SCALA: Scaling algorithm for multi-class imbalanced classification

A novel algorithm specifically designed for multi-class multiple minority imbalanced data problems.

Ala O. Barzinji*
School of Computing & Mathematical Sciences, University of Greenwich, London, UK

Jixin Ma
School of Computing & Mathematical Sciences, University of Greenwich, London, UK

Chaoying Ma
School of Computing & Mathematical Sciences, University of Greenwich, London, UK

ABSTRACT

Most of the existing techniques for solving data imbalance problems are geared towards binary classification problems, hence a novel strategy capable of natively handling multi-class classification problems is required. Existing implementations mainly employ a one-versus-rest approach to support multi-class problems and this generalisation hinders its effectiveness in datasets with multiple minority classes. On the contrary, a one-versus-one approach avoids such generalisation and provides finer control over the balancing strategy. In this paper, we propose a novel SCALA algorithm capable of handling imbalanced data with multiple minority class labels with a multi-class output. We introduce a user-defined set of scaling factors which are then integrated with a one-versus-one balancing strategy. Our results show that SCALA demonstrated a significant improvement compared to ADASYN and SMOTE in model performance metrics used to validate balancing techniques. SCALA can balance these datasets without allowing minority classes to overshadow other minority classes. This preserves the information needed by the training algorithm to distinguish between the classes to a high precision.

CCS CONCEPTS

• **Computing methodologies** → Machine learning; Machine learning algorithms; Machine learning; Learning paradigms; Supervised learning; Supervised learning by classification; Machine learning; Machine learning approaches; Classification and regression trees.

KEYWORDS

Imbalanced data, imbalance learning, multiple minority classes

ACM Reference Format:

Ala O. Barzinji, Jixin Ma, and Chaoying Ma. 2023. SCALA: Scaling algorithm for multi-class imbalanced classification: A novel algorithm specifically designed for multi-class multiple minority imbalanced data problems.. In *2023 8th International Conference on Machine Learning Technologies (ICMLT 2023), March 10–12, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3589883.3589893

*To whom correspondence should be addressed at: School of Computing & Mathematical Sciences, University of Greenwich, London SE10 9LS, UK. Tel: +44 (0) 208 331 8539; Email: A.O.Barzinji@greenwich.ac.uk

ICMLT 2023, March 10–12, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9832-9/23/03…$15.00
https://doi.org/10.1145/3589883.3589893

1 INTRODUCTION

In machine learning, classification refers to the power of prediction. The imbalanced dataset problem has been burgeoning in recent years. It is an increasing challenge in both industry and academia [1]. With wide applications of machine learning in the real world, imbalanced data has become a leading concern. Many datasets for real-life scenarios are likely to suffer from a balancing problem. Hence, it is crucial for the machine learning practitioner to be aware of this problem and to possess skills to sufficiently handle this type of data [2]. The fields of study affected by imbalanced datasets range from test classification, bioinformatics, speech recognition, and telecommunications [3, 4], anomaly detection [5], fault diagnosis, and spam filtering [6, 7]. In disease prediction and medical diagnosis, imbalanced classes are almost always present and are a major risk to the suitability of the model. This can be attributed to most available health datasets suffering from unlabelled instances, with labelled datasets being very expensive and difficult to obtain [8]. Furthermore, in the medical domain, data are collected over a time period and evolve from simple sets to more refined categories, which necessitates a multi-class output instead of binary [9]. Class imbalance occurs whenever the labels of the dataset are distributed in such a way that one or more class labels dominate the instance space [10]. Classification models suffer from some degree of imbalance problems if the classes of interest are very scarce (minority classes) while other classes dominate the dataset (majority classes). The phenomenon of imbalanced data can also be defined as the skewness of the class label distribution. The extent of the imbalance is defined as the imbalance ratio. The representation of the majority class (or the normal case) and the minority class (or the abnormal case) imposes a major error in the classification model performance. This leads to minority class examples being falsely misclassified as the majority class.

An intuitive example is classification of body-mass index (BMI) in children predicting four distinct classes: Underweight, Healthy, Overweight, and Obese. Most examples will be the Healthy class, and Obese and Overweight will be the minority classes, with Underweight somewhere in between. If there are only four obese cases in one thousand records, then the machine learning model will lean towards predicting the Healthy class for all inputs, and the model will be falsely considered very accurate. Thus, the model will most likely learn to predict the majority classes (Healthy) no matter what the input data is. The model is effectively unsuitable, irrespective of the high accuracy. To combat this problem, the dataset must be balanced to similar amounts of minority and majority examples. The interpretation of misclassification errors may differ across the classes. It depends on the extent of the imbalance, which makes most machine learning algorithms perform poorly, and the complexity of the problem domain where the examples were drawn from. For example, misclassifying an instance from the majority class as an instance from the minority class is (a false-positive error) and is undesired in most cases. However, it is less critical than classifying an instance from the minority class as an instance of the majority class (a false-negative error). This is known as the unequal cost of misclassification and is another challenge of imbalance problems. These challenges, skewness of data and unequal cost of misclassification errors, exacerbate the imbalanced data problem. In addition, there are other concerns such as data size, data distribution, and anomaly detection which are factors for worsening the imbalanced data [11]. In contrast, some studies suggest that an imbalance ratio is not the only factor for deficient performance in learning from imbalanced data [12]. The minority class is under-represented and hence it is much harder to learn than the majority class. Thus, misclassifying the minority class examples are more costly. This is because there is an expectation for the data to be balanced or have equal misclassification costs. The aim of the classifier, in this case, is to maximise overall model accuracy. This often leads to the classifier predicting examples as the majority classes and under-representing the minority classes. In real world cases, this leads to up to 100% accuracy of the majority classes and a very low accuracy of the minority classes, between 0% - 10% [13]. Class imbalance data can negatively impact the performance of machine learning algorithms that train on the data. In scenarios with rare cases, events occur infrequently. Because of this, the machine learning models that we build to identify these rare cases will perform badly. Learning from imbalanced data will produce biased machine learning models towards the majority class. Standard classification algorithms do not perform well on imbalance datasets especially with small sample sizes [14]. They should not be used as they may lead to critical mistakes. Instead, a classifier with adjustable output threshold needs to be used [15].

Most widely used balancing algorithms, such as SMOTE and ADASYN, have been shown to balance the classes in the dataset such that the trained machine learning model predicts all classes to a high precision. However, these algorithms were designed primarily for binary classification problems. The algorithms themselves do not natively support multi-class inputs, instead code-based implementations (such as those found in scikit-learn) provide support for multi-class inputs using a one-versus-rest approach. One-versus-rest effectively collapses the scenario into a binary classification problem by retaining one class and grouping all other classes into a second class. The original algorithm is then computed on the grouped data. This grouping leads to a loss of data and loss of fine control needed to differentiate between the classes.

2 RELATED WORKS

There is a proliferation of studies available on class imbalance. Various methods and techniques have been developed to address imbalanced datasets. In those methods and techniques, data engineering has been employed in attempts to address the issue. However, recent studies show that data engineering has no impact on classification performance compared to the data augmentation method. However, combining various augmentation methods will result in deterioration of the classifier performance in imbalanced data problems [16] Class imbalance can be addressed at the data level or the algorithm level. Ensemble-based methods leveraging both data- and algorithm-level techniques are also available.

Firstly, data level approaches have been proposed and lean towards sampling methods. These aim to correct the distribution of the training data by some mechanisms, to obtain a relatively balanced class distribution. The mechanisms include under-sampling, over-sampling, and some hybrid methods. Under-sampling and hybrid under-sampling methods work by reducing the number of examples from the majority classes [17–19]. However, these methods tend to be avoided as it removes potentially important class examples for the majority class. There is no evidence that under-sampling techniques address imbalance data problems to a high degree [20]. In contrast, over-sampling methods create new samples for the minority classes. There are two types of over-sampling: random over-sampling, and synthetic over-sampling. In random over-sampling techniques, the existing examples in the minority classes are replicated randomly. Synthetic over-sampling techniques generate new artificial samples from the minority class examples. Several over-sampling techniques have been introduced such as: random over-sampling with replacement, synthetic minority over-sampling technique (SMOTE) [21], Borderline-SMOTE [22], SMOTEBoost [23], AWSMOTE [24], MSMOTE [25], Safe-level SMOTE [26], and Adaptive Synthetic Sampling Technique (ADASYN) [27]. Both under-sampling and over-sampling techniques have their own advantages as well as drawbacks [28]. Another approach is a hybrid approach combining under- and over-sampling techniques [29]. Secondly, algorithm level approaches have been proposed such as cost-sensitive learning [30] and kernel-based methods [31]. Finally, ensembled-based methods have been proposed such as those in the study of He et al. and Devi et al. [20, 32]. The Adaptive Synthetic (ADASYN) algorithm is an over-sampling technique with a focus on harder-to-learn examples [27]. The ADASYN algorithm works by calculating a density distribution for each minority class example, using the ratio of majority neighbours to total neighbours, through the K Nearest Neighbour method. This provides an advantage over previous over-sampling techniques such as SMOTE, Borderline-SMOTE, and SMOTEBoost which apply their over-sampling algorithms uniformly across the data. Therefore, ADASYN can selectively over-sample examples only when needed, reducing the likelihood of overfitting and bias. Due to the nature of ADASYN, outlier minority class examples are not over-sampled, which reduces bias in the model. This approach was built for binary classification problems with one majority class and one minority class. Implementations of the ADASYN algorithm such as the scikit-learn Python library use a one-versus-rest approach to support multi-class and multi-label problems. However, these approaches tend to break down when there are multiple minority classes in one neighbourhood because the current implementations continue to use majority-to-total neighbour ratio when calculating the density distribution.

In contrast, our approach calculates independent density distributions for each neighbouring class label at each minority class example. This is also known as a one-versus-one approach. This allows the algorithm to fully understand the composition of the surrounding examples as opposed to previous implementations which were only aware of majority-to-total neighbours. Our approach also introduces the novel scaling factors as another input parameter to the algorithm. These scaling factors are applied to the independent density distributions to either penalise or promote competition of the minority class example with the other class labels. This enables finer tuning of the algorithm and can be configured such that a minority class example surrounded by neighbours primarily from other minority classes will experience less over-sampling than minority class examples surrounded by majority classes.

3 SCALA: SCALING ALGORITHM

To improve the prediction power of classification algorithms, we propose a novel SCALA algorithm in this paper. Building upon the ADASYN algorithm, our approach improves the prediction power of datasets with multiple minority class labels. The proposed algorithm for solving multi-class classification problems is described as follows:

3.1 The intuition of our novel algorithm

The key elements of the novel approach are implementation of a one versus one approach, and introduction of the scaling factors which is a distinct input parameter.

In the first step, we begin with a similar procedure to ADASYN by calculating the K-nearest neighbours surrounding the minority class example in question. This makes use of the Euclidean distance to determine the closest points. Knowledge of the surrounding points and which labels they have is necessary to proceed with the algorithm. In previous one-versus-rest approaches the exact composition of the neighbouring labels were not needed. Therefore, the algorithm does not need to be aware of the total set of labels within the dataset. However, with the one-versus-one approach each label is put against all other labels in the dataset except itself. For this reason, the algorithm must be able to determine a set containing all labels in the dataset, except the current example's label. In the second step, we describe this procedure. First a set C is determined containing all labels in the dataset. Then, the second set D is defined as set C with the current example's label excluded. This set D is then used to continue with the algorithm. In ADASYN, a ratio is computed dependent on the proportion of majority to total examples. This approach is not compatible with a one-versus-one approach, as the proportion of majority to total examples would always be the same for all classes surrounding a minority class example.

In our approach, to make better use of the data now available to the algorithm, the third step calculates the ratio dependent on the proportion of each label against the total examples and provides a better summary of the labels surrounding the current example. In the fourth step, the ratios are normalised such that the sum of all the ratios is one. This forms a density distribution so that G can be multiplied by each component to form the final number of samples to generate.

The scaling factor parameter is introduced in step 5 of the algorithm. Each normalised ratio is multiplied by its respective scaling factor corresponding to its class label. This skews the ratios to either boost or reduce the values depending on the user defined input. In most cases, the scaling factors will be used to penalise the majority classes and boost the minority classes. By combining the novel scaling factors with the adaptive ratios, you can achieve an elevated level of oversampling only where it is needed.

Algorithm 1 SCALA Algorithm

Input
Training dataset
Scaling factors, F
Procedure
1. Calculate the number of synthetic data examples that need to be generated:
$G = (m_b - m_s) \times \beta \qquad (1)$
Where $\beta \in [0, 1]$ is a parameter used to specify the desired balance level after generation of the synthetic data. m_s is the number of examples in the smallest minority class.
2. Determine C as the set of labels in the dataset Y and, D the difference of C without the class label of x_i as
$D = C \setminus \{ y_i \} \qquad (2)$
3. Calculate the ratio against the other classes divided by total number of neighbours. For each j in D, do:
$r_i^{(j)} = \frac{\Delta_i^{(j)}}{K},\ i = 1, \ \ldots, \ m_s,\ j \in D \qquad (3)$
where $\Delta_i^{(j)}$ is the number of examples in the K nearest neighbours against other classes and (j) represents an iteration element of set D.
4. Normalise these ratios according to the sum of all the ratios, such that:
$\widehat{r}_i^{(j)} = r_i^{(j)} \ / \ \sum_{i=1}^{j} r_i \qquad (4)$
5. Apply the scaling factors to these ratios, such that:
$\widehat{s}_i^{(j)} = \widehat{r}_i^{(j)} \ \times F_j \qquad (5)$
6. Use the ratios to calculate the number of the synthetic samples to generate for this point
$g_i = \ \sum_{j}^{D} (\widehat{s}_i^{(j)} \ \times G) \qquad (6)$
where j is the first class in the set of all classes D.
7. For each minority class example x_i, generate g_i synthetic data examples according to the following steps:
Do the **Loop** from 1 to g_i:
i) Randomly choose one minority data example, x_{zi} from the K nearest neighbours for data x_i
ii) Generate the synthetic data sample:
$s_i = \ x_i + (x_{zi} - \ x_i) \ \times \ \lambda \qquad (7)$
where $(x_{zi} - \ x_i)$ is the difference vector in n dimensional spaces, and λ is a random number $\lambda \ \in [0, \ 1]$.
End **Loop**

3.2 Description of the dataset

To evaluate the effectiveness of our algorithm, we used the Southampton Women's Survey (SWS) dataset which had a highly imbalanced nature. The data consists of 3,157 rows consisting of 215 columns. Not every entry had a complete set of data, meaning some columns are empty. The output column of interest was body-mass index (BMI) which consists of four classes, of which two are deemed minorities. The value counts are as follows: Underweight, with 1,434 records; Healthy, with 588 records; Overweight, with 190 records; and Obese, with 107 records. The dataset poses a challenge consisting of severely imbalanced data with multiple minorities, with a multi-class output model.

4 SIMULATION ANALYSIS AND REASONING FOR CREATING A NEW ALGORITHM

We used the random forest classifier to build the machine learning model that will be trained with the SWS data. We selected random forest due to its robust nature and intrinsic ability to deal with imbalanced data better than other algorithms [33]. This is partly due to bootstrapping, where subsets of the data are distributed randomly among individual trees and the final output is decided upon by voting on the individual tree predictions. The first step in evaluation was to establish a baseline using only random forest without any balancing algorithms. As expected, it could not perform well with the degree of the imbalanced data in the SWS dataset, despite hyper-parameter tuning. For this reason, we looked towards balancing techniques. Due to the low number of records in the minority classes, we could not rely on under-sampling techniques. For this reason, we selected two widely used over-sampling techniques, SMOTE and ADASYN. It was evident that ADASYN outperformed SMOTE, and this can be attributed most likely to its adaptive nature. At this stage, we also tried amending hyper-parameters such as K-neighbours but could not achieve any significant improvements. Therefore, we selected ADASYN as the starting point for our improved algorithm, SCALA. For each experiment, we split the data into training and testing counterparts at a ratio of 30% with stratification. ADASYN and SMOTE parameters remained as default as K_neighbours = 5, and B = 1. For our algorithm hyperparameters, we set K_neighbours = 5 and B = 1. We also used scaling factors as follows: Healthy = 2, Underweight = 6, Obese = 0.1, Overweight = 0.1. These values were attained using trial and error.

4.1 Evaluation Metrics for Multi-class Imbalanced Datasets

Traditionally, the performance of machine learning algorithms is evaluated by overall accuracy against test cases. However, this measure is often insufficient when dealing with imbalanced datasets [34]. Most of the machine learning algorithms perform better when datasets are balanced, hence machine learning algorithms expect balanced data. Overall accuracy, a popular choice of evaluating a classifier's performance, might not be appropriate when we encounter imbalanced classification [35]. In addition to the confusion matrix, we also evaluate our algorithm using different performance metrics such as AUC-ROC [34], precision, recall, F1-score, Kappa [36], and MCC [37, 38], which are proven to be the most suitable evaluation measurements for multi-class imbalance classification.

5 EXPERIMENTAL RESULTS

The important outcome of learning algorithms is to improve the recall, without sacrificing the precision. The best measure for evaluating the learning algorithm towards the minority class is the F-score, which is the integration of both precision and recall [23]. These metrics are all derived through a combination of formulae which use values from the confusion matrix. The confusion matrix of the same data before and after application of SCALA are shown in Table 1 and Table 2 respectively.

The results show that the false positive and false negative occurrences are reduced after application of SCALA.

The area under the curve score was also calculated using the receiver operating characteristic (ROC) function, known as the ROC-AUC score. The ROC-AUC score was found to be 84%, which was equal to the ROC-AUC for ADASYN and SMOTE. Though our model saw increased metrics such as MCC and Kappa as shown in Table 3. In each model, the highest performing algorithm is highlighted for each metric. From these results, we can see that the original ADASYN algorithm attained the highest overall accuracy (OA) score of 0.69. However, as mentioned previously, OA is not a suitable metric when the problem at hand has a multi-class output. Instead, we should look towards Cohen's Kappa coefficient (Kappa) and Mathew's correlation coefficient (MCC). SCALA attained the best performing Kappa (0.43) and MCC (0.44) out of the algorithms, showing that it is more suited for multi-class imbalanced data than ADASYN and SMOTE. The results also show that, apart from three metrics, SCALA achieved the highest precision, recall, and F1-score, especially Obese and Overweight classes which were the minority classes. This means that the novel SCALA algorithm exhibits increased performance for all classes while achieving its advantageous ability of balancing datasets with multiple minority classes. We can confidently conclude that SCALA provides many advantages over the ADASYN algorithm, both in its ability to balance multiple minority imbalance datasets, and its raw performance improvements across the board in the metrics measured. Applying SCALA to an imbalanced dataset yields a predictor capable of distinguishing between all classes to a high precision.

6 CONCLUSION

In this paper, a novel Scaling algorithm (SCALA) was proposed for solving multiple minority imbalanced dataset phenomena. The results show that SCALA can improve the performance of the classifiers for all classes in a multi-class multi-minority scenario. SCALA is an improvement over ADASYN and SMOTE as a new over-sampling approach to handle class imbalance. SCALA was tested on SWS data to predict BMI classes into Healthy, Obese, Overweight and Underweight. SCALA, in tandem with the random forest machine learning method, was found to perform better than random forest with ADASYN or SMOTE. Our algorithm was shown to have greater predictive power for minority classes in imbalanced data, which tend to be the classes of interest. Without compromising AUC-ROC score, we were able to increase the precision, recall, and F1 of the minority classes while preserving good predictive power of majority classes as well. Our model showed an increased Kappa and MCC score to the other algorithms. The introduction of the scaling factors input provides opportunity for fine control over the

Table 1: Confusion matrix for the classes in our multi-class classification, which are Healthy, Obese, Overweight and Underweight, without the application of any balancing techniques.

Confusion matrix		Predicted class			
		Healthy	Obese	Overweight	Underweight
Actual class	Healthy	73	0	0	104
	Obese	3	5	1	23
	Overweight	12	1	1	43
	Underweight	24	3	1	402

Table 2: Confusion matrix for the classes in our multi-class classification, which are Healthy, Obese, Overweight and Underweight, with the application of SCALA.

Confusion matrix		Predicted class			
		Healthy	Obese	Overweight	Underweight
Actual class	Healthy	88	0	10	79
	Obese	1	16	2	13
	Overweight	12	6	15	24
	Underweight	34	9	8	379

Table 3: The performance measures for SCALA and the comparison with ADASYN and SMOTE. Precision, Recall, F-score, KAPPA and MCC

Algorithm	OA	Classes	Precision	Recall	F1	Kappa	ROC-AUC	MCC
SMOTE	0.68	Healthy	0.59	0.49	0.53	0.37	0.84	0.37
		Obese	0.42	0.41	0.41			
		Overweight	0.26	0.14	0.18			
		Underweight	0.75	**0.85**	**0.81**			
ADASYN	**0.69**	Healthy	**0.61**	0.53	0.57	0.39	0.84	0.39
		Obese	0.37	0.34	0.35			
		Overweight	0.31	0.19	0.24			
		Underweight	0.76	0.84	0.80			
SCALA	0.67	Healthy	0.51	**0.76**	**0.61**	**0.43**	0.84	**0.44**
		Obese	**0.44**	**0.47**	**0.45**			
		Overweight	**0.39**	**0.28**	**0.33**			
		Underweight	**0.84**	0.70	0.76			

scaling strategy. Therefore, we can conclude that our algorithm is best suited to improve predictive power of minority classes in severely imbalanced datasets, where those minority classes are the classes of interest. Our suggestion for future work, to improve the algorithm, is to devise a secondary algorithm to efficiently select scaling factors for a particular input dataset and to deploy and refine the algorithm on a wider range of datasets.

ACKNOWLEDGMENTS

The authors thank the women of Southampton and their children who gave their time to take part in the Southampton Women's Survey and the research nurses and other staff who collected and processed the data. All phases of the SWS were approved by Southampton and South-West Hampshire Local Research Ethics Committee, and written consent was obtained from all participants.

REFERENCES

[1] F Provost. 2000. Machine learning from imbalanced data sets 101. Invited paper for the AAAI, Workshop on Imbalanced Data Sets, Menlo Park, CA (2000).
[2] Max Kuhn. 2018. Applied Predictive Modeling 1st ed. 2013, Corr. 2nd.
[3] Maria Rocha Sousa, João Gama, and Elísio Brandão. 2016. A new dynamic modeling framework for credit risk assessment. Expert Syst Appl 45, (March 2016), 341–351. DOI:https://doi.org/10.1016/j.eswa.2015.09.055
[4] Hang Zhang, Weike Liu, Shuo Wang, Jicheng Shan, and Qingbao Liu. 2019. Resample-Based Ensemble Framework for Drifting Imbalanced Data Streams. IEEE Access 7, (2019), 65103–65115. DOI:https://doi.org/10.1109/ACCESS.2019.2914725
[5] Jordi Meseguer, Vicenç Puig, and Teresa Escobet. 2010. Fault Diagnosis Using a Timed Discrete-Event Approach Based on Interval Observers: Application to Sewer Networks. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 40, 5 (September 2010), 900–916. DOI:https://doi.org/10.1109/TSMCA.2010.2052036
[6] Kyosuke Nishida, Shohei Shimada, Satoru Ishikawa, and Koichiro Yamauchi. 2008. Detecting sudden concept drift with knowledge of human behavior. In 2008 IEEE International Conference on Systems, Man and Cybernetics, IEEE, 3261–3267. DOI:https://doi.org/10.1109/ICSMC.2008.4811799

[7] Shuo Wang, Leandro L. Minku, and Xin Yao. 2014. A multi-objective ensemble method for online class imbalance learning. In 2014 International Joint Conference on Neural Networks (IJCNN), IEEE, 3311–3318. DOI:https://doi.org/10.1109/IJCNN.2014.6889545
[8] Yanping Yang and Guangzhi Ma. 2010. Ensemble-based active learning for class imbalance problem. J Biomed Sci Eng 03, 10 (2010), 1022–1029. DOI:https://doi.org/10.4236/jbise.2010.310133
[9] Tengyue Li, Simon Fong, Sabah Mohammed, Jinan Fiaidhi, Steven Guan, and Victor Chang. 2022. Empowering multi-class medical data classification by Group-of-Single-Class-predictors and transfer optimization: Cases of structured dataset by machine learning and radiological images by deep learning. Future Generation Computer Systems 133, (August 2022), 10–22. DOI:https://doi.org/10.1016/j.future.2022.02.022
[10] Haibo He and E.A. Garcia. 2009. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng 21, 9 (September 2009), 1263–1284. DOI:https://doi.org/10.1109/TKDE.2008.239
[11] Jorma Laurikkala. 2001. Improving Identification of Difficult Small Classes by Balancing Class Distribution. . 63–66. DOI:https://doi.org/10.1007/3-540-48229-6_9
[12] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. Learning from Imbalanced Data Sets (1st ed. 2018 ed.). Springer International Publishing.
[13] Miroslav Kubat, Robert C Holte, and Stan Matwin. 1998. Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Mach Learn 30, 2 (1998), 195–215. DOI:https://doi.org/10.1023/A:1007452223027
[14] Yong Hu, Dongfa Guo, Zengwei Fan, Chen Dong, Qiuhong Huang, Shengkai Xie, Guifang Liu, Jing Tan, Boping Li, and Qiwei Xie. 2015. An Improved Algorithm for Imbalanced Data and Small Sample Size Classification. Journal of Data Analysis and Information Processing 03, 03 (2015), 27–33. DOI:https://doi.org/10.4236/jdaip.2015.33004
[15] S Kubat, M. and Matwin. 1997. Addressing the curse of imbalanced training sets: one-sided selection. (1997).
[16] Ahsan bin Tufail, Kalim Ullah, Rehan Ali Khan, Mustafa Shakir, Muhammad Abbas Khan, Inam Ullah, Yong-Kui Ma, and Md. Sadek Ali. 2022. On Improved 3D-CNN-Based Binary and Multiclass Classification of Alzheimer's Disease Using Neuroimaging Modalities and Data Augmentation Methods. J Healthc Eng 2022, (February 2022), 1–14. DOI:https://doi.org/10.1155/2022/1302170
[17] Małgorzata Bach, Aleksandra Werner, and Mateusz Palt. 2019. The Proposal of Undersampling Method for Learning from Imbalanced Datasets. Procedia Comput Sci 159, (2019), 125–134. DOI:https://doi.org/10.1016/j.procs.2019.09.167
[18] Mahmudul Hasan Popel, Khan Md. Hasib, Syed Ahsan Habib, and Faisal Muhammad Shah. 2018. A Hybrid Under-Sampling Method (HUSBoost) to Classify Imbalanced Data. In 2018 21st International Conference of Computer and Information Technology (ICCIT), IEEE, 1–7. DOI:https://doi.org/10.1109/ICCITECHN.2018.8631915
[19] Debashree Devi, Saroj K. Biswas, and Biswajit Purkayastha. 2019. Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique. Conn Sci 31, 2 (April 2019), 105–142. DOI:https://doi.org/10.1080/09540091.2018.1560394
[20] Debashree Devi, Saroj K. Biswas, and Biswajit Purkayastha. 2020. A Review on Solution to Class Imbalance Problem: Undersampling Approaches. In 2020 International Conference on Computational Performance Evaluation (ComPE), IEEE, 626–631. DOI:https://doi.org/10.1109/ComPE49325.2020.9200087
[21] N. v. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, (June 2002), 321–357. DOI:https://doi.org/10.1613/jair.953
[22] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. . 878–887. DOI:https://doi.org/10.1007/11538059_91
[23] Nitesh v. Chawla, Aleksandar Lazarevic, Lawrence O. Hall, and Kevin W. Bowyer. 2003. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. . 107–119. DOI:https://doi.org/10.1007/978-3-540-39804-2_12
[24] Naga Prasanthi Bobbili and Ana-Maria Cretu. 2018. Adaptive Weighting with SMOTE for Learning from Imbalanced Datasets: A Case Study for Traffic Offence Prediction. In 2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), IEEE, 1–6. DOI:https://doi.org/10.1109/CIVEMSA.2018.8439957
[25] Shengguo Hu, Yanfeng Liang, Lintao Ma, and Ying He. 2009. MSMOTE: Improving Classification Performance When Training Data is Imbalanced. In 2009 Second International Workshop on Computer Science and Engineering, IEEE, 13–17. DOI:https://doi.org/10.1109/WCSE.2009.756
[26] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. 2009. Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. . 475–482. DOI:https://doi.org/10.1007/978-3-642-01307-2_43
[27] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, 1322–1328. DOI:https://doi.org/10.1109/IJCNN.2008.4633969
[28] Oded Maimon and Lior Rokach. 2010. Data Mining and Knowledge Discovery Handbook. Springer US, Boston, MA. DOI:https://doi.org/10.1007/978-0-387-09823-4
[29] Enislay Ramentol, Yailé Caballero, Rafael Bello, and Francisco Herrera. 2012. SMOTE-RSB *: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowl Inf Syst 33, 2 (November 2012), 245–265. DOI:https://doi.org/10.1007/s10115-011-0465-6
[30] Mengjun Kang, Yue Liu, Mengqi Wang, Lin Li, and Min Weng. 2022. A random forest classifier with cost-sensitive learning to extract urban landmarks from an imbalanced dataset. International Journal of Geographical Information Science 36, 3 (March 2022), 496–513. DOI:https://doi.org/10.1080/13658816.2021.1977814
[31] Jue Zhang and Li Chen. 2019. Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. Computer Assisted Surgery 24, sup2 (October 2019), 62–72. DOI:https://doi.org/10.1080/24699322.2019.1649074
[32] Haibo He and E.A. Garcia. 2009. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng 21, 9 (September 2009), 1263–1284. DOI:https://doi.org/10.1109/TKDE.2008.239
[33] Chao Chen, Andy Liaw, and Leo Breiman. 2004. Using Random Forest to Learn Imbalanced Data.
[34] F. Provost and T. Fawcett. 1997. Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. In proceedings of the Third international conference on Knowledge discovery and data mining, Menlo park, CS. AAAI Press, 43-48, 1997 (1997).
[35] Nitesh v. Chawla. 2009. Data Mining for Imbalanced Datasets: An Overview. In Data Mining and Knowledge Discovery Handbook. Springer US, Boston, MA, 875–886. DOI:https://doi.org/10.1007/978-0-387-09823-4_45
[36] J R Landis and G G Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33, 1 (March 1977), 159–74. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/843571
[37] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognit 91, (July 2019), 216–231. DOI:https://doi.org/10.1016/j.patcog.2019.02.023
[38] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. 2021. The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. IEEE Access 9, (2021), 78368–78381. DOI:https://doi.org/10.1109/ACCESS.2021.3084050