Multiclass Oversampling via Optimum-Path Forest for Tree Species Classification from Street-view Perspectives

Danilo Samuel Jodas*†, Leandro Aparecido Passos‡, Giuliana Del Nero Velasco†,
Mariana Hortelani Carneseca Longo†, Aline Ribeiro Machado†, João Paulo Papa*
*Department of Computing
São Paulo State University, Bauru-SP, Brazil 17033-360
Email: danilojodas@gmail.com, joao.papa@unesp.br
†Institute for Technological Research
University of São Paulo, São Paulo-SP, Brazil 05508-901
Email: {danilojodas,velasco,marihc,asribeiro}@ipt.br
‡CMI Lab, School of Engineering and Informatics
University of Wolverhampton, Wolverhampton, England, United Kingdom
Email: l.passosjunior@wlv.ac.uk

***Abstract*—Urban forest surveillance relies on several aspects that involve the analysis of green area preservation and the monitoring of individual trees. Urban trees are essential to maintain the good quality of the cities and reduce the effects of carbon dioxide emissions in the atmosphere. In this sense, one can cite the tree species diversity as essential to ensuring the preservation and proper functioning of the urban ecosystem and the conservation of the wildlife species in the urban forest environment. Furthermore, tree species play an essential role in assessing the tree risk of falling since the species are related to the wood density, thus providing further details for the tree structural analysis. However, tree species classification involves a time-consuming process that requires allocating human resources for fieldwork. Also, the tree species are quite imbalanced in the urban landscape, requiring a more efficient approach to provide accurate results for minority species. Therefore, computer-aided methods are helpful to support the rapid analysis of the tree species for tasks involving inventory and analysis of the tree conditions. This paper proposes a multiclass extension of the O^{2}PF, an Optimum-Path Forest-based oversampling method, to generate synthetic samples based on features extracted from images of five urban tree species. Further, we present the so-called "Street Level Tree Species Classification", a novel dataset for tree species classification based on tree images from the ground-view perspective. Four variants of the multiclass O^{2}PF were tested and compared to several state-of-the-art oversampling methods found in the literature. The obtained results confirm the effectiveness and superior accuracy of the proposed approaches in most cases.**

***Index Terms*—Optimum-Path Forest; deep learning; convolutional neural networks; tree classification; urban forest.**

I. Introduction

Tree management in the urban environment becomes essential for monitoring the green area conservation, analyzing the tree structure conditions, and cataloging the tree species for inventory purposes [1], [2]. Regarding the latter aspect, one can cite the counting of the tree species to assess the urban forest diversity. Furthermore, in combination with other physical measures, the species is one of the factors that assist the forecast of the possible risk of the rupture of the trunk and branches of the tree. However, manual observation and counting of the tree species is a laborious and time-consuming task that demands effort from the field staff. Also, trees in the urban landscape are diverse in species and sometimes imbalanced depending on aspects of the region's climate and geography. Therefore, computer-aided methods are desirable and essential to accelerate the tasks related to urban tree assessment.

In this context, machine learning and deep learning-based approaches have been successfully employed for similar tasks in several application domains. In forestry management, one can cite the tree detection and the tree health assessment in remote sensing-based images [3]–[5]. Furthermore, tree species classification is frequently employed for tree inventory and forest diversity analysis. Usually, tree species classification relies on satellite imagery and data derived from the aerial point of view. Regarding the latter strategy, Light Detection and Ranging (LiDAR) [6] and aerial RGB images [7] are the long-established data used to map vegetation regions and detect the tree canopies using image processing and machine learning-based algorithms. From the ground-view perspective, Terrestrial Laser Scanner (TLS) is also an alternative for urban tree species classification [8]–[10]. However, street-view RGB images are not fully explored yet beyond the detection and segmentation of trees. Street-level pictures are affordable, easy to capture, and accessible with reasonable resolutions. Furthermore, photographs from the ground-view perspective can offer additional data for the tree condition analysis, like the extraction of the tree dendrometry.

Regardless of the employed data for tree species classification, the performance of such models is severely impacted by a well-known drawback related to machine learning so-

lutions, i.e., data imbalance. The data imbalance problem is characterized by datasets composed of an unbalanced number of samples per class, i.e., some categories have much more instances than others, thus leading the model to familiarize itself with the majority classes' behavior and neglect patterns present in minority classes.

Many works addressed the abovementioned problem using data undersampling [11], [12], which consists of pruning the dataset and rebalancing it by removing selected samples, and data oversampling, which comprises a mechanism for synthetic data generation. Regarding the latter, some researchers addressed the problem using Generative Adversarial Networks [13], [14] in the context of artificial image generation. For general scenarios, Chawla et al. [15] proposed the Synthetic Minority Over-Sampling Technique (SMOTE), a successful method for data oversampling considering the interpolation of minority class samples to generate new instances. Further, several works improved the technique with a diverse range of variations [16]–[19], to cite a few. However, such methods usually produce some noise data since labels among all classes are not considered in the interpolation process.

Recently, Passos et al. [20] proposed an Optimum-Path Forest (OPF) [21]-based algorithm for data oversampling, namely O^{2}PF, to tackle this problem. The Optimum-Path Forest is a graph-based framework developed for supervised [21] and unsupervised [22] learning with successful application in a wide variety of fields [23]–[25]. Further, Passos et al. [26] proposed a set of OPF-based solutions for data imbalance, comprising four variants of the O^{2}PF, as well as four undersampling and three hybrid models. Despite the success observed in the experiments, such techniques lack in the sense that they are suitable for binary problems only.

Therefore, this paper proposes a novel dataset composed of street-level tree images, namely "Street Level Tree Species Classification", for the task of tree species classification. Further, it also proposes a multiclass oversampling approach based on the OPF framework for tree species classification. The proposed method extends the O^{2}PF model to handle the class imbalance issue in the multiclass domain, particularly in urban tree analysis. The technique employs the unsupervised OPF to capture the inherent aspects of the minority class samples. Afterward, new synthetic instances are created for the training set based on the Gaussian distribution of the minority sample's features. Compared to the previous binary O^{2}PF, the proposed approach is iteratively applied to all minority classes of the dataset.

This paper provides the following three contributions:

- To propose a model for tree species classification based on images captured from the ground-view perspective;
- To extend the O^{2}PF for oversampling in the multiclass domain and evaluate its impact on the class imbalance aspect of the tree species classification;
- To provide a novel dataset composed of street-level tree images, namely "Street Level Tree Species Classification."

The remainder of this paper is presented as follows. Section II provides a theoretical background regarding the unsupervised OPF and the O^{2}PF, as well as its variants, while Section III introduces the proposed approach for tree species classification from the ground-view perspective and the O^{2}PF variation for multiclass problems. Further, Section IV describes the novel **Street Level Tree Species Classification** dataset and the experimental setup. Finally, Sections V and VI state the experimental results and conclusions, respectively.

II. Theoretical Background

This section provides a brief description of the O^{2}PF and its variants, as well as the unsupervised OPF, the basis on which O^{2}PF was built.

A. Unsupervised Optimum-Path Forest

The main objective of the Optimum-Path Forest [22] unsupervised version is to represent every dataset's sample as a node in a graph for further grouping such instances into clusters with similar properties. The clustering procedure is conducted by connecting the training samples to their k-nearest neighbors through edges, whose weights are given by their distances d in the euclidian space. Besides, the nodes are also weighted by a probability density function (pdf), as follows:

$$\rho(\boldsymbol{s}) = \frac{1}{\sqrt{2\pi\phi^2}k} \sum_{\forall \boldsymbol{r} \in \mathcal{A}_k(\boldsymbol{s})} \exp\left(\frac{-d^2(\boldsymbol{s},\boldsymbol{r})}{2\phi^2}\right), \quad (1)$$

where ϕ is given by $\frac{m_w}{3}$, with m_w describing the maximum weight among all edges, and $\mathcal{A}_k(\boldsymbol{s})$ denotes the k-neighborhood of sample $\boldsymbol{s}$.

Supposing a graph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$, where $\mathcal{V}$ denotes the set of vertices (instances) and $\mathcal{A}$ the set of edges connecting such vertices, one can estimate the probability density by discovering the optimum number of nearest neighbors $k^* \in \{1 \leq k_{\max} \leq |\mathcal{V}|\}$, i.e., the value of k^* that minimizes the graph cut over $\mathcal{V}$. In this case, k_{max} is a hyperparameter representing the maximum possible value of k.

Further, the algorithm has to select a set of prototypes $\mathcal{P}$ containing one element per maximum of the pdf. The node $\boldsymbol{r}$ is attached to the path whose f_{min}, i.e., the minimum density value along its course, is maximum. The f_{min} value is computed as follows:

$$\begin{aligned} f_{min}(\langle \boldsymbol{r} \rangle) &= \begin{cases} \rho(\boldsymbol{r}) & \text{if } \boldsymbol{r} \in \mathcal{P} \\ \rho(\boldsymbol{r}) - \delta & \text{otherwise,} \end{cases} \\ f_{min}(\langle \phi_{\boldsymbol{s}} \cdot \langle \boldsymbol{s}, \boldsymbol{r} \rangle \rangle) &= \min\{f_{min}(\phi_{\boldsymbol{s}}), \rho(\boldsymbol{r})\}, \end{aligned} \quad (2)$$

where δ is a constant small value.

B. O^{2}PF

Recently, Passos et al. [20] proposed the O^{2}PF, an algorithm for data oversampling based on the unsupervised Optimum-Path Forest. The algorithm comprises two main steps to generate synthetic samples: (i) producing synthetic elements with

coherent characteristics; and (ii) introducing sample variability to avoid subsets of similar features. In the first step, the algorithm clusters the minority class training samples to extract intrinsic properties of the class, i.e., the samples' average position and variance. In the sequence, O^{2}PF assumes that all characteristics from a class follow a normal distribution, thus generating a new sample $\boldsymbol{z} \in \mathbb{R}^m$ by sampling such a distribution from some of the discovered clusters considering a proportion of samples by cluster size. The distribution is computed as follows:

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma), \tag{3}$$

where $\boldsymbol{\mu} \in \mathbb{R}^m$ stands for the cluster mean, m is the number of features per instance, and $\Sigma \in \mathbb{R}^{m \times m}$ is the covariance matrix, computed as follows:

$$\Sigma = \frac{1}{n-1}(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^T, \tag{4}$$

such that $X \in \mathbb{R}^{m \times n}$ is a matrix comprising a concatenation of all n cluster feature vectors.

Further, Passos et al. [26] proposed a set of methods to tackle data imbalance, comprising oversampling, undersampling, and hybrid approaches. Among such practices, the authors proposed four O^{2}PF variants, described as follows:

1) O^{2}PF$_{RI}$: The O^{2}PF with Radius Interpolation replaces the cluster mean by its geometric median, making the model more robust to outliers.
2) O^{2}PF$_{MI}$: The O^{2}PF with Mean Interpolation generates the new sample $\boldsymbol{z}$ and then interpolates it with its nearest neighbor within the cluster.
3) O^{2}PF$_P$: The O^{2}PF Prototype employs the prototype sample instead of using the cluster's mean as a parameter of the Gaussian distribution.
4) O^{2}PF$_{WI}$: The O^{2}PF with Weight Interpolation employs a strategy that weights each sample according to its density (Equation 1). Therefore, the cluster's mean stands for the weighted average of its samples. The sampling process remains the same as O^{2}PF$_{MI}$.

III. Proposed Approach

This section presents the proposed model for tree species classification and the multiclass oversampling approach based on the O^{2}PF.

A. Tree species classification

Figure 1 depicts the entire process of producing the feature vector of the input image. The image passes through a MobileNet CNN architecture that performs the feature extraction and dimensionality reduction using a set of pre-trained weights. The model has been previously trained to detect the tree elements in images captured from the ground view perspective of the urban landscape [27]. This process helps identify the inherent aspects of the trees while using transfer learning and inductive bias of the pre-trained network's weights. The last layer comprises $1,024$ neurons whose outcome is then used by classical machine learning algorithms to classify the tree species.

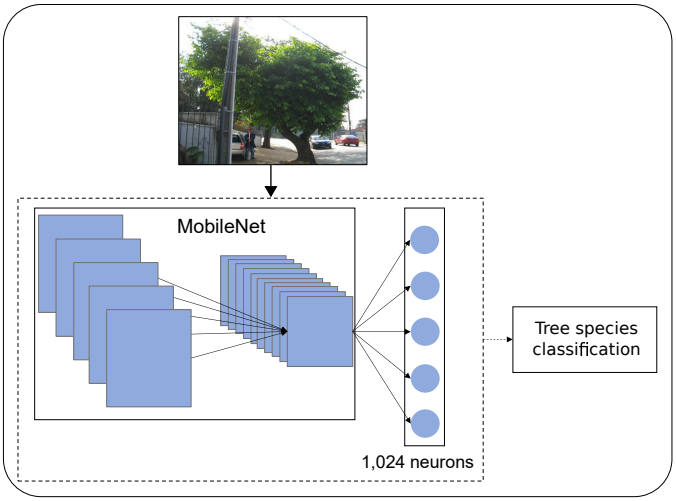

Fig. 1: Illustration of the proposed model for tree species classification: Top: an image of the entire tree; Bottom: illustration of the network proposed for feature extraction and tree species classification.

The feature extraction is performed by the same architecture used by Jodas et al. [28]. The architecture comprises convolutional blocks that execute a sequence of depthwise and pointwise convolutions with batch normalization and Rectified Linear Unit (ReLU) activation function. This approach reduces the network size without decreasing the predictions' efficacy [29]. Moreover, residual blocks are employed to avoid gradient vanishing and reinforcing the feature maps for the subsequent layers. Ultimately, the model ends with a Max Pooling and a Global Average Pooling on the last feature map. In total, the entire CNN model is composed of $3,228,864$ parameters.

B. Multiclass oversampling

The proposed approach extends the binary O^{2}PF to each minority class until the dataset is wholly balanced according to the number of samples in the majority class. The oversampling steps are summarized as follows:

1) Generate clusters from the minority class samples;
2) For each minority class cluster, estimate a Gaussian distribution according to Equation 3;
3) Generate n new synthetic samples using the mean $\boldsymbol{\mu}$ and the covariance matrix Σ as shown in Equation 3, where n stands for the number of samples in the majority class;
4) Repeat steps $1-3$ for each minority class of the dataset.

Compared to the binary O^{2}PF, the algorithm above includes a fourth step for the oversampling in the multiclass domain.

IV. Methodology

This section presents the proposed "Street Level Tree Species Classification" dataset and experimental setup adopted in the experiments.

A. Street Level Tree Species Classification Dataset

This paper proposes the "Street Level Tree Species Classification" dataset, which comprises 727 street-level photographs obtained from five species of trees in the city of São Paulo, Brazil [1]. The images are clippings from bounding boxes that enclose the trees in the whole photography (Please refer to Jodas et al. [27] for more details.). All images were resized to a 416×416 resolution to fit the input shape of the MobileNet architecture for feature extraction. Table I presents the number of images in each tree species.

TABLE I: Number of images in each tree species.

Species	ID	Number of images
Cenostigma pluviosum	CP	273
Holocalyx balansae	HB	48
Ligustrum lucidum	LL	46
Pleroma granulosum	PG	109
Tipuana tipu	TT	251

B. Experimental setup

The cross-validation procedure was used to split the image set into twenty folds of training and test sets with stratified sampling. Support Vector Machine (SVM) and Random Forest were used for the supervised classification of the original and the oversampled training sets at each fold. Notice that the standard OPF classifier was not employed to avoid biased results since the oversampling mechanism follows similar principles. A fine-tuning procedure was employed to find the best hyperparameter values that maximize the prediction performance of the machine learning models. Grid Search and Randomized Search are the gold-standard methods for hyperparameter optimization. However, Grid Search is time-consuming on large datasets and high hyperparameter spaces. Randomized Search is prone to high variance because of the random selection of the hyperparameter values, although it is relatively fast compared to Grid Search. The third approach is called Bayesian Optimization, and it relies on previous decisions to find the next set of hyperparameters that optimize an objective function [30]. In this sense, we perform the Bayesian Optimization to find the best hyperparameter values that optimize each classifier. An inner cross-validation approach is performed for each fold of the outer cross-validation. This method splits the training set into five folds of subsets for training and validation. The Bayesian Optimization is applied to the subsets at each inner fold to find the best hyperparameters that maximize the F1-Score. At last, the best-performing model is selected for predictions on the test set of the outer fold. This process is repeated for each fold of the external cross-validation. Figure 2 illustrates the strategy for one fold of the outer cross-validation.

Table II presents the hyperparameter space assigned to each classifier.

[1]https://github.com/recogna-lab/datasets/tree/master/TreeSpecies

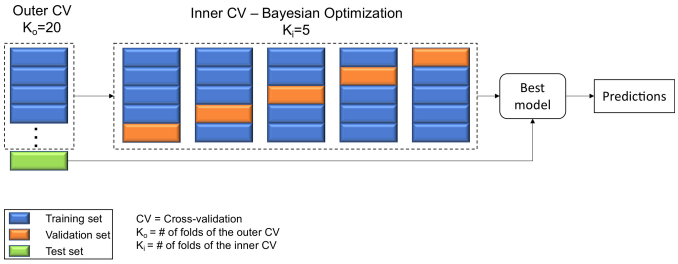

Fig. 2: Pipeline for the hyperparameter optimization.

TABLE II: Hyperparameter space for each classifier.

Model	Hyperparameter	Values
Random Forest	Number of trees	[10,100,200,500]
Support Vector Machine	Cost C	[1,2,3,...,20]
	Kernel	[Radial Basis Function, Sigmoid, Linear, Polynomial]
	Degree*	[1,2,3,...,20]

*Only applied to the Polynomial kernel

We compared the multiclass O^{2}PF and its variants against three state-of-the-art oversampling methods: Synthetic Minority Oversampling Technique (SMOTE) [15], BorderlineSMOTE [31], and Majority Weighted Minority Oversampling Technique (MWMOTE) [17]. The oversampling methods' hyperparameters where set to default values, i.e., $k_{max} = 5$ for the O^{2}PF and its variants, $k = 5$ for SMOTE [15] and BorderlineSMOTE [31], and $k_1 = k_2 = k_3 = 5$ considering MWMOTE [17]. We used the macro F1-Score as the standard metric to evaluate the effectiveness of the oversampling methods. The evaluation process was performed over 20 runs to compute the mean and the standard deviation from each oversampling technique and classifier. Then, the Wilcoxon signed-rank test was employed to evaluate the statistical similarity among O^{2}PF and the baselines over each classifier with 5% significance. Ultimately, we computed the confusion matrix from predictions obtained by the best-performing classifier and oversampling technique to confirm the prediction improvements of the minority class samples. For each external fold, a confusion matrix is computed from predictions on the test set. Afterward, the sum of all cross-validation matrices yields a general confusion matrix with hits and misses for each classifier.

The O^{2}PF relies on the Python-based OPFython library [32], and the source code is available at the Github repository [2].

V. Experiments and Results

Table III presents the average F1-Scores obtained from the original dataset after employing the oversampling techniques for both classifiers. The highest average values are stressed

[2]Available at https://github.com/Leandropassosjr/OpfImb

in bold, while the underlined values denote similar results according to the Wilcoxon signed-rank test for both classifiers, i.e., Random Forest and SVM.

TABLE III: Average F1-Score for each oversampling technique and classifier.

Dataset version	Random Forest	SVM
Original	0.4365±0.0896	0.4325±0.0623
BorderlineSMOTE-1	0.5284±0.1135	0.5956±0.1064
BorderlineSMOTE-2	0.5299±0.1063	0.6002±0.1349
MWMOTE	0.4532±0.0676	0.5276±0.0941
O^{2}PF	**0.5441±0.1238**	0.6088±0.1069
O^{2}PF$_{MI}$	0.5247±0.1258	0.5953±0.1146
O^{2}PF$_{P}$	0.4789±0.1037	0.5510±0.0950
O^{2}PF$_{RI}$	0.4431±0.0778	0.5576±0.1079
O^{2}PF$_{WI}$	0.5306±0.0971	**0.6170±0.1081**
SMOTE	0.5319±0.1111	0.6160±0.1127

For both classifiers, the multiclass O^{2}PF attained the best results among the other oversampling approaches. The SVM model obtained the highest score with the O^{2}PF$_{WI}$ as the underlying oversampling strategy, thus achieving an average F1-Score of 0.6170 ± 0.1081. For the Random Forest classifier, the average F1-Score increased by 10% when the O^{2}PF was applied, while the SVM model improved the prediction ability by 18% considering the best oversampling technique (O^{2}PF$_{WI}$). Compared to MWMOTE, which employs a similar cluster-based analysis of the minority class samples, the O^{2}PF-based approaches attained remarkable results over the original version of the dataset. Moreover, MWMOTE requires more hyperparameters in contrast to all O^{2}PF variants, which require only one parameter, i.e., k_{max} for clustering the samples.

The statistical analysis also presents similarities between the O^{2}PF and the other methods used for comparison purposes. The O^{2}PF was statistically similar to SMOTE and the two versions of the BorderlineSMOTE when considering the results obtained by the Random Forest classifier. The same behavior is also noticeable in the SVM results. Furthermore, since the O^{2}PF variants work similarly and share the same mechanism for the oversampling task, a statistical similarity is also expected in some circumstances. One can notice the statistical similarity among all O^{2}PF results, except for the O^{2}PF$_{RI}$ obtained from the Random Forest classifier, and the statistical similarity between O^{2}PF, O^{2}PF$_{MI}$, and O^{2}PF$_{WI}$ for the SVM.

Figure 3 shows the confusion matrices computed from predictions of the SVM classifier before and after applying the oversampling with the O^{2}PF$_{WI}$. The significant improvement is perceptible in predictions of tree species with few samples. For the sake of comparison, no sample of the *Holocalyx balansae* was correctly predicted in the original version of the dataset. On the other hand, the number of successful predictions increased a lot for the oversampled version with the O^{2}PF$_{WI}$. The same behavior is noticeable for the other minority classes. Despite the improvement in detecting the minority class samples, the increase in errors is noticeable for the species *Cenostigma pluviosum* (CP) and *Tipuana tipu* (TT). Since some tree species are similar to each other, the feature vector of new synthetic samples may induce some errors in the balanced version of the dataset.

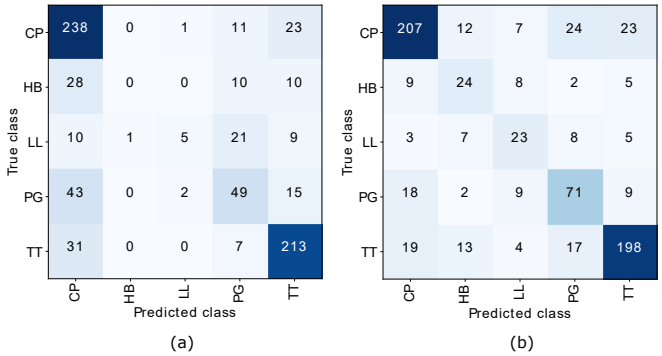

Fig. 3: Confusion matrix computed from predictions of the SVM classifier: a) Original dataset; b) Predictions after applying the oversampling with O^{2}PF$_{WI}$.

Finally, Figure 4 depicts the sample's distributions before and after applying the best-performing oversampling technique (O^{2}PF$_{WI}$).

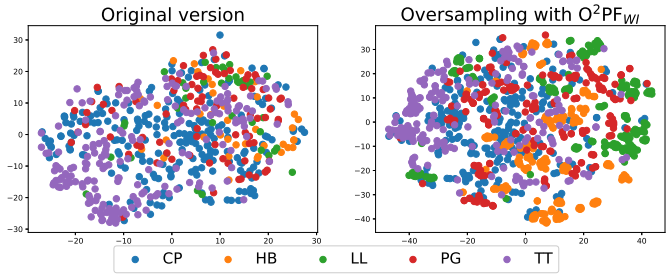

Fig. 4: Distribution of the samples before (left) and after (right) applying the oversampling with the O^{2}PF$_{WI}$. Projection performed using t-SNE.

The data distribution reveals regions with well-defined sparse clusters after using the O^{2}PF$_{WI}$ as the underlying oversampling approach. The oversampled distribution follows a similar behavior as presented by the original version of the dataset, where the data clusters also show a sparse arrangement for each class of tree species. Notice that the new synthetic samples are created for each cluster yielded from each minority class illustrated in the original version of the dataset. We used the t-Distributed Stochastic Neighbor Embedding (t-SNE) [33] method to project the sample's feature vector into a two-dimensional space.

VI. Conclusions and Future Works

This paper presented a new multiclass oversampling approach based on the Optimum-Path Forest framework for tree species classification. The method relies on the normal distribution of minority class clusters generated by the unsupervised

Optimum-Path Forest algorithm. The previous method, named O^2PF, was extended to cope with datasets in the multiclass domain through a simple iterative process that oversamples the minority class samples of each generated cluster. Similar and even superior results confirmed the effectiveness of the multiclass O^2PF against three state-of-the-art methods in the tree species classification domain. Future studies will be conducted to extend the proposed approach to over- and undersampling tasks in multiclass datasets of general purpose.

ACKNOWLEDGMENTS

The authors are grateful to FAPESP grants #2014/12236-1 and #2019/18287-0, CNPq grant 308529/2021-9, and Engineering and Physical Sciences Research Council (EPSRC) grant EP/T021063/1.

REFERENCES

[1] H. C. de Lima Araújo, F. S. Martins, T. T. P. Cortese, and G. M. Locosselli, "Artificial intelligence in urban forestry—A systematic review," *Urban Forestry & Urban Greening*, vol. 66, p. 127410, 2021.

[2] S. Beery, G. Wu, T. Edwards, F. Pavetic, B. Majewski, S. Mukherjee, S. Chan, J. Morgan, V. Rathod, and J. Huang, "The Auto Arborist Dataset: A Large-Scale Benchmark for Multiview Urban Forest Monitoring Under Domain Shift," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 294–21 307.

[3] P. Torres, M. Rodes-Blanco, A. Viana-Soto, H. Nieto, and M. García, "The Role of Remote Sensing for the Assessment and Monitoring of Forest Health: A Systematic Evidence Synthesis," *Forests*, vol. 12, no. 8, p. 1134, 2021.

[4] A. R. Shahtahmassebi, C. Li, Y. Fan, Y. Wu, M. Gan, K. Wang, A. Malik, G. A. Blackburn *et al.*, "Remote sensing of urban green spaces: A review," *Urban Forestry & Urban Greening*, vol. 57, p. 126946, 2021.

[5] S. S. Hanapi, S. Shukor, and J. Johari, "A review on remote sensing-based method for tree detection and delineation," in *IOP Conference Series: Materials Science and Engineering*, vol. 705, no. 1. IOP Publishing, 2019, p. 012024.

[6] M. Michałowska and J. Rapiński, "A review of tree species classification based on airborne LiDAR data and applied classifiers," *Remote Sensing*, vol. 13, no. 3, p. 353, 2021.

[7] C. Zhang, K. Xia, H. Feng, Y. Yang, and X. Du, "Tree species classification using deep learning and RGB optical images obtained by an unmanned aerial vehicle," *Journal of Forestry Research*, vol. 32, no. 5, pp. 1879–1888, 2021.

[8] M. Wang, M. S. Wong, and S. Abbas, "Tropical Species Classification with Structural Traits Using Handheld Laser Scanning Data," *Remote Sensing*, vol. 14, no. 8, p. 1948, 2022.

[9] M. Åkerblom and P. Kaitaniemi, "Terrestrial laser scanning: a new standard of forest measuring and modelling?" *Annals of Botany*, vol. 128, no. 6, pp. 653–662, 2021.

[10] L. Terryn, K. Calders, M. Disney, N. Origo, Y. Malhi, G. Newnham, P. Raumonen, H. Verbeeck *et al.*, "Tree species classification using structural features derived from terrestrial laser scanning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 168, pp. 170–181, 2020.

[11] J. Li, Y. Wu, S. Fong, A. J. Tallón-Ballesteros, X.-s. Yang, S. Mohammed, and F. Wu, "A binary PSO-based ensemble under-sampling model for rebalancing imbalanced training data," *The Journal of Supercomputing*, vol. 78, no. 5, pp. 7428–7463, 2022.

[12] A. Guzmán-Ponce, J. S. Sánchez, R. M. Valdovinos, and J. R. Marcial-Romero, "DBIG-US: A two-stage under-sampling algorithm to face the class imbalance problem," *Expert Systems with Applications*, vol. 168, p. 114301, 2021.

[13] J. Engelmann and S. Lessmann, "Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning," *Expert Systems with Applications*, vol. 174, p. 114582, 2021.

[14] L. A. Souza, L. A. Passos, R. Mendel, A. Ebigbo, A. Probst, H. Messmann, C. Palm, and J. P. Papa, "Fine-tuning generative adversarial networks using metaheuristics," in *Bildverarbeitung für die Medizin 2021*. Springer, 2021, pp. 205–210.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[16] L. Camacho, G. Douzas, and F. Bacao, "Geometric SMOTE for regression," *Expert Systems with Applications*, p. 116387, 2022.

[17] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 405–425, 2012.

[18] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, "A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data," *Information Sciences*, vol. 572, pp. 574–589, 2021.

[19] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Information Sciences*, vol. 465, pp. 1–20, 2018.

[20] L. A. Passos, D. S. Jodas, L. C. F. Ribeiro, T. Moreira, and J. P. Papa, "O^2PF: Oversampling via optimum-path forest for breast cancer detection," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 2020, pp. 498–503.

[21] J. P. Papa, A. X. Falcão, and C. T. N. Suzuki, "Supervised pattern classification based on optimum-path forest," *International Journal of Imaging Systems and Technology*, vol. 19, no. 2, pp. 120–131, 2009.

[22] L. M. Rocha, F. A. M. Cappabianco, and A. X. Falcão, "Data clustering as an optimum-path forest problem with applications in image analysis," *International Journal of Imaging Systems and Technology*, vol. 19, no. 2, pp. 50–68, 2009.

[23] D. S. Jodas, M. Roder, R. Pires, M. C. S. Santana, L. A. de Souza Jr, and L. A. Passos, "Detecting atherosclerotic plaque calcifications of the carotid artery through optimum-path forest," in *Optimum-Path Forest*. Elsevier, 2022, pp. 137–154.

[24] M. A. Bertoni, G. H. d. Rosa, and J. R. Brega, "Optimum-path forest stacking-based ensemble for intrusion detection," *Evolutionary Intelligence*, pp. 1–18, 2021.

[25] L. C. S. Afonso, L. A. Passos, and J. P. Papa, "Enhancing brain storm optimization through optimum-path forest," in *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. Ieee, 2018, pp. 000 183–000 188.

[26] L. A. Passos, D. S. Jodas, L. C. F. Ribeiro, M. Akio, A. N. De Souza, and J. P. Papa, "Handling imbalanced datasets through optimum-path forest," *Knowledge-Based Systems*, vol. 242, p. 108445, 2022.

[27] D. S. Jodas, T. Yojo, S. Brazolin, G. D. N. Velasco, and J. P. Papa, "Detection of Trees on Street-View Images Using a Convolutional Neural Network," *International Journal of Neural Systems*, vol. 32, no. 01, p. 2150042, 2022.

[28] D. S. Jodas, S. Brazolin, T. Yojo, R. A. De Lima, G. D. N. Velasco, A. R. Machado, and J. P. Papa, "A deep learning-based approach for tree trunk segmentation," in *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2021, pp. 370–377.

[29] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," 2017.

[30] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on bayesian optimization," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, 2019.

[31] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *International Conference on Intelligent Computing*. Springer, 2005, pp. 878–887.

[32] G. H. de Rosa and J. P. Papa, "OPFython: A Python implementation for Optimum-Path Forest," *Software Impacts*, p. 100113, 2021.

[33] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.