Journal of Autonomous Intelligence (2024) Volume 7 Issue 3
doi: 10.32629/jai.v7i3.966

REVIEW ARTICLE

An extensive analysis of several methods for classifying unbalanced datasets

Sharaf Alzoubi[1], Khaled Aldiabat[2], Mofleh Al-diabat[3], Laith Abualigah[3,4,5,6,7,8,9,*]

[1] *College of Computing and Informatics, Amman Arab University, Amman 11953, Jordan*
[2] *Department of Management Information Systems, Ajloun National University, Ajloun 26810, Jordan*
[3] *Computer Science Department, Al al-Bayt University, Mafraq 25113, Jordan*
[4] *Department of Electrical and Computer Engineering, Lebanese American University, Byblos 13-5053, Lebanon*
[5] *Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman 19328, Jordan*
[6] *MEU Research Unit, Middle East University, Amman 11831, Jordan*
[7] *Applied Science Research Center, Applied Science Private University, Amman 11931, Jordan*
[8] *School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang 11800, Malaysia*
[9] *School of Engineering and Technology, Sunway University Malaysia, Petaling Jaya 27500, Malaysia*
*** Corresponding author:** Laith Abualigah, aligah.2020@gmail.com

ABSTRACT

In large-scale data applications, handling unbalanced data is a major issue. In order to gather the uneven data at the fastest pace feasible, the imbalanced data categorization system was created. Numerous neural methods have been developed to accurately categorize unbalanced data. However, because of the intricacy of the data, the classification process becomes more challenging due to increased resource utilization, computing costs, and algorithm complexity. As a result, this research has provided specifics on the performances of many classification models in various unbalanced datasets. Ultimately, a performance study was conducted to evaluate each model's categorization performance. For this reason, the precision, specificity, accuracy, and sensitivity have been used to measure the robustness. Each model's advantages and disadvantages are also thoroughly covered. The categorization models then offered future approaches to enhance the unbalanced data based on the drawbacks.

Keywords: imbalanced data; data mining; deep learning; classifiers; over and under sampling; optimization algorithms

ARTICLE INFO

Received: 11 July 2023
Accepted: 12 October 2023
Available online: 5 January 2024

1. Introduction

When one or more classes are underrepresented in the dataset, class imbalance classification is used in data mining and machine learning[1]. Class imbalance is present in many real-world classification tasks, which poses a serious challenge to the data mining community[2]. The main problem with these problems is the skewed distribution, which reduces the efficacy of conventional classification techniques since most learning algorithms presume a training dataset, which makes minority class situations more difficult to predict[3,4]. Several attempts have been made in the last several years to solve the issue of binary unbalanced classes, which only consist of two classes. Nonetheless, a number of domains, such as text classification, human impact recognition, and diagnostics, use multi-class unbalanced classification[5]. Unfortunately, it may not be appropriate to apply the solutions offered for two unequal class problems straight to the number expansion of classes. It could now handle multi-class problems by

using certain algorithms[6]. Thankfully, decomposition techniques have been developed by the academic community to handle the multi-class classification problem[7]. In this stage, the multi-class categorization problems are reduced to two class subtasks, which are much simpler to distinguish. One versus one and one against all are two well-known strategies. Due to the establishment of a synthetic class imbalance, the situation has become one where one entity is pitted against several others[8].

Furthermore, it is advised against resolving difficulties with distorted initial distributions[9]. Recent study indicates that the performance of a multi-class breakdown, which includes one classifier, is significantly enhanced by using a structured curriculum and addressing data-level challenges[10].

The aforementioned methodologies has several captivating uses when dealing with imbalanced datasets. Multi-class imbalanced data classification has been highly valued by the scientific community in recent times[11,12]. Nevertheless, the majority of existing methods for addressing imbalanced classes rely on oversampling approaches. Undersampling is advantageous in binary volatile scenarios, since it may mitigate certain limitations associated with resampling[13]. There is a lack of specific under-sampling methods that include many categories and can deduce their connections[14]. Multi-class imbalanced issues pose significant difficulties in solving them owing to the large number of classes that need to be analyzed and the complex relationships between these classes[15]. Furthermore, the conventional methods designed for binary problems may become impractical or ineffective since they are unable to replicate this complex challenge[16]. Currently, there is a scarcity of specialized multi-class approaches, and further testing is necessary in this domain. Static-Synthetic-Minority The SMOTE methodology employs an M-step resampling method, where M is the class count[17,18]. The resampling approach selects the class with the least size for each cycle and duplicates a few instances from that original dataset class[19]. An established trend in this domain highlights the crucial significance of taking into account the distinct attributes of SMOTE categories. Their training complexity has increased while working with unbalanced oversampling data in a multi-class setting. It is suggested to use a data-driven approach that can be used to any statistical multi-class solution[20].

Current oversampling strategies focus on using data from many classes concurrently while limiting the impact of overlapping and noisy elements[21]. The purpose of prototype selection in a sampling operation is to minimize the classifier's comparison set, hence increasing efficiency and reducing storage requirements. However, in an imbalanced situation, the objective shifts to prioritize the balancing of data throughout the distribution process, as shown by previous studies[22,23]. The support vectors have prioritized generating a significant under-sampled dataset via a search guided by a genetic algorithm[24]. To begin, many randomly under-sampled information subsets are created and then refined to improve the fitness value of the already well-represented dataset[25]. Similarly, in every evolutionary process, the arrangement in which genomes reflect solutions is crucial. Therefore, a comprehensive analysis has been carried out on unbalanced data classification to ascertain the fundamental features and limitations of each model.

2. Unbalanced categorization of the dataset

Unbalance class is one of the hottest study topics in machine learning. Nevertheless, previous and current studies have shown that class duplication negatively impacts classifier performance more than anything else. This paper examines class overlapping in detail and impartially from the standpoint of imbalanced data and how it affects the classifier.

Initially, we carry out an extensive experimental assessment of imbalances and overlaps across classes. Unlike previous studies, this experiment was carried out across a wide range of unbalanced class degrees and throughout the whole class overlapping spectrum. Second, using current techniques for handling unbalanced datasets, a thorough technical examination was carried out[26]. Class sharing and class coinciding tactics are

two categories into which the current approaches are divided after careful analysis. Furthermore, current advancements and new techniques are thoroughly reviewed.

Furthermore, the experimental results in this study are consistent with other studies in that they clearly show that the effectiveness of the training algorithm decreases with increasing levels of category overlap. On the other hand, unbalanced data has no effect at all. The review emphasizes how crucial it is to do further study on class overlap. Two noteworthy contributions have been made by Minh Dang et al.[27]: first, a tuned deep network for sewer fault diagnosis that is based on a box architectural style and consists of a series of fully connected layers that can effectively extract complex patterns from defective zones; and second, combination attachments of the developed framework that address the highly imbalanced problem using both an outfit strategy and an expense learning-based technique. The results of the trial showed that the proposed design performed better than previous sewage flaw-sensing devices and was not affected by unbalanced data problems. In real-world sewage fault analysis applications, the suggested defect prediction framework may promote more efficient faults provided in equations and ease the degree of connection of deep neural-based methodologies.

2.1. Oversampling in the categorization of unbalanced data

In the past, farmers or extension service workers from the agricultural management have often carried out this detection by carefully inspecting and supervising cassava plantations, after which the plantations are reported to Agricultural Advisory Services for further evaluation. It is expensive and time-consuming, however. It is unable to detect cassava illness in time to let farmers implement preventive measures to healthy leaves and seeds in order to boost yields, expand the market in Africa, and end hunger. Furthermore, convolutional neural networks (CN), one of the unbalanced datasets, were tested for robustness using food marketing databases[28]. Ultimately, it achieved a 93 percent accuracy rate in the illness affection prediction. Regression-vector-based voting-classifier (RVVC) is an ensemble method that has been developed to identify potentially dangerous comments on social media. The ensemble combines logic regression and support vector classifiers under benign voting circumstances. To evaluate the proposed model's operational performance, a series of tests are run on both balanced and unbalanced datasets. Furthermore, an artificial minority oversampling technique is used to the imbalanced dataset in order to bring the data back into balance.

To evaluate their suitability, two extraction techniques—maximum likelihood, bag-of-words, and document frequency—are also used. The recall, accuracy, F-score, and precision of the proposed method are compared to the results of many machine classifiers. For oversampling, the recorded accuracy is 0.95, the loss is 0.11, while for undersampling, the recorded accuracy is 0.88, the gain is 0.26. The regression strategy has also received an accuracy of 0.93 and a loss of 0.13 in the absence of the Sampling model[29].

Therefore, the goal of this study is to improve SMOTE's capacity to identify noise in synthetic data produced by treating imbalanced data with the Local-Outlier-Factor (LOF). In order to get the best appropriate classification result utilizing the imbalanced data, Asnicar et al. created the model LOF[30]. Furthermore, the results show that compared to traditional SMOTE, the developed SMOTE-LOF provides better accuracy and f-measurement. However, the AUC value for SMOTE was likely better at handling skewed data for datasets with fewer test data samples. Decision trees are gathered into collections called random forests (RF), where each tree is evenly and individually picked depending on a variable. Furthermore, the generalization error coheres to the limit if the number of trees is increased. Furthermore, each tree's intensity inside the RF is directly correlated with the generalization error of the tree classifiers.

Using logistic regression (LR) analysis, the relationship between a single binary classifier and a collection of alternative (explanatory) components is examined. Furthermore, when the answer variable only contains two input components, like 0 and 1. Thus, this input element provides a yes-or-no response to the logistical inquiries. Additionally, the multimodal LR may describe the multi-objective classes in a single run since it

contains more input components. It is also used for detection in a number of digital applications. A classifier is expressed by a recurrent split of the data space called a decision tree (DT).

In addition, the clustering technique consists of vertices that form a driven, grounded tree devoid of incoming edges. There is just one incoming edge per node that is still in existence. Furthermore, an internal node provides benefits that are outward-facing. “Leaf” refers to each extra node. Additionally, depending on the discrete process of the input feature values, each inner leaf node splits the occurrence region into two or more sub-spaces. The k-nearest-neighbors (kNN) algorithm is a simple but effective method for classifying objects. The main drawback of k-means is how ineffective it is in dynamic situations. Furthermore, in order to get the best prediction or classification outcomes, it is dependent on other sub-learning models. But this KNN model is simpler to use and less sophisticated.

2.2. Profound understanding

Skin disease diagnosis using deep convolutional-based neural networks (DCNM) has been thoroughly studied. Therefore, some techniques provide the best diagnostic results, on par with or even better than dermatologists. Furthermore, the incompleteness and imbalance of publicly accessible skin-lesion datasets hinder the broad use of DCNM in diagnosing skin disorders. Therefore, for small and imbalanced datasets, Peng Yao et al.[31] have provided a novel method for categorizing skin lesions using a specific model. In order to show that models of intermediate complexity performed better than models of higher complexity, numerous DCNMs are first taught on a variety of tiny and imbalanced datasets. In order to solve the issues of sample underrepresentation in the short dataset, a Revised Rand Augment technique is offered, and regularization Drop Block and Drop Out are added as a second measure to reduce overfitting.

The surface imperfection zone of strip steel is quite small, with a variety of fault shapes and intricate gray structures. Existing machine vision techniques are unable to distinguish flaws in various types of steel strips because of the high frequency of false faults and interference from edge light. In order to train the network for picture recognition using deep learning, a large number of photographs are needed. Using the other hand, the widely used deep network training tasks may be completed on a smaller sample set that contains imbalanced class errors. As a result, Xiang Wan et al.[32] have offered a collection of methods based on fast pre-process algorithms and transfer learning mechanisms for full-strip steel fault diagnosis. These methods have made it possible to quickly screen the surface, extract features from flaws, balance the categories in a sample dataset, forecast defeats, amplify data, and categorize results. Furthermore, it was found that the improved VGG19 network achieved 97.8 percent recognition accuracy.

The sample class is very skewed, and allocation is a challenging task. Repetition bias affects many common machine learning and statistical classification methods, making it challenging to learn how to distinguish between majority and minority classes. K. Ruwani M et al.[33] introduced a class rebalancing strategy based on a dynamically balanced class with weights given based on the class occurrence and the expected probability of regression coefficients class in order to address the unbalanced class distributions in deep neural ideas. With the dynamic weighted scheme's ability to deliberately modify its values in response to choices, the system may be tuned for a range of complicated scenarios, producing gradient updates driven by intricate class samples.

Gradient boost (GB) is being introduced for regression models and machine learning classifiers[34,35]. Additionally, GB is the only form of Ml that contributes to improving the classification parameters[36]. The GB has adjusted the classification parameter to the required level when it is applied to any classifier model. Therefore, the optimal output will be obtained. Furthermore, the GB and XGboost (XGB) worked together to provide a portable enough library function that allowed the GB to run on all systems for a variety of applications and uses[37]. The neural techniques were used with the cross-entropy (CE) models for the

smoothing process. The operation and procedure of the CE will vary depending on the characteristics of the neural approach[38].

The data balance of multi-label categorization is intrinsically off. The issue of class imbalance has persisted in being a barrier to inter-categorization in spite of several investigations. Think about the sixteen-label classification task. There are several viable options in the identifier subset. Therefore, it is not feasible to get a balanced database for every combination of labels. Many studies on inter-categorization either ignore the imbalance or try to rescale the dataset to balance it. Nevertheless, it is challenging to adapt the under- and over-sampling algorithms to this setting since they were not designed for multi-label classification. Using the inverse subclass frequency per weighted category is one common heuristic. Furthermore, the Dynamically Weighted (DW) model may balance the training loss when used with machine learning techniques[34].

Furthermore, in order to switch the class imbalance data, the updated Focal loss has incorporated both the dice and cross entropy based on losses. Nevertheless, there were two main problems with the improved Focal loss approaches in real-world applications[35]. In order to balance the data and validate the likelihood of errors occurring in each data point, cross-entropy, or CE, is a very helpful tool. However, the error calculation rate decreases with unstructured data[36].

2.3. Models for optimization in unequal data categorization

Elevated imbalanced data makes it difficult to create a suitable classifier, which severely reduces classifier efficiency[37,38]. Despite the fact that several techniques, such ensemble learning models, cost-sensitive models, and oversampling, have been created to deal with incorrect statistics, they are hindered by complete data that contains redundancy and distortion. Therefore, an adaptive subspace optimum ensemble technique (ASOEM) for high-dimensional imbalanced data categorization was developed to get over limitations[39]. To generate several robust and discriminative embeddings, a novel adaptive-subspace optimization (ASO) method and rotational-subspace optimization (RSO) are developed. Once again, the optimized subspace is resampled to provide class-balanced data for every classifier. Several experiments have been carried out to show the usefulness of the created ASOEM. The Whale optimization (WO) for the unbalanced data specification was first presented by Eslam et al.[40]. The whale is one of the most ostentatious animals in the water. They are acknowledged as the planet's biggest creatures. They were considered in relation to the categorization of unbalanced data.

The seven main species of this enormous mammal are the minke, killer, Sei, right, humpback, blue, and finback[41]. Most people think of whales as predators. Furthermore, despite breathing air from the water, it never sleeps. Actually, only half of the brain is used during sleep. Whales may live in groups or alone, which makes their social behavior another fascinating topic. They are often seen in groups. The technique of classifying unbalanced data has made use of this social behavior and hunting fitness. The validation has produced the maximum AUC of 99 percent for the WO. However, the average classification exactness score that it has recorded is 81%. Furthermore, the categorization procedure has become more complicated as a result of the unbalanced data. Thus, before beginning data mining or any other prediction process, the data must be balanced. For the aim of majority class definition, Sayan Surya Shaw et al.[42] have presented a particle swarm algorithm (PSA). Thus, 15 real-time unbalanced datasets were used to assess the SA's performance. In several instances, the swarm paradigm has produced subpar outcomes. So, the parameters were documented after the Ring theory with PSA (RTSA)[42] was run. Additionally, the real-time databases were examined using the Ring theory on evaluation-learning (RTEL)[41].

A popular meta-heuristic for local search that is often used to address intermittent and lower degree continuous optimization problems is called simulated annealing, or SA. Furthermore, the primary benefit of the SA is that it allows local optima to utilize permission to climb hills in order to reach the global optimum.

For the problems with unbalanced data classification, this ideal approach is applied[43]. As a result, 96.63 is the reported G-mean score.

Forecasting relies heavily on customer loyalty, and the telecoms and banking sectors are closely intertwined. Consequently, several sectors used diverse measures to foster a robust relationship with their clientele and reduce the rate of user attrition[44]. The variables that lead to churning and the dynamic consumer hierarchy are critical learning criteria to attain maximum client loyalty. Additionally, the rain optimization (RO) for the customer churn prediction unbalanced data has been simulated by Irina et al.[44].

RO is a new kind of heuristic that draws inspiration from the way raindrops fall on small areas when they hit the ground. Furthermore, this approach can detect both local and global extremes if its parameters are adjusted appropriately.

3. Analysis of performance

A number of techniques for classifying the unbalanced data were presented. As a result, several data kinds from industries including business, e-learning, medical agriculture, and so forth were employed. Based on the calculated categorization rate, each approach has unique properties. Furthermore, the uneven data in the data mining area is increasingly complicated, therefore classifying the imbalanced data requires an effective neural model. Furthermore, each approach's efficacy is verified by the estimation of critical metrics including accuracy, precision, recall, F-measure, and AUC.

To calculate each dataset's classification exactness score, the parameter correctness has been evaluated. The accuracy has been computed by dividing the value by the total number of samples. Therefore, Equation (1) equates the accuracy formulation, while Equation (2) describes the precision formulation (2).

$$Accuracy = \frac{True\_positives + True\_negative}{Total \quad Samples} \tag{1}$$

$$Precision = \frac{True\_positives}{False\_positives + True\_positives} \tag{2}$$

The accuracy metrics was computed in order to quantify the positive values in the prediction scenario.

$$recall = \frac{True \quad Positives}{false \quad negative + \quad true \quad positives} \tag{3}$$

Recall is also known as sensitivity, and the robustness score based on several datasets is measured by confirming the sensitivity. Equation describes the formulation of the model, which has the best classification exactness score if it has the greatest recall validation (3).

$$F - score = 2 \times \frac{recall \times precision}{Recall + precision} \tag{4}$$

The F-measure parameters, which are expressed in Equation, were measured in order to determine the mean value of the metrics recall and accuracy (4).

In this case, the RF technique received the following scores: recall 78 percent, F-measure 83 percent, accuracy 94 percent, and unbalanced data specification exactness score of 92 percent. For the purpose of categorizing unbalanced data, the KNN model has improved its accuracy by 89%, precision by 86%, F-score by 78%, and recall by 74%. The method DT has produced a 91 percent unbalanced data classification score, 84 percent accuracy, 85 percent F-measure, and 85 percent sensitivity. In addition, the LR scheme has obtained a 91 percent unbalanced data classification precision score, a 94 percent accuracy, an 89 percent f-score, and an 87 percent sensitivity. The unbalanced data classification exactness score, as reported by the RV0VC model, is 93%, accuracy is 91%, sensitivity is 85%, and the f-measure is 88%. The C4,5 scheme produced a 71 percent

accuracy for classifying unbalanced data, a 58 percent precision, a 59 percent F-score, and a 60 percent recall. With unbalanced data specification, the model NB has improved its exactness score to 76%, accuracy to 67%, F-score to 63.80%, and sensitivity to 60%. In conclusion, the support-vector model (SWM) has improved its unbalanced data classification accuracy by 77.25 percent, precision by 73.4 percent, f-score by 62.5 percent, and recall by 54.5 percent. The knowledge-based system[26] has achieved the highest accuracy rate overall, at 99 percent.

The metrics Area-Under-Curve (AUC) have been tallied in order to quantify the positive values in the prediction results. The model performs well in the prediction function if it has increased to the highest AUC rate. This review's primary goal is to examine the advantages and disadvantages of uneven data categorization methods. Thus, after examining a number of techniques, certain optimization models have achieved a 100% F-measure. Furthermore, SA and PSA are the models. Based on the data complexity and unstructured rate, the two algorithms haven't achieved a 100% F-score for all datasets; hence, the F-score may vary. Additionally, such methods have taken longer to finish all of the iterations necessary to arrive at the ideal answer.

4. Conclusion and future work

By examining the many available classification methods, this review paper seeks to determine the future path for the categorization of unbalanced data. AUC, precision, F-measure recall, accuracy, and other important metrics were also used to determine the literature's efficiency. As a result, the model SA has earned the 100% f-measure via comparison of these metrics, indicating that it satisfies the unbalanced data classification applications. The greatest results in categorizing the unbalanced data have been obtained using a number of ML and DL techniques. Additionally, different classifiers are used in various ways by combining optimizations and oversampling methods. Nevertheless, a lot of the algorithms did not use enough resources. Thus, creating hybrid deep learning models with hybrid optimization models in the future will provide several functions used for multi-imbalanced data categorization with the best exactness score.

Conflict of interest

The authors declare no conflict of interest.

References

1. Yin X, Liu Q, Pan Y, et al. Strength of Stacking Technique of Ensemble Learning in Rockburst Prediction with Imbalanced Data: Comparison of Eight Single and Ensemble Models. Natural Resources Research. 2021, 30(2): 1795-1815. doi: 10.1007/s11053-020-09787-0
2. Dogan A, Birant D. Machine learning and data mining in manufacturing. Expert Systems with Applications. 2021, 166: 114060. doi: 10.1016/j.eswa.2020.114060
3. Thakkar H, Shah V, Yagnik H, et al. Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis. Clinical eHealth. 2021, 4: 12-23. doi: 10.1016/j.ceh.2020.11.001
4. Pan Y, Zhang L. A BIM-data mining integrated digital twin framework for advanced project management. Automation in Construction. 2021, 124: 103564. doi: 10.1016/j.autcon.2021.103564
5. Espadinha-Cruz P, Godina R, Rodrigues EMG. A Review of Data Mining Applications in Semiconductor Manufacturing. Processes. 2021, 9(2): 305. doi: 10.3390/pr9020305
6. Jedrzejowicz J, Jedrzejowicz P. GEP-based classifier for mining imbalanced data. Expert Systems with Applications. 2021, 164: 114058. doi: 10.1016/j.eswa.2020.114058
7. Liu P, Qingqing W, Liu W. Enterprise human resource management platform based on FPGA and data mining. Microprocessors and Microsystems. 2021, 80: 103330. doi: 10.1016/j.micpro.2020.103330
8. Al-Hashedi KG, Magalingam P. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. Computer Science Review. 2021, 40: 100402. doi: 10.1016/j.cosrev.2021.100402
9. Sanad Z, Al-Sartawi A. Financial Statements Fraud and Data Mining: A Review. Lecture Notes in Networks and Systems. Published online 2021: 407-414. doi: 10.1007/978-3-030-77246-8_38
10. Shabtay L, Fournier-Viger P, Yaari R, et al. A guided FP-Growth algorithm for mining multitude-targeted item-sets and class association rules in imbalanced data. Information Sciences. 2021, 553: 353-375. doi: 10.1016/j.ins.2020.10.020

11. Aminian E, Ribeiro RP, Gama J. Chebyshev approaches for imbalanced data streams regression models. Data Mining and Knowledge Discovery. 2021, 35(6): 2389-2466. doi: 10.1007/s10618-021-00793-1
12. Korycki Ł, Krawczyk B. Low-Dimensional Representation Learning from Imbalanced Data Streams. Lecture Notes in Computer Science. 2021, 629-641. doi: 10.1007/978-3-030-75762-5_50
13. Grzyb J, Klikowski J, Woźniak M. Hellinger Distance Weighted Ensemble for imbalanced data stream classification. Journal of Computational Science. 2021, 51: 101314. doi: 10.1016/j.jocs.2021.101314
14. Lu N, Yin T. Transferable common feature space mining for fault diagnosis with imbalanced data. Mechanical Systems and Signal Processing. 2021, 156: 107645. doi: 10.1016/j.ymssp.2021.107645
15. Sisodia D, Sisodia DS. Data sampling strategies for click fraud detection using imbalanced user click data of online advertising: An empirical review. IETE Technical Review. 2021, 39(4): 789–798. doi: 10.1080/02564602.2021.1915892
16. Alican D, Birant D. Machine learning and data mining in manufacturing. Expert Systems with Applications 2021, 166: 114060.
17. Mirzaei B, Nikpour B, Nezamabadi-pour H. CDBH: A clustering and density-based hybrid approach for imbalanced data classification. Expert Systems with Applications. 2021, 164: 114035. doi: 10.1016/j.eswa.2020.114035
18. Chen S xia, Wang X kang, Zhang H, et al. Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine. Expert Systems with Applications. 2021, 173: 114756. doi: 10.1016/j.eswa.2021.114756
19. Zhu S. Analysis of the severity of vehicle-bicycle crashes with data mining techniques. Journal of Safety Research. 2021, 76: 218-227. doi: 10.1016/j.jsr.2020.11.011
20. Yang K, Yu Z, Chen CLP, et al. Incremental weighted ensemble broad learning system for imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 2021, 34(12): 5809-5824. doi: 10.1109/TKDE.2021.3061428
21. Pradipta GA, Wardoyo R, Musdholifah A, et al. Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning From Imbalanced Data. IEEE Access. 2021, 9: 74763-74777. doi: 10.1109/access.2021.3080316
22. Wang W, Sun D. The improved AdaBoost algorithms for imbalanced data classification. Information Sciences. 2021, 563: 358-374. doi: 10.1016/j.ins.2021.03.042
23. Hou C, Wu J, Cao B, et al. A deep-learning prediction model for imbalanced time series data forecasting. Big Data Mining and Analytics. 2021, 4(4): 266-278. doi: 10.26599/bdma.2021.9020011
24. Pereira RM, Costa YMG, Silla Jr. CN. Toward hierarchical classification of imbalanced data using random resampling algorithms. Information Sciences. 2021, 578: 344-363. doi: 10.1016/j.ins.2021.07.033
25. Wang X, Xu J, Zeng T, et al. Local distribution-based adaptive minority oversampling for imbalanced data classification. Neurocomputing. 2021, 422: 200-213. doi: 10.1016/j.neucom.2020.05.030
26. Vuttipittayamongkol P, Elyan E, Petrovski A. On the class overlap problem in imbalanced data classification. Knowledge-Based Systems. 2021, 212: 106631. doi: 10.1016/j.knosys.2020.106631
27. Dang LM, Kyeong S, Li Y, et al. Deep learning-based sewer defect classification for highly imbalanced dataset. Computers & Industrial Engineering. 2021, 161: 107630. doi: 10.1016/j.cie.2021.107630
28. Sambasivam G, Opiyo GD. A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks. Egyptian Informatics Journal. 2021, 22(1): 27-34. doi: 10.1016/j.eij.2020.02.007
29. Rupapara V, Rustam F, Shahzad HF, et al. Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification Using RVVC Model. IEEE Access. 2021, 9: 78621-78634. doi: 10.1109/access.2021.3083638
30. Asniar, Maulidevi NU, Surendro K. SMOTE-LOF for noise identification in imbalanced data classification. Journal of King Saud University-Computer and Information Sciences. 2021, 34(6): 3413-3423. doi: 10.1016/j.jksuci.2021.01.014
31. Yao P, Shen S, Xu M, et al. Single model deep learning on imbalanced small datasets for skin lesion classification. IEEE Transactions on Medical Imaging. 2021, 41(5): 1242-1254. doi: 10.1109/TMI.2021.3136682
32. Wan X, Zhang X, Liu L. An Improved VGG19 Transfer Learning Strip Steel Surface Defect Recognition Deep Neural Network Based on Few Samples and Imbalanced Datasets. Applied Sciences. 2021, 11(6): 2606. doi: 10.3390/app11062606
33. Fernando KRM, Tsokos CP. Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. IEEE Transactions on Neural Networks and Learning Systems. 2022, 33(7): 2940-2951. doi: 10.1109/TNNLS.2020.3047335
34. Yilmaz SF, Kaynak EB, Koç A, et al. Multi-Label Sentiment Analysis on 100 Languages With Dynamic Weighting for Label Imbalance. IEEE Transactions on Neural Networks and Learning Systems. 2023, 34(1): 331-343. doi: 10.1109/TNNLS.2021.3094304
35. Kim Y, Lee Y, Jeon M. Imbalanced image classification with complement cross entropy. Pattern Recognition Letters. 2021, 151: 33-40. doi: 10.1016/j.patrec.2021.07.017
36. Yan Z, Wen H. Electricity Theft Detection Base on Extreme Gradient Boosting in AMI. IEEE Transactions on Instrumentation and Measurement. 2021, 70: 1-9. doi: 10.1109/tim.2020.3048784

37. Nguyen HTT, Chen LH, Saravanarajan VS, et al. Using XG Boost and Random Forest Classifier Algorithms to Predict Student Behavior. 2021 Emerging Trends in Industry 40 (ETI 40). 2021. doi: 10.1109/eti4.051663.2021.9619217
38. Dong Y, Shen X, Jiang Z, et al. Recognition of imbalanced underwater acoustic datasets with exponentially weighted cross-entropy loss. Applied Acoustics. 2021, 174: 107740. doi: 10.1016/j.apacoust.2020.107740
39. Xu Y, Yu Z, Chen CLP, et al. Adaptive Subspace Optimization Ensemble Method for High-Dimensional Imbalanced Data Classification. IEEE Transactions on Neural Networks and Learning Systems. 2023, 34(5): 2284-2297. doi: 10.1109/tnnls.2021.3106306
40. Hassib EslamM, El-Desouky AliI, Labib LabibM, et al. WOA + BRNN: An imbalanced big data classification framework using Whale optimization and deep neural network. Soft Computing. 2019, 24(8): 5573-5592. doi: 10.1007/s00500-019-03901-y
41. Li Z, Zhang Q, He Y. Modified group theory-based optimization algorithms for numerical optimization. Applied Intelligence. 2022, 1-24.
42. Shaw SS, Ahmed S, Malakar S, et al. Hybridization of ring theory-based evolutionary algorithm and particle swarm optimization to solve class imbalance problem. Complex & Intelligent Systems. 2021, 7(4): 2069-2091. doi: 10.1007/s40747-021-00314-z
43. Desuky AS, Hussain S. An Improved Hybrid Approach for Handling Class Imbalance Problem. Arabian Journal for Science and Engineering. 2021, 46(4): 3853-3864. doi: 10.1007/s13369-021-05347-7
44. Pustokhina IV, Pustokhin DA, Nguyen PT, et al. Multi-objective rain optimization algorithm with WELM model for customer churn prediction in telecommunication sector. Complex & Intelligent Systems. 2021, 9(4): 3473-3485. doi: 10.1007/s40747-021-00353-6