Adaptive Subspace Sampling for Class Imbalance Processing

Yu-Ting Liu, Nikhil R. Pal, *Fellow, IEEE*, Shang-Lin Wu, Tsung-Yu Hsieh and Chin-Teng Lin, *Fellow, IEEE*

***Abstract*— This paper presents a novel oversampling technique that addresses highly imbalanced data distribution. At present, the imbalanced data that have anomalous class distribution and underrepresented data are difficult to deal with through a variety of conventional machine learning technologies. In order to balance class distributions, an adaptive subspace self-organizing map (ASSOM) that combines the local mapping scheme and globally competitive rule is proposed to artificially generate synthetic samples focusing on minority class samples. The ASSOM is conformed with feature-invariant characteristics, including translation, scaling and rotation, and it retains the independence of basis vectors in each module. Specifically, basis vectors generated via each ASSOM module can avoid generating repeated representative features that offer nothing but heavy computational load. Several experimental results demonstrate that the proposed ASSOM method with supervised learning manner is superior to other existing oversampling techniques.**

I. Introduction

Learning from imbalanced data has attracted a growing attention in the research society in recent years as it is present in a variety of real-world application problems, e.g., medical diagnosis [1], anomaly detection [2], [3], financial fraud detection [4] and biomedical engineering [5], [6]. Under these circumstances, the use of computationally intelligent methods is good potential to play an essential role for solving these problems; however, there are still many new challenges for this research topic.

Specifically, a classification task can regard as an imbalanced problem whenever some types of data distribution significantly dominate the others. In this paper, for simplicity, we focus on the two-class imbalanced classification problem, which is a topic of major interest in research community. The underlying challenge manifests itself in two common forms, relative imbalance and absolute imbalance. Relative imbalance occurs when minority samples are well represented but severely outnumbered by majority samples whereas absolute imbalance arises in datasets in which minority samples are definitely scarce and underrepresented. Either form of imbalance poses a great challenge to conventional classification algorithms as it becomes extremely hard to detect minority class samples. The reason comes from the fact that the algorithm tends to favor the majority class samples or simply omit the minority class samples in the training process and thereby results in a biased classifier. This phenomenon becomes troublesome when the detection of minority class samples is crucially important, such as cancer diagnosis.

Current solutions to the imbalance problem can be divided into two categories: the internal method and the external method. The internal method targets imbalance problem by modifying the underlying classification algorithm. A popular approach in this category is the cost-sensitive learning [7]. It uses a cost-matrix for different types of errors or instance to facilitate the learning direction from an imbalance data set. A higher cost for misclassifying a minority class sample compensate for the scarcity of the minority class. In [8], a cost-sensitive framework for applying support vector machine is proposed. In [9], Zhou and Liu investigated the applicability of cost-sensitive neural networks on imbalance classification problem. By contrast, the external method aims at dealing with imbalance problem by manipulating the input data to form a more balance data set. The external method can further be divided into undersampling and oversampling. Undersampling methods compensate for the imbalance problem by reducing the instances of the majority class. In [10], a cluster-based undersampling approach is proposed. In contrast to undersampling methods that remove majority class sample, the oversampling methods balance the data set by generating synthetic samples for minority classes. The synthetic minority over-sampling technique (SMOTE) [11] algorithm generates an arbitrary number of synthetic minority samples to eliminate the classifier learning bias. A collection of extension works based on the SMOTE algorithm has been proposed to deal with imbalance classification problem, e.g., the Borderline-SMOTE [12], SMOTE-Boost [13], MWMOTE [14], ADASYN [15]. In [16], an oversampling method based on the combination of multivariate Gaussian distribution and interpolation-based algorithm is developed. In this paper, we proposed an adaptive subspace self-organizing map (ASSOM) oversampling method to address the imbalance problem. The ASSOM holds the feature-invariant characteristic, including translation, scaling and rotation. By assuring independence of basis vectors of each module, we can generate representative synthetic samples for the minority class.

II. Adaptive-Subspace Self-Organizing Map

The ideal of using subspaces, which is a subset of the largest principal components, for data generation, is an

Y. T. Liu is with the Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia (email: yu-ting.liu@uts.edu.au).

N. R. Pal is with the Electronics and Communication Sciences Unit, Indian Statistical Institute, Calcutta, India (email: nikh-il@isical.ac.in).

S. L. Wu and T. Y. Hsieh are with the Institute of Electrical Control Engineering, National Chiao Tung University, Hsinchu, Taiwan (email: slwu19870511@gmail.com; aaron.eecs98@g2.nctu.edu.tw).

Chin-Teng Lin is with the Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia; the Brain Research Center and the Lifetime Chair Professor of Electrical Engineering and Computer Science, National Chiao Tung University, Hsinchu, Taiwan; the International Faculty of Center for Advanced Neurological Engineering University of California, San Diego, USA; Honorary Professorship of University of Nottingham, England, UK (email: chin-teng.lin@uts.edu.au).

emerging technology. Since the eigenvectors of the input correlation matrix are called as the principal components, which composed by the corresponding linear subspaces. As shown in Fig. 1, the proposed ASSOM model, which is extended by the concept of self-organizing map (SOM) that is used to model neural functions, has attracted much attention on our insight. Therefore, in this section, we describe the algorithm of ASSOM in the presence of structure and learning scheme.

An invariant feature of the input vector $\vec{X}$ represents the signal subspaces. A linear subspace L of dimensionality H is in general defined as given the linearly independent basis vectors $b_1, \ldots b_H$, and the reconstructed signal is obtained as shown in Eq. (1); however, there exist infinitely many equivalent combination of the b_h, which is not unique, for the same L. In the parameter learning phase, this study utilizes a gradient descent (GD) algorithm to achieve updating fashion. The detailed functions of each layer are described below.

A. ASSOM Structure

The inputs in Layer 1 are crisp values. Only the current states $\vec{X} = (\mathrm{x}_1, \cdots, \mathrm{x}_\mathrm{n})$ are fed as inputs to this layer.

Each node in layer 2 can be represented as a linear-subspace neural unit. Each node is a linearly independent basis vector. The output function of layer 2 is written as:

$$\hat{x} = \left(\sum_{h=1}^{H} b_h b_h^T \right) x = \sum_{h=1}^{H} \left(b_h^T x \right) b_h \tag{1}$$

where b_h is denoted by orthonormal form and H is denoted by the number of hidden nodes. Here, a set of equivalent orthonormal basis vectors for L can be computed by the familiar Gram-Schmidt process. The reconstructed signal relies on orthonormal basis; in other words, reconstructed signal $\hat{x}$ belongs to L is the orthogonal projection of x onto L.

We expect that the reconstructed signal is approximately similar to the original signal; thus, the criterion using Euclidean distance as $\|\tilde{x}\| = \|x - \hat{x}\|$ is presented to determine whether they are similar or even the same. Finally, a projection operator matrix P is defined as Eq. (2) and its property holds $P^2 = P$ and $P^\mathrm{T} = P$.

$$P = \sum_{h=1}^{H} b_h b_h^T \tag{2}$$

where $\hat{x} = Px$ and $\tilde{x} = (I - P)x$, in which I represents an identity matrix.

B. Learning Scheme

Due to inherit the learning mechanism from an SOM, an ASSOM also possesses the abilities of competitive learning for parameter learning, which are vital contributions leading the effectiveness and robustness of system. Firstly, we would like to describe the procedure of competitive learning. The different modules are made to compete on the input signal subspaces to find the minimum distance as winner is an important information represented as a given signal subspace best wins, and consequently, the updated weight vectors in each module followed by the representative winner. As the modules in the neighborhood of the winner are adapted to represent the input better, the neighboring modules gradually

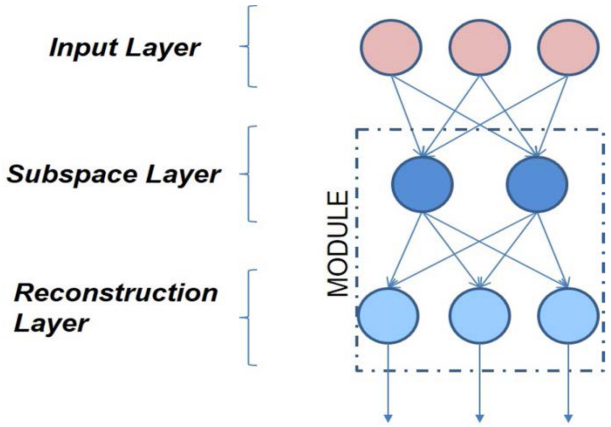

Figure 1. Architecture of ASSOM.

approximate the winner of inputs. The representative winner can be defined by the expression:

$$c = \arg\max_i \left\{ \sum_{t \in S} \left\| \hat{x}^i(t) \right\|^2 \right\} \tag{3}$$

or equivalently,

$$c = \arg\min_i \left\{ \sum_{t \in S} \left\| \tilde{x}^i(t) \right\|^2 \right\} \tag{4}$$

where i and S are denoted by the number of modules and the total number of input samples, respectively.

After obtaining the winning module via a competitive learning, then free parameters of other modules must be adjusted dependently by the factor in terms of distance between their input subspace and the subspace of the winning module to effectively achieve the phase of learning. Therefore, we define the objective to minimize the error function. The error function, which is considered two factors corresponding to the neighborhood factors $\mathrm{h}_\mathrm{c}^\mathrm{i}$ as follows:

$$h_c^i = \exp\left(-\frac{(c-i)^2}{2\sigma^2} \right) \tag{5}$$

The values of the projection error $\tilde{\mathrm{x}}^\mathrm{i}$ is

$$E = \sum_i h_c^i \sum_{t \in S} \left\| \tilde{x}^i(t) \right\|^2 \tag{6}$$

Consequently, this GD algorithm is performed for each piece of incoming datum. By using the GD algorithm for the updated basis vectors of each module, we have

$$b_h^i(t+1) = b_h^i(t) - \eta \frac{\partial E}{\partial b_h^i(t)} \tag{7}$$

where the factor η is a learning rate and the derivation is computed as:

$$\frac{\partial E}{\partial b_h^i(t)} = -2h_c^i \sum_{t \in S} \left(x(t)x(t)^T \right) b_h^i \tag{8}$$

Based on the Eqs. (7) and (8), the basis vectors are updated as follows:

$$b_h^i(t+1) = (I + \eta h_c^i XX^T) b_h^i(t) \tag{9}$$

In the case of rotation operation, the learning rate η should be such that it guarantees a monotonically increasing function of $\|\tilde{x}^i\|$ or a monotonically decreasing function of $\|\hat{x}^i\|$. For the monotonic correction, we must be proportional to $x(t)^T b_h^i$, and thus, one of the simplest ways is to divide the learning rate η by the crisp value $\|\hat{x}^i\|/\|x\|$. Let us note the learning rate as λ, and Eq. (9) then is computed as:

$$b_h^i(t+1) = \left[I + \bar{\lambda}^i \frac{x(t)x(t)^T}{\|\hat{x}^i(t)\|\|x(t)\|} \right] b_h^i(t) \tag{10}$$

where $\bar{\lambda}^i = \lambda h_c^i$.

During the learning process, we set the magnitude of small components of the basis vectors b_h^i to zero for reducing those degrees of freedom; thus, the b_h^i is forced to approximate the dominant frequency components. If we denote the basis vectors $b_h^i = [b_{h1}^i, \cdots, b_{hN}^i]^T$, then the corrected values $\bar{b}_h^i$ by a dissipation effect can be described by

$$\bar{b}_{hj}^i = \mathrm{sgn}(b_{hj}^i)\max(0, |b_h^i| - \varepsilon) \tag{11}$$

where ε is a small fraction of magnitude can be expressed as the following equation.

$$\varepsilon = \alpha |b_h^i(t) - b_h^i(t-1)| \tag{12}$$

where α is a small constant. Based on the dissipation effect, it must be applied after the process of the GD algorithm and prior to normalization. Finally, the learning steps of an ASSOM are concluded as follows:

Once we receive each piece of training data, the procedure will be divided into the following steps.

Step 1: Find the winning module by Eq. (3) or Eq. (4).
Step 2: Update the basis vectors of each module via a gradient descent algorithm.
Step 3: Orthonormalize the basis vectors of each module via the Gram-Schmidt process.

B. Supervised Learning

To effectively display the performance of using different oversampling methods, we apply artificial neural networks (ANNs) during the supervised learning in recognition tasks. ANNs are one of the nonparametric learning methods, and its mathematical model is motivated in accordance with biological neural networks, which imitates the structure and behavior of biological neurons. Specifically, ANNs can be used for solving a learning problem even if there are lacks of any mathematical models from the problems. In addition, ANNs have been successfully used in solving function approximation, pattern recognition, classification or signal and image processing. In this study, we use feedforward ANNs for verifying the improved performance after using different oversampling methods.

III. EXPERIMENTAL RESULT

Five benchmark datasets from UCI machine learning repository [17] and KEEL dataset [18] are employed to tentative the proposed method compared with other existing oversampling or synthetic data generation technologies. In order to significantly reveal the advantage of the proposed method, four assessment metrics, including recall, precision, G-mean, and F-score, are achieved such this intent. Finally, the results demonstrate that we need to take the oversampling techniques into account to avoid the classification of minority party will be dominated by majority party. Specifically, the information contributions from minority party are more important than those from majority party.

This section presents the performance of the ASSOM and compares it with other state-of-the-art methods. The proposed ASSOM in this paper has been successfully validated on nine real-world imbalanced problems from the UCI machine learning repository [17] and the KEEL dataset repository [18] , including Abalone, Breast cancer, Ecoli, Phoneme and Glass,. These sets are chosen in such a way that they have different characteristic of samples, features, classes, and imbalanced ratios. Some of these datasets possesses samples of more than two classes. For simplicity, these datasets are transformed in to a two-class problem in this study. Table I describes the relevant items associated with data attributes and properties. As shown in Table I, there exist highly imbalanced ratios in the presence of the problems of two categories.

Extensive experiments using the ANN as the supervised learning method demonstrate the performance of each dataset on the classification task after employing different

TABLE I. GENERAL INFORMATION OF THE IMBALANCED DATA SETS

Data set name	# of total examples	# of attributes	Minority class	Majority class	# of minority examples	# of majority examples
Abalone	731	7	Class of '18'	Class of '9'	42	689
Breast cancer	683	9	Class of 'malignant'	Class of 'benign'	239	444
Ecoli	336	7	Class of 'im'	All other classes	77	259
Phoneme	5404	5	Class of '1'	Class of '0'	1586	3818
Glass	214	9	Class of '5,6,7'	All other classes	51	163

oversampling approaches. The proposed method is evaluated by the before-and after test to show the improvement compared to the classifiers constructed based on primitive datasets, which these datasets do not be oversampled. After the before-and after test, the ASSOM is further compared the state-of-the-art oversampling approaches, including SMOTE [11] and ADASYN, to show the improvement realized by the proposed method.

For each comparative model in the validation process, 70% of the data are randomly selected to build the training data set, while the remaining serves as test data. To maintain the imbalanced ratio in each dataset, the selection of majority and minority samples are processed from the original dataset, respectively. Further, to prevent the bias of initial states of parameters during the supervised learning procedure, the classification task has been conducted 50 times to evaluate each comparative classifier. This overall process of validation is repeated 5 times; hence, the average of the total 250 runs is compared against other methods.

The validation results with different oversampling approaches on the five datasets are shown in Table II. The best performance in Table II is shown in bold face. The results show that our proposed ASSOM achieves top three performing methods in each assessment matrix. The ASSOM exhibits better performance than SMOTE and ADASYN for most of the real-world problems.

To better show the improvement of the proposed ASSOM, all the comparative approaches are ranked based on the result of each assessment metric. Under each assessment metric, the oversampling algorithm with the best performance is scored 4 points, and the worst one is scored 1 point. Consequently, we compute the average rank of the four assessments metrics across the nine datasets to quantify the relative performance. By further averaging these four assessment metrics, an overall assessment matrix is integrated to evaluate these comparative approaches. The best performance which possesses the highest points are shown in the last row of Table II. The average overall rank of the ASSOM is 3.55, which is higher than any of the other state-of-the-art approaches. These experimental results suggest that our proposed ASSOM model can bring a significant improvement in performance for the imbalance correction.

IV. Conclusion

In this paper we proposed a promising and powerful method, ASSOM, which can effectively evolve useful samples using invariant features associated with rotation, translation and scaling. To solve the imbalanced issues in the recognition problems, synthetic data are intuitively generated into the minority class to reach the amount of majority samples; therefore, classifiers via such an oversample technology strategy is able to obtain a superior performance compared with those ones that are train with imbalanced samples.

The principal contributions of ASSOM contain twofold. One is the learning ability, and the other one is the use of the subspace concept. The learning procedure of an ASSOM is extended from SOM. As a result, the distinguished abilities of ASSOM yet include competitive learning and adaptive

TABLE II. Average Performance Comparison for Different Oversampling Methods

Dataset	Measure	Original	SMOTE	ADASYN	ASSOM
Abalone	Recall	0.401±0.156	**0.765±0.125**	0.511±0.112	0.622±0.093
	Precision	0.414±0.132	0.355±0.049	0.345±0.085	**0.446±0.094**
	F value	0.394±0.123	0.483±0.061	0.407±0.086	**0.513±0.076**
	G mean	0.606±0.132	**0.832±0.068**	0.687±0.078	0.766±0.057
Breast cancer	Recall	0.862±0.065	0.94±0.036	0.902±0.057	**0.958±0.025**
	Precision	0.937±0.031	0.934±0.022	0.936±0.025	**0.947±0.027**
	F value	0.896±0.039	0.937±0.02	0.918±0.031	**0.952±0.021**
	G mean	0.913±0.035	0.952±0.018	0.933±0.029	**0.964±0.016**
Ecoli	Recall	0.714±0.094	0.864±0.072	0.734±0.089	**0.887±0.074**
	Precision	0.645±0.089	0.664±0.082	0.655±0.075	**0.681±0.066**
	F value	0.674±0.073	0.746±0.052	0.689±0.063	**0.766±0.044**
	G mean	0.791±0.056	0.863±0.034	0.803±0.049	**0.879±0.034**
Glass	Recall	0.817±0.111	0.863±0.068	0.79±0.104	**0.88±0.108**
	Precision	0.8±0.102	0.842±0.078	**0.857±0.089**	0.836±0.096
	F value	0.8±0.067	0.849±0.054	0.817±0.074	**0.852±0.079**
	G mean	0.87±0.056	0.903±0.038	0.867±0.059	**0.908±0.074**
Phoneme	Recall	0.717±0.023	0.869±0.017	**0.901±0.016**	0.837±0.027
	Precision	**0.743±0.015**	0.648±0.017	0.604±0.014	0.662±0.018
	F value	0.73±0.014	**0.742±0.013**	0.723±0.012	0.739±0.012
	G mean	0.802±0.012	**0.836±0.01**	0.824±0.01	0.829±0.01
Average	Recall	0.7022	**0.8602**	0.7676	0.8368
	Precision	0.7078	0.6886	0.6794	0.7144
	F value	0.6988	0.7514	0.7108	0.764
	G mean	0.7964	**0.8772**	0.8228	0.8692
Average Rank	Recall	1.2	3.2	2.2	**3.4**
	Precision	2.4	2.2	2	**3.4**
	F value	1.2	3.2	1.8	**3.8**
	G mean	1.2	3.4	1.8	**3.6**
Average overall rank		1.5	3	1.95	**3.55**

learning strategy to effectively adjust all free parameters. The competitive learning is able to locate the optimal model, and the updated weights followed by the winner of which is with the smallest distance to the signal subspace of input domain. Subsequently, the learning procedure of ASSOM mainly adjusts the subspace of the winning module in order to make free weights of each module for approaching the raw signal in the input subspace.

In addition, unlike the comparative methods that mostly exploit the notion of KNN to artificially generate useful samples in this study, the ASSOM using subspace concept is the first proposed to substitute conventional KNN evolved approaches. Experimental results demonstrated the proposed ASSOM with the learning mechanism and the use of subspace concept is much more effective and robust and outperforms its rivals.

ACKNOWLEDGMENT

This work was supported in part by the UST-UCSD International Center of Excellence in Advanced Bio-engineering sponsored by the Taiwan Ministry of Science and Technology under Grant Number partially by MOST 104-2627-E-009-001 and MOST 105-2221-E-009-191, and partially by the Aiming for the Top University Plan of National Chiao Tung University and the Ministry of Education, Taiwan under Contract 105W963. Research was also sponsored in part by the Army Research Laboratory and performed under Cooperative Agreement Number W911NF-10-2-0022. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the US Government. The US Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation herein.

The authors would like to thank Prof. Jyh-Yeong, Chang and all members at Brain Research Center, National Chiao Tung University, Taiwan.

REFERENCES

[1] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Networks*, vol. 21, no. 2–3, pp. 427–436, 2008.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, pp. 1–58, 2009.

[3] S. C. Tan, J. Watada, Z. Ibrahim, and M. Khalid, "Evolutionary Fuzzy ARTMAP Neural Networks for Classification of Semiconductor Defects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 933–950, 2015.

[4] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Decis. Support Syst.*, vol. 50, no. 3, pp. 602–613, 2011.

[5] H. Yu and J. Ni, "An Improved Ensemble Learning Method for Classifying High-dimensional and Imbalanced Biomedicine Data," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, pp. 1–1, 2014.

[6] P. Yang, P. D. Yoo, J. Fernando, B. B. Zhou, Z. Zhang, and A. Y. Zomaya, "Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications," *IEEE Trans. Cybern.*, vol. 44, no. 3, pp. 445–455, 2014.

[7] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[8] R. Batuwita and V. Palade, "FSVM-CIL: Fuzzy support vector machines for class imbalance learning," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 3, pp. 558–571, 2010.

[9] Z. H. Zhou and X. Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 63–77, 2006.

[10] S. J. Yen and Y. S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5718–5727, 2009.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[12] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *Adv. Intell. Comput.*, vol. 17, no. 12, pp. 878–887, 2005.

[13] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," *Knowl. Discov. Databases (PKDD 2003)*, pp. 107–119, 2003.

[14] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE - Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, pp. 405–425, 2014.

[15] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," *IEEE Int. Jt. Conf. Neural Networks (IJCNN 2008)*, pp. 1322–1328, 2008.

[16] H. Cao, X. L. Li, D. Y. K. Woon, and S. K. Ng, "Integrated Oversampling for Imbalanced Time Series Classification," *IEEE Trans. Knowl. Data Eng.*, vol. 25, pp. 2809–2822, 2013.

[17] M. Lichman, "UCI Machine Learning Repository." 2013.

[18] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Mult. Log. Soft Comput.*, vol. 17, no. 2–3, pp. 255–287, 2011.