
Fla	Flare	1066	11	6	147/211/239/95/43/331	7.70
Hay	Hayes-roth	160	4	3	65/64/31	2.10
Led	Led7digit	500	7	10	45/37/51/57/52/52/47/57/53/49	1.54
New	New-thyroid	215	5	3	150/35/30	5.00
Pag	Page-blocks	5472	10	5	4913/329/28/87/115	175.46
Sat	Satimage	6435	36	6	1533/703/1358/626/707/1508	2.45
Spl	Splice	3190	60	3	767/768/1655	2.16
Thy	Thyroid	720	21	3	17/37/666	39.18
Win	Wine	178	13	3	59/71/48	1.48
Wqr	Wine-Quality-Red	1599	11	6	10/53/681/638/199/18	68.10

TABLE II: Parameters of the classification and the sampling algorithms used throughout the experimental study.

Algorithm	**Parameters**

OCSV-US	kernel type = RBF;
	C = $\in [0.5, 1.0, \cdots, 5]$;
	γ = $\in [0.001, 0.02, \cdots, 0.01]$;
	frac. rejected = 0.05;
	population size = $\in [10, 20, \cdots, 100]$;
	number of eval. = $\in [100, 200, \cdots, 1000]$;
	HUX = $\in [0.05, 0.10, \cdots, 0.30]$

S-SMOTE [30]	k-nearest neighbors = $\in [3, 5, \cdots, 11]$; oversampling ratio $\in [50, 100, \cdots, 500]$
MDO [33]	$K1 \in [1, 2, \cdots, 10]$; $K2 \in [2, 4, \cdots, 20]$; oversampling ratio $\in [50, 100, \cdots, 500]$
SMOM [34]	$K1 \in [2, 4, \cdots, 20]$; $K2 \in [1, 2, \cdots, 10]$; $rTh \in [0.1, 0.2, \cdots, 1]$; $rTh \in [1, 2, \cdots, 10]$; $w1, w2, r1, r2 \in [0.1, 0.2, \cdots, 1]$; k-nearest neighbors = 5; oversampling ratio $\in [50, 100, \cdots, 500]$

TABLE III: Average accuracy (AvAcc) results for examined methods. The best result is highlighted in bold.

Dataset	S-SMOTE	MBO	SMOM	MC-EUS	OCSV-US
Bal	66.92	85.81	71.23	83.45	**86.71**
Car	**85.08**	54.83	69.41	82.18	82.98
Cle	30.33	31.11	29.68	30.62	**32.04**
Con	47.36	47.21	44.75	47.82	**48.13**
Der	95.00	**96.45**	92.38	90.56	91.27
Fla	60.57	54.45	59.88	61.06	**62.39**
Hay	**85.39**	76.42	82.49	78.74	78.89
Led	**72.10**	59.45	70.05	69.72	70.68
New	88.42	89.05	**91.33**	90.00	90.35
Pag	**80.75**	48.97	70.09	77.60	78.41
Sat	87.49	88.16	**90.11**	88.38	88.86
Spl	94.14	74.10	87.03	94.82	**95.94**
Thy	95.46	64.97	90.32	96.83	**98.05**
Win	93.13	97.58	95.64	97.37	**98.42**
Wqr	35.73	31.84	33.61	39.41	**41.07**
Avg. rank	2.65	3.20	3.05	4.15	1.95

where M is the number of clases, $mat_{i,j}$ stands for the number of instances of the true class i that were predicted as class j, $P_i = \frac{\sum_{j=1}^{M} mat_{i,j}+mat_{j,i}}{2\cdot\sum_{i,l=1}^{M} mat_{k,l}}$, and $CEN_i = -\sum_{j=1,i\neq j}^{M}\left(P_{i,j}^{i}\log_{2(C-1)}(P_{i,j}^{i}) + P_{j,i}^{i}\log_{2(C-1)}(P_{j,i}^{i})\right)$. Additionally, for CEN we have $P_{i,i}^{i} = 0$, $P_{i,j}^{i} = mat_{i,j}/\left(\sum_{j=1}^{C} mat_{i,j} + mat_{j,i}\right)$, and $i \neq j$.

Statistical analysis. Friedman ranking test and Wilcoxon post-hoc test with significance level $\alpha = 0.05$ were used to compare examined algorithms over multiple datasets.

B. Results and discussion

Comparison with state-of-the-art resampling methods. Tables III – VI depict the performance of OCSV-US and reference algorithms according to four performance metrics, while Table VII present the outcomes of Wilcoxon post-hoc statistical analysis. We can see that the proposed two-step undersampling leads to improved results compared to all of reference methods on the vast majority of examined benchmarks. This confirms the observations made in literature on binary datasets that undersampling may lead to improvements in model accuracy. This can be contributed to a selective sampling of the classes and selecting most relevant instances, while all of remaining techniques do not take into consideration the nature of examples. Interestingly, OCSV-US is capable of outperforming oversampling approaches - which shows that they suffer from well-known limitations such as incorrect placement of artificial instances or lack of robustness to noise [1]. The proposed undersampling alleviates those limitations, leading to creating compact, yet accurate representations of multiple classes.

Investigating the role of evolutionary prototype selection. When compared to a undersampling initialized with the entire training set, our proposed methods display significantly improved accuracy on all four used metrics (see Tables III – VI and Tab. VII). This shows that the smart initialization of the search procedure will lead to improved instance representation and robustness in multi-class undersampling. Table VIII presents the average training time of C5.0 classifier on undersampling from full dataset and reduced dataset

TABLE IV: Class Balanced Accuracy (CBA) results for examined methods. The best result is highlighted in bold.

Dataset	S-SMOTE	MBO	SMOM	MC-EUS	OCSV-US
Bal	62.81	81.23	70.01	79.92	**84.33**
Car	**82.11**	50.02	64.28	77.29	81.14
Cle	27.49	28.92	26.44	27.99	**30.13**
Con	43.65	42.88	40.07	42.99	**45.20**
Der	92.01	92.45	**92.58**	87.47	90.05
Fla	56.82	49.87	54.91	57.61	**60.01**
Hay	**81.28**	72.47	77.52	73.24	76.18
Led	68.92	54.77	66.51	63.66	**69.99**
New	85.19	83.98	**86.72**	84.11	86.08
Pag	**77.45**	41.90	66.80	70.91	75.22
Sat	82.87	84.18	85.82	84.91	**87.51**
Spl	90.07	71.11	82.99	89.13	**92.66**
Thy	90.03	59.89	89.71	90.82	**93.61**
Win	88.86	91.74	92.19	90.08	**95.91**
Wqr	30.81	27.31	30.89	34.44	**38.19**
Avg. rank	2.95	3.60	3.35	3.55	1.55

TABLE V: Multi-class G-measure (mGM) results for examined methods. The best result is highlighted in bold.

Dataset	S-SMOTE	MBO	SMOM	MC-EUS	OCSV-US
Bal	64.21	82.98	69.21	68.08	**84.11**
Car	**83.21**	51.92	66.98	77.11	81.21
Cle	27.81	29.48	27.51	24.33	**30.17**
Con	45.08	44.92	42.66	43.04	**46.09**
Der	92.77	**93.84**	90.11	89.72	90.48
Fla	58.02	51.93	56.74	54.19	**60.88**
Hay	**82.49**	73.28	80.14	74.21	77.30
Led	**70.02**	57.19	66.48	65.37	68.93
New	85.82	86.21	**89.64**	84.71	88.72
Pag	**77.93**	47.91	67.22	66.89	75.14
Sat	85.36	85.36	**87.91**	84.28	85.90
Spl	91.68	71.03	84.52	87.69	**92.88**
Thy	92.73	61.18	87.02	86.79	**96.17**
Win	89.81	94.08	93.17	92.81	**96.45**
Wqr	31.88	29.64	30.39	31.82	**40.33**
Avg. rank	2.65	3.20	3.05	4.15	1.95

TABLE VI: Confusion Entropy (CEN) results for examined methods. The best result is highlighted in bold.

Dataset	S-SMOTE	MBO	SMOM	MC-EUS	OCSV-US
Bal	0.34	0.18	0.31	0.21	**0.16**
Car	**0.18**	0.48	0.36	0.23	0.20
Cle	0.78	0.72	0.74	0.73	**0.67**
Con	0.57	0.59	0.60	0.58	**0.51**
Der	0.08	0.07	**0.06**	0.13	0.09
Fla	0.44	0.51	0.46	0.43	**0.38**
Hay	**0.22**	0.29	0.24	0.26	0.23
Led	0.34	0.46	0.35	0.37	**0.30**
New	0.15	0.17	**0.13**	0.17	0.15
Pag	**0.33**	0.69	0.46	0.48	0.35
Sat	0.18	0.16	015	0.17	**0.11**
Spl	0.09	0.28	0.16	0.13	**0.06**
Thy	0.08	0.42	0.14	0.11	**0.07**
Win	0.12	0.11	0.10	0.10	**0.05**
Wqr	0.71	0.73	0.72	0.74	**0.58**
Avg. rank	3.25	3.50	3.15	3.85	1.25

TABLE VII: Wilcoxon post-hoc test for comparison between the proposed approach and reference methods according to four evaluation metrics. Symbol '+' stands for situation in which our proposal is superior and '=' for classifiers without significant differences.

Hypothesis	p-value			
	AvAcc	CBA	mGM	CEN
OCSV-US vs S-SMOTE	= (0.1406)	+ (0.0479)	= (0.1723)	+ (0.0417)
OCSV-US vs MDO	+ (0.0089)	+ (0.0106)	+ (0.0094)	+ (0.0055)
OCSV-US vs SMOM	= (0.0710)	+ (0.0395)	= (0.0603)	+ (0.0402)
OCSV-US vs MC-EUS	+ (0.0399)	+ (0.0252)	+ (0.0403)	+ (0.0106)

TABLE VIII: Training time [s.] comparison between EUS using the full training set and a proposed two-step approach .

Dataset	MC-EUS	OCSV-US
Bal	104.21	33.18
Car	208.41	59.74
Cle	54.02	12.05
Con	138.49	50.06
Der	46.43	11.27
Fla	113.25	30.88
Hay	20.89	5.72
Led	45.93	10.49
New	29.89	9.58
Pag	397.95	98.49
Sat	502.74	130.69
Spl	284.01	80.22
Thy	112.03	35.09
Win	19.48	8.27
Wqr	145.62	54.11

(support vector prototypes). When comparing computational times required to run tested procedures, one may clearly see the advantage of using reduced support vector representation instead of the whole data set. Training M OCSVMs and using EUS on support vector representation is still much faster than using EUS on the full dataset. Furthermore, one must remember that undersampling is a filter method. Therefore, the balancing procedure is conducted once and then we may train any number of classifiers based on the balanced dataset. The showed times saving will become even more important, when multiple classifiers need to be trained on the balanced dataset for a large-scale study.

V. Conclusions and Future Works

Concluding summary. In this paper, we have presented a novel approach for handling multi-class imbalanced data based on undersampling. We proposed a two-step procedure for obtaining a reduced set of meaningful instances. In the first phase, the M-class imbalanced problem was decomposed using M one-class Support Vector Machines. They were used to obtain support vectors for each class, which would represent their boundaries. Such a set of support vectors was used to initialize a multi-class evolutionary undersampling in order to further balance the training set and remove redundant vectors. Finally, a standard multi-class classifier was trained on this balanced dataset. Experimental results confirmed high

accuracy returned by this approach and computational gains from using two-step reduction.

Future works. Multi-class imbalanced data undersampling can be further potentially improved by including additional information about instance-level and class-level difficulties during the selection procedure. Understanding what learning difficulties are specific to each considered problem may lead to more accurate predictions which regions in the decision space should be more or less densely represented. Finally, understanding why some instances were retained and other removed may lead to explainable approach for learning from imbalanced data.

REFERENCES

[1] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in AI*, vol. 5, no. 4, pp. 221–232, 2016.

[2] S. Das, S. Datta, and B. B. Chaudhuri, "Handling data irregularities in classification: Foundations, trends, and future challenges," *Pattern Recognition*, vol. 81, pp. 674–693, 2018.

[3] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 31:1–31:50, 2016.

[4] J. Tian, S. Chen, X. Zhang, and Z. Feng, "A graph-based measurement for text imbalance classification," in *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, ser. Frontiers in Artificial Intelligence and Applications, vol. 325. IOS Press, 2020, pp. 2188–2195.

[5] A. A. Alani, G. Cosma, and A. Taherkhani, "Classifying imbalanced multi-modal sensor data for human activity recognition in a smart home using deep learning," in *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*. IEEE, 2020, pp. 1–8.

[6] J. Kaufmann, K. Asalone, R. Corizzo, C. Saldanha, J. Bracht, and N. Japkowicz, "One-class ensembles for rare genomic sequences identification," in *Discovery Science - 23rd International Conference, DS 2020, Thessaloniki, Greece, October 19-21, 2020, Proceedings*, ser. Lecture Notes in Computer Science, A. Appice, G. Tsoumakas, Y. Manolopoulos, and S. Matwin, Eds., vol. 12323. Springer, 2020, pp. 340–354.

[7] P. Trajdos and M. Kurzynski, "A correction method of a binary classifier applied to multi-label pairwise models," *Int. J. Neural Syst.*, vol. 28, no. 9, pp. 1 750 062:1–1 750 062:19, 2018.

[8] Q. Li, Y. Song, J. Zhang, and V. S. Sheng, "Multiclass imbalanced learning with one-versus-one decomposition and spectral clustering," *Expert Syst. Appl.*, vol. 147, p. 113152, 2020.

[9] X. Gao, Y. He, M. Zhang, X. Diao, X. Jing, B. Ren, and W. Ji, "A multiclass classification using one-versus-all approach with the differential partition sampling ensemble," *Eng. Appl. Artif. Intell.*, vol. 97, p. 104034, 2021.

[10] B. Krawczyk, M. Galar, M. Wozniak, H. Bustince, and F. Herrera, "Dynamic ensemble selection for multi-class classification with one-class classifiers," *Pattern Recognit.*, vol. 83, pp. 34–51, 2018.

[11] B. Krawczyk, M. Koziarski, and M. Wozniak, "Radial-based oversampling for multiclass imbalanced data classification," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 8, pp. 2818–2831, 2020.

[12] W. C. Sleeman and B. Krawczyk, "Multi-class imbalanced big data classification on spark," *Knowl. Based Syst.*, vol. 212, p. 106598, 2021.

[13] M. Koziarski, "Radial-based undersampling for imbalanced data classification," *Pattern Recognition*, vol. 102, p. 107262, 2020.

[14] P. Vuttipittayamongkol and E. Elyan, "Neighbourhood-based undersampling approach for handling imbalanced and overlapped data," *Inf. Sci.*, vol. 509, pp. 47–70, 2020.

[15] J. A. Sáez, M. Galar, and B. Krawczyk, "Addressing the overlapping data problem in classification using the one-vs-one decomposition strategy," *IEEE Access*, vol. 7, pp. 83 396–83 411, 2019.

[16] S. Alam, S. K. Sonbhadra, S. Agarwal, and P. Nagabhushan, "One-class support vector classifiers: A survey," *Knowl. Based Syst.*, vol. 196, p. 105754, 2020.

[17] I. Triguero, M. Galar, H. Bustince, and F. Herrera, "A first attempt on global evolutionary undersampling for imbalanced big data," in *2017 IEEE Congress on Evolutionary Computation, CEC 2017, Donostia, San Sebastián, Spain, June 5-8, 2017*. IEEE, 2017, pp. 2054–2061.

[18] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Springer, 2018. [Online]. Available: https://doi.org/10.1007/978-3-319-98074-4

[19] L. Korycki and B. Krawczyk, "Online oversampling for sparsely labeled imbalanced and non-stationary data streams," in *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*. IEEE, 2020, pp. 1–8.

[20] S. Sharma, C. Bellinger, B. Krawczyk, O. R. Zaïane, and N. Japkowicz, "Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance," in *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*. IEEE Computer Society, 2018, pp. 447–456.

[21] C. Bellinger, R. Corizzo, and N. Japkowicz, "Remix: Calibrated resampling for class imbalance in deep learning," *CoRR*, vol. abs/2012.02312, 2020. [Online]. Available: https://arxiv.org/abs/2012.02312

[22] C. Bellinger, P. Branco, and L. Torgo, "The CURE for class imbalance," in *Discovery Science - 22nd International Conference, DS 2019, Split, Croatia, October 28-30, 2019, Proceedings*, ser. Lecture Notes in Computer Science, P. K. Novak, T. Smuc, and S. Dzeroski, Eds., vol. 11828. Springer, 2019, pp. 3–17.

[23] P. S. Akash, M. E. Kadir, A. A. Ali, and M. Shoyaib, "Inter-node hellinger distance based decision tree," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, Ed. ijcai.org, 2019, pp. 1967–1973.

[24] P. Ksieniewicz, "Standard decision boundary in a support-domain of fuzzy classifier prediction for the task of imbalanced data classification," in *Computational Science - ICCS 2020 - 20th International Conference, Amsterdam, The Netherlands, June 3-5, 2020, Proceedings, Part IV*, ser. Lecture Notes in Computer Science, vol. 12140. Springer, 2020, pp. 103–116.

[25] C. Zhang, K. C. Tan, H. Li, and G. S. Hong, "A cost-sensitive deep belief network for imbalanced classification," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 1, pp. 109–122, 2019.

[26] M. Wozniak, M. Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Inf. Fusion*, vol. 16, pp. 3–17, 2014.

[27] J. Díez-Pastor, J. J. Rodríguez, C. I. García-Osorio, and L. I. Kuncheva, "Diversity techniques improve the performance of the best imbalance learning ensembles," *Inf. Sci.*, vol. 325, pp. 98–117, 2015.

[28] S. Wojciechowski and M. Wozniak, "Employing decision templates to imbalanced data classification," in *Hybrid Artificial Intelligent Systems - 15th International Conference, HAIS 2020, Gijón, Spain, November 11-13, 2020, Proceedings*, ser. Lecture Notes in Computer Science, vol. 12344. Springer, 2020, pp. 120–131.

[29] P. Zyblewski, R. Sabourin, and M. Wozniak, "Preprocessed dynamic classifier ensemble selection for highly imbalanced drifted data streams," *Inf. Fusion*, vol. 66, pp. 138–154, 2021.

[30] F. Fernández-Navarro, C. Hervás-Martínez, and P. A. Gutiérrez, "A dynamic over-sampling procedure based on sensitivity for multi-class problems," *Pattern Recognition*, vol. 44, no. 8, pp. 1821–1833, 2011.

[31] J. A. Sáez, B. Krawczyk, and M. Wozniak, "Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets," *Pattern Recognition*, vol. 57, pp. 164–178, 2016.

[32] M. Koziarski, M. Wozniak, and B. Krawczyk, "Combined cleaning and resampling algorithm for multi-class imbalanced data with label noise," *Knowl. Based Syst.*, vol. 204, p. 106223, 2020.

[33] L. Abdi and S. Hashemi, "To combat multi-class imbalanced problems by means of over-sampling techniques," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 238–251, 2016.

[34] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognition*, vol. 72, pp. 327–340, 2017.

[35] P. Branco, L. Torgo, and R. P. Ribeiro, "Relevance-based evaluation metrics for multi-class imbalanced domains," in *Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part I*, ser. Lecture Notes in Computer Science, J. Kim, K. Shim, L. Cao, J. Lee, X. Lin, and Y. Moon, Eds., vol. 10234, 2017, pp. 698–710.

View publication stats