DOI: 10.24996/ijs.2024.65.3.37

ISSN: 0067-2904

A Survey Study on Proposed Solutions for Imbalanced Big Data

Shaymaa Ahmed Razoqi [1]*, Ghayda A.A. Al-Talib[2]

[1]Department of Computer Science, College of Education for Pure Science, University of Mosul, Mosul, Iraq

[2]Department of Computer Science, College of Computer Sciences and Mathematics, University of Mosul, Mosul, Iraq

Received: 25/9/2022 Accepted: 27/3/2023 Published: 30/3/2024

Abstract

Learning from imbalanced data has been a focus of studies for more than two decades of continuous development. Training data is considered imbalanced when the size of the positive (minority) class is neglected because of the large size of the negative (majority) class, in addition to the problem of deviating distributions of binary tasks. The appearance of big data brings new problems and challenges to the imbalance problem. Big Data announces the challenges with 5V: volume, velocity, veracity, value, and variety. This study relied on dividing the solution to the problem of data imbalance into three levels: data level, algorithm level, and hybrid approaches. First, the standard solutions for this problem that were proposed were mentioned, and in addition, the most important metrics adopted for measuring the classification efficiency of imbalanced data were identified. In this survey study, 27 studies were reviewed during the period 2015–2022, distributed according to the levels of treatment of the imbalance problem. They also reviewed the performance metrics that were used in these studies and the sources of the datasets to which these solutions were applied. The study makes it easier for researchers and scholars to see the solutions to addressing the problem of data imbalance and the hybrid approaches recently used for that, and to take advantage of them in improving the classification process.

Keywords: Imbalanced Data, Machine Learning, Resampling methods, Classifier Performance metrics, Ensemble classifiers.

دراسة مسحية حول حلول مقترحة لحالة عدم التوازن في البيانات الضخمة

شيماء احمد رزوقي[1]*, غيداء عبدالعزيز[2]

[1]قسم علوم الحاسوب, كلية التربية للعلوم الصرفة، جامعة الموصل, الموصل, العراق

[2]قسم علوم الحاسوب, كلية علوم الحاسوب والرياضيات, جامعة الموصل, الموصل, العراق

الخلاصة

ظل التعلم من البيانات غير المتوازنة محور تركيز الدراسات بالرغم من مرور أكثر من عقدين من التطوير المستمر. بدءًا من مشكلة انحراف توزيعات المهام الثنائية ، فإن بيانات التدريب تكون غير متوازنة عندما يطغى عدد العينات في فئة (الأقلية) الإيجابية على عدد العينات في فئة (الأغلبية) السلبية. كما ان ظهور البيانات الضخمة استحدث مشاكل وتحديات جديدة لمشكلة اختلال التوازن الطبقي في البيانات، حيث ان البيانات الضخمة شكلت خمس تحديات اطلق عليها 5 :V الحجم والسرعة والدقة والاهمية والتنوع .

*Email: shymaa.raazoqi@uomosul.edu.iq

اعتمدت هذه الدراسة على تقسيم حلول مشكلة عدم توازن البيانات إلى ثلاثة مستويات: مستوى البيانات ، ومستوى الخوارزمية ، والطرق الهجينة. حيث تم أولاً ذكر الحلول المعيارية المقترحة لهذه المشكلة، إضافة إلى تحديد أهم المقاييس المعتمدة في قياس كفاءة تصنيف البيانات غير المتوازنة. تمت مراجعة 27 دراسة في هذه الدراسة الاستقصائية خلال الفترة من 2015 إلى 2022 موزعة حسب مستويات علاج مشكلة عدم التوازن، كما استعرضت الدراسة الحالية مقاييس الأداء التي تم استخدامها في هذه الدراسات ومصادر البيانات التي تم تطبيق هذه الحلول عليها. تسهل الدراسة على الباحثين والدارسين معرفة طرق معالجة مشكلة عدم توازن البيانات ، والتقنيات الهجينة المستخدمة حديثًا لذلك ، للاستفادة منها في تحسين عملية التصنيف.

1. Introduction

In the last two decades, there has been special interest in classification approaches for imbalanced data sets. Classifiers such as Naïve-Bayes (NB), Decision Tree (DT), K-Nearest Neighbor (KNN), and neural networks perform competently if the class distribution is balanced in the given classification problem dataset. But in real-world issues like text classification, network intrusion detection, fault detection, spam review detection [1], fraud detection, or medical classification, the data set is often highly imbalanced [2]. Realistic data sets are imbalanced in that one class contains a smaller number of instances (a minority) than another class (a majority), in which the number of instances is predominantly very large.

Supervised learning in the training process is often based on a balanced distribution of classes. In the case where the distribution of the classes is imbalanced, the unavailability of the target class data, compared to the required detection data, presents a challenge to detection and leads to poor performance and uncertainty in the results of applying these approaches [3]. The majority class makes the performance of data learning algorithms skew towards it, as most algorithms measure performance accuracy based on the proportion of correct instances classified. Thus, the results are often deceptive because minority classes have little influence on overall accuracy. Then, the performance of the majority class overshadows the insignificant performance of the minority class [4]. The imbalanced data classification generally uses accuracy, precision, recall, G-mean, F-measure, and receiver operating characteristic (ROC) curves as metrics to estimate performance for the learning algorithms [3, 5].

Three levels of solutions can be identified to deal with the imbalance of the data set: the data level, the algorithm level, and hybrid approaches. The policy of the data level includes resampling methods used to minimize the imbalance rate in classification classes. Two essential resampling methods are random over sampling (ROS) and random under sampling (RUS). ROS causes minority class samples to be duplicated randomly, while RUS causes samples of the majority class to be randomly removed as a way to modify class distribution, where oversampling can guide the classification operation to an overfitting problem because it makes the same copies of minority instances, while at the same time, little sampling results in information loss for the majority class in instances that are ignored [2]. The algorithm level focused especially on adjusting the classification algorithm and its learning method so the classifier may be adjusted corresponding to the imbalanced textures of data and improve the remember capability of the classifier for positive instances, also increasing the accuracy of the implemented classifier [1]. Hybrid approaches combine data-level and algorithm-level approaches.

The study aims to give a comprehensive idea of the proposed solutions to the problem of data imbalance. For this reason, Paragraph 2 is devoted to defining standard solutions at the three levels and the most important approaches used for this. In Paragraph 3, the common standards that are relied upon to measure classification efficiency in imbalanced data were also discussed. Paragraphs 4 and 5 dealt with a review of 27 previous studies, the techniques they used, the data sources, and the classification measures they relied on.

2. The imbalanced solutions

To study the solutions developed to solve this problem, solutions can be roughly divided into three types of study approaches: data-level, algorithm-level, and hybrid approaches [1].

2.1 Data-Level Methods

Mainly, the primary objective of this approach was to balance data that had a skewed distribution of instances in classes [6]. The sampling methods were implemented by resizing the training datasets to balance the class distribution [7]. In those methods, only training data selection is changed, and no changes are done in classifiers [8].

Solutions at the data level include many different forms of resampling, such as selecting samples from the minority to reconstruct a new sample from, selecting the majority samples to be deleted, sampling the minority by generating new synthetic data from an original group, and combinations of more than one method [9]. Several methods were proposed at the data level during the previous period to deal with imbalanced data.

ROS is a non-heuristic method that balances the class distribution through the random replication of positive instances. It is the simplest method to increase the size of the minority class by replicating existing examples corresponding to the minority class. With a ROS, overfitting is more likely to occur [10, 11]. Synthetic minority over-sampling technique (SMOTE) is a process of increasing samples in minority samples by creating new artificial instances using existing samples based on the KNN algorithm [7]. Several modifications of the original SMOTE method were proposed during the previous period, such as the use of statistical methods, fuzzy logic [12], and linear optimization. While the SMOTE approach does not deal with all attributes of a data set, it has been generalized to deal with imbalanced data sets with continuous and nominal attributes [10].

The adaptive synthetic over-sampling technique (ADASYN) increases the learning accuracy in the data distributions by using weighted distributions for different minority examples according to their level of difficulty in learning [7, 13]. Borderline-SMOTE alters SMOTE, and the border instances belonging to the minority class were selected only to perform SMOTE on them. The algorithm computes the majority neighbors’ quantity for every minority instance and then splits up the minority instances according to that into three groups: the noise group, the danger group, and the safe group. Danger group instances are just used to generate artificial instances [14].

The random undersampling method was the most important for sampling imbalanced data. RUS attempts to balance the class distribution by randomly removing an instance of the majority class. This generates the problem of losing valuable information [15]. Because of its efficiency, simplicity, and speed, RUS has shown very good performance, and it is used in boosting for these reasons [7, 11]. Tomek Link (T-Link) is an under-sampling sampling strategy announced by Tomek. T-Link is based on upgrading the Nearest Neighbor Regulation (NNR). T-link has also been classified as an improved nearest neighborhood rule. T-Link technology may be an undersampled method by removing the detected majority instance closest to the minority class by involving the nearest neighbors rule for selecting examples [7]. This method can also be over/under-sampled when removing instances from minority and majority classes; this is due to the difficulty in identifying well-defined boundary regions. Edited Nearest

Neighbor (ENN) undersampling of the majority class is prepared by eliminating instances whose class label varies from a majority of its k nearest neighbors [16].

2.2 Algorithm-Level Approaches

Algorithm-level approaches are usually called internal approaches. At the algorithmic level, the solution focuses on the capability of improving the classifier to learn about minority classes [7]. These approaches concentrate on comprehending what exact learning operation degrades the performance of a classifier when dealing with imbalanced data. This needs an in-depth analysis of a given algorithm. Because these approaches are generally specific only to a given learner, these techniques offer high effectiveness at the cost of reduced flexibility [17].

To produce a solution at the algorithm level, an understanding of the learning algorithm and the domain of the application related to the problem classifier is needed, primarily a thorough understanding of the failure of the learning method with the imbalanced datasets [9]. At the algorithmic level, some solutions contain modifying the effective costs for different classes to solve the misbalancing [6]. For example, decision trees can be modified at the algorithm level in different ways; one technique is to modify the probabilistic appreciation for a tree's leaves, and the strategy of developing new pruning procedures is also used [9]. One-class learning techniques can be used at the algorithm level to recognize instances of one class and deny others. This approach optimizes the performance of the learning algorithm on unseen data [15].

In recent years, there has been a lot of interest in neural networks and deep learning, as well as solving the imbalance problem. At the algorithm level, for example, weights are upgraded by minimizing widespread error in a standard backpropagation algorithm, which is donated mainly by the majority class. Consequently, we accumulate classification results that are partisan [18].

2.3 Hybrid approaches

At the hybrid approach level, various resampling methods and boosting classification techniques from the data and algorithm levels must be combined. It can be seen as a wrap-up of other levels' methods. This method comprises preprocessing before data training and tuning the initial imbalanced data [6]. Whereas the data resampling method is designed to treat problems caused by imbalanced classes, the hybrid method can upgrade the performance of any weakened classifier that happened regardless of imbalance in the dataset [7]. Ensemble classifiers can be considered hybrid methods. The essential concept of a learner in the ensemble classifier is to form multiple classifiers that deal with the original data and later collect the prediction results of those classifiers to classify unknown samples [9]. Ensemble approaches apply the synthesis of different methods, and this commonly uses solutions based on bagging and boosting with class imbalances [7]. Bagging utilizes and integrates a number of learners using an averaging technique to reduce variation and bias, and these approaches give fine results [19]. This may show that a low learner can transform into a powerful learner, and this solution may also be slightly better than guessing randomly. Unlike bagging, each sample of data boosts weights. AdaBoost is the family's most influential algorithm. Boosting needs bootstrapping, which means that some samples will be run more frequently than others [7, 18].

In general, every classifier has its drawbacks and can make mistakes if it is trained on a limited set of data. It is assumed that the patterns that have been incorrectly classified by different classifiers are not necessarily the same. Therefore, using more than one classifier and studying the effect of results in terms of statistical concepts of bias and variance gives higher efficiency in the process of classifying unbalanced data [9]. For example, Chawla, Nitesh V., et al. (2003) proposed SMOTEBoost, using SMOTE with a standard boosting procedure.

While creating synthetic examples from the minority class indirectly changes the updating weighted [20]. SEIFFERT, Chris, et al. (2009) designed RUSBoost as an alternative to SMOTEBoost to solve problems of complexity and increased training time because RUSBoost was simple and had a faster training time but suffered from information loss [21].

The Locality Informed Underboosting (LIUBoost) method was the first to combine both sampling and cost-sensitive learning in 2017. LIUBoost does not suffer from information loss. Both majority and minority instances are divided into categories, and the hard instances are given special importance in the imbalanced dataset [22].

3. Performance metrics

Most of the studies in these imbalanced areas essentially focus on a two-class problem. By categorizing the minority class's label as positive and the majority class's label as negative [23]. By the confusion matrix for a binary classification problem, a related list of simple performance metrics is explained as follows:

- True positive (TP) means the positive instances' number, which is correctly classified as a positive class label.
- True negative (TN) means the negative instances' number, which is correctly classified as a negative class label.
- False positive (FP) means the negative instances' number, which is incorrectly identified and has the wrong label as positive.
- False negative (FN) means the positive instances' number, which is incorrectly identified and has the wrong label as negative.
- TP-rate, also known as Recall or Sensitivity, is equal to TP / (TP + FN).
- TN-rate, also known as Specificity, is equal to TN / (TN + FP).

The most common performance metric used is accuracy. But it does not distinguish between positive and negative instances when computed. Accuracy represents the rate of instances that are correctly classified compared to the total number of instances classified, as in Eq. (1), and this is not wanted in this case, so we must use additional metrics widely.

$$Accuracy= (TP+TN)/ (TP+FN+FP+TN) \quad (1)$$

The other performance metric is G-Mean, which indicates how well the model performs at the threshold where TP-rate and TN-rate are equal, as shown in Eq. (2) [14].

$$G_{\_mean} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \quad (2)$$

Furthermore, F-score (F-measure) is a famous evaluation metric when dealing with imbalance problems as in Eq. (3) [24]. It combined recall and precision, which are categorized as effective metrics in imbalanced data. For a more general F-score Fβ as shown in Eq. (4), the F-value must be high when both recall and precision are increased; the values of it can be adjusted by changing the β value, which is usually set to 1. The β value corresponds to the relative significance of precision versus recall in this equation [25].

$$F\text{-}score = (2*Recall * Precision) / (Precision+ Recall) \quad (3)$$

$$f_{\_measure} = \frac{(1+\beta\times\beta)\times recall \times Precision}{\beta\times\beta\times recall \times Precision} \quad (4)$$

Perhaps the most common metric is the area under the receiver operating characteristic (ROC) curve (AUC) analysis, which assesses overall classification performance, It is not placing more priority on one class than other classes, and this must make it not biased against the minority class. The ROC curve is a two-dimensional graph, where the FP rate represents the x-axis and the TP rate represents the y-axis. For the learner, the ideal point is (0, 1). The ROC curve depicts relative trade-offs between benefits, denoted by the TP rate, and costs, which are represented by the FP rate. ROC curves are produced by sampling, changing a threshold for the decision, or varying the matrix of the cost. Regardless of the approach used to generate ROC curves, the problem of selecting the best singular method and the best classifier for implementation in an intelligent application system still remained [23].

4. Reviewing the Related Works

With the emergence of big data in the recent period, interest has increased in the problem of class imbalance, which is a major obstacle to a good classifier in machine learning algorithms and reduces its efficiency. Previous studies dealt with comparisons between approaches to dealing with imbalanced data problems and used datasets with several different models in terms of the degree of randomness and at different levels. Classifiers such as K-nearest neighbors (KNN), random forest (RF), neural networks, and support vector machines (SVM) were also used. This section reviews some of these studies with a summary of each one. The studies are divided into three tables according to the level of handling of the problem of data imbalance. The later sections study the metrics and datasets used.

In Table 1, the studies that adopted the development of approaches for solving the problem of data imbalance are shown. Through the temporal progress of the studies, we note that the studies in the first date focused on the method of resampling instances in one of the classes (over-sampling minority or under-sampling majority), using one or more methods in the level of data. In recent years, studies have appeared that adopt a more complex method, focus on resampling instances in both classes, and use more than one method for each class. Thus obtaining higher efficiency results when classifying, while at the same time addressing the problem of over-fitting caused by new minority instances and reducing information loss for the majority class instances removed.

Table 1: Data level solutions

Year	Title	Algorithm used	Description
2015	An Improved Algorithm for Imbalanced Data and Small Sample Size Classification [26]	WRO with SVM and KNN as classifiers	Changing class distribution through adding virtual samples generated by the windowed regression oversampling method. The efficiency was not high because of the challenge of solving the regression coefficients in the local window. Disadvantages: many variables are used. It does not deal with high-dimensional data.
2016	Classification of Imbalance Data using Tomek Link Combined with Random Under-sampling as a Data Reduction Method [27]	Tomek-Link with LR, RF, ANN classifiers	Tomek-Link is used in the preprocessing phase as a method of data cleaning to remove noise. Removing noise observation from the majority class followed by resampling methods reduces the chance of information loss. Disadvantages: Depending very much on the way the data is distributed, some original data may be considered noise.
2017	A Novel Technique on Class Imbalance Big Data using Analogous over Sampling Approach [28]	OSIBD	Analyzing the minority samples and removing noisy or borderline instances using a recursive analogous oversampling strategy to improve knowledge discovery from the class imbalance big data Disadvantages: Depending very much on the way the data is distributed, some original data may be considered noise.
2020	Improving k-Nearest Neighbors Algorithm for Imbalanced Data Classification [29]	Improved KNN	Using resampling methods (ROS, SMOTE, and Borderline SMOTE) as preprocessing before a classification with KNN and comparing their results Disadvantages include information loss in the majority class due to random selection.
2021	Improved Sampling Data Workflow Using Smtmk To Increase The Classification Accuracy Of Imbalanced Dataset [5]	SMTMK with RF, LR and extreme GB as classifiers	Combination of SMOTE with Tomek-Link for resampling the dataset. The additional step is removing majority data based on the ratio of minority proposed. Disadvantages: deals with numeric attributes in this research.
2021	Comprehensive analysis for class imbalance data with concept drift using ensemble-based classification [30]	MOA	Massive Online Analysis (MOA) is used to compare learners. The synthetic data stream (SEA and KDD datasets) was implemented by learners. Disadvantages: It is affected by the pace of data flow and the nature of its change.
2021	Handling Imbalance Classification Virtual Screening Big Data Using Machine Learning Algorithms [31]	KSMOTE	The samples of the majority are clustered into K clusters using K-mean. Each majority cluster company used all minority samples to make K separate training datasets, and then SMOTE was applied to each dataset. Disadvantages: expensive processing, in addition to creating duplicate elements in more than one group.
2021	FDR2-BD: A Fast Data Reduction Recommendation Tool for Tabular Big Data Classification Problems [32]	FDR2-DB	Analyze data in a dual way (vertical and horizontal) using Spark. The first step was feature selection, the other drop-duplicate samples were processed using RUS, and then the DT classifier was used. Disadvantages include information loss in the majority class due to random selection.
2022	Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset [3]	SMOTE, Tomek-Link	SMOTE combined with Tomek-Link is used in condition monitoring on electrical machines. After that, study the performance of using NB, SVM, and KNN classifiers on monitoring systems. Disadvantages: high computation and information loss in majority class.

Since processing imbalanced data at the algorithm level requires in-depth analysis and study of the algorithm under modification, compared to the improvement obtained in classifying imbalanced data, few studies have addressed this type of method. Table 2 shows some of these studies, which were considered for this review.

Table 2: Algorithm-level solutions

Year	Title	Algorithm used	Description
2018	Convolutional Neural Network-Based Classification of Histopathological Images Affected by Data Imbalance [33]	CNN with NCL and RBO	Neighborhood Cleaning Rule (NCL) and Radial Based Oversampling (RBO) were used as preprocessing and resampling steps on bread cancer camera microscopic images. Weighted loss strategy of assigning a weight associated with the misclassification of an object based on its class. Disadvantages: the accuracy of the processed images is low, and it takes a lot of time and processing.
2018	Imbalanced data classification algorithm with support vector machine kernel extensions [34]	KE-SVM	the suggested algorithm trained the greatest margin classification of the SVM to achieve the initial classification results, and after that, another novel kernel function was obtained. The algorithm then used the Chi-square test when training the samples again with the new novel SVM to gain better classifier accuracy. Disadvantages: features must be independent. In each data sample, all classes must appear.
2019	A novel multi-module neural network system for imbalanced heartbeats classification [35]	CNN BLSM, CTFM and 2PT	BLSM is used to synthesize virtual samples linearly around the minority samples. CTFM consists of a DAE-based feature extraction part and a QRS-based feature selection part, which apply 2PT to processed data by feeding it into a convolutional neural network (CNN). Disadvantages: increasing the size of the data without fixing 33 problems with the BLSM method.
2021	Minority Class oriented Active Learning for Imbalanced Datasets [36]	Active-Learning	Design a new acquisition method for iterative active learning that focuses the selection process toward minority samples. Disadvantages: incompetent with high dimensional data.
2022	KNNOR: An oversampling technique for imbalanced datasets [37]	KNNOR	In K-Nearest Neighbor Oversampling, which studies the location of the minority class, the sample that is closer to their neighbors has higher priority; a sample that is farther apart is not totally ignored but has lower importance in generating new samples. Disadvantages: expensive processing; calculating the distance between minority elements more than once.

Since the results of modifying classification algorithms do not lead to significant improvements, especially with big data, researchers resort to using the hybrid method or the ensemble method. The use of resampling methods at the data level and then the forwarding of the resulting samples into the modified classifier results in higher efficiency, especially with the availability of financial capabilities to implement this in the recent period. Currently, hybrid studies have appeared remarkably, and this also applies to the solutions proposed to solve the problem of data misalignment when classifying. Table 3 shows the studies collected in this review that use hybrid solutions; it is noted that most of these studies were presented recently. Most of these studies are based on bagging and boosting, and we see the use of the Random Forest classifier as the most widely used when dealing with imbalanced data. This classifier gives high efficiency when dealing with imbalanced data, as it is designed by adopting the "bagging" principle, which is one way to address the problem of imbalanced data classification.

Table 3: Hybrid solutions

Year	Title	Algorithm used	Description
2019	Uncertainty based under- sampling for learning Naive Bayes classifiers under imbalanced data sets [38]	Under-sample with Naïve Bayes	A small, balanced subset of the available training set is picked at random and trains NB. The remaining set gets balanced using SMOTE and acts as a pool for active selection. Disadvantages: random selection of data samples and NB variables, it did not achieve high efficiency results despite being time-consuming.
2019	Classification of Imbalanced Big Data using SMOTE with Rough Random Forest [39]	SMOTE with RRF	The model consists of ensemble learning by rough random forest (RRF). The model uses a selection of random features and methods of bagging to generate a number of DTs using the approach of the overall boundary region. Disadvantages: allocated to a specific dataset, samples were chosen randomly, and the trees continued to grow.
2019	Improving Detection Accuracy for Imbalanced Network Intrusion Classification using Cluster-based Under-sampling with Random Forests [40]	Cluster-Based under-sampling	Multilayer classification with imbalanced network intrusion. Cluster-based undersampling was applied, and then ensemble RF was used to address the overfitting problem. Disadvantages: allocated to a specific dataset, results are based on class centering, and random selection is used.
2020	The Effects of Data Sampling with Deep Learning and Highly Imbalanced Big Data[41]	DNN-2, DNN-4 With ROS-RUS	Use deep neural networks with 2 and 4 hidden layers after using the hybrid sampling method (ROS-RUS) for detecting fraud in severely class-imbalanced data. Disadvantages: threshold values must be carefully chosen programmatically, which is expensive.
2020	Performance of RUS and SMOTE Method on Twitter Spam Data Using Random Forest [42]	RUS-SMOTE	Design an Imbalanced Classification Ensemble The big-data method uses sampling (RUS and SMOTE) on spam data as pre-processing before using a RF classifier. Disadvantages: incompetent with wildly imbalanced data.
2020	WOA + BRNN: An imbalanced big data classification framework using Whale optimization and deep neural network [43]	WOA with BRNN	Using while optimization with a bidirectional recurrent neural network in three steps: feature selection using WOA, preprocessing SMOTE, and training Deep NN using WOA. Disadvantages: expensive processing as WOA is used to select the attributes and choose the weights of the neural network.
2021	Classification of Imbalanced Data of Medical Diagnosis using Sampling Techniques [44]	MLPUS with MWMOTE	Weighted minority oversample (MWMOTE) is used for generating synthetic samples for the minority class, and Multi-Layer Perceptron Under-Sample (MLPUS) preserves the distribution information of the majority class. Disadvantages: incompetent with high-dimensional data. It's taking time.
2021	Hybrid Algorithm Based on Simulated Annealing and Bacterial Foraging Optimization for Mining Imbalanced Data [45]	BFO with Simulation Annealing	Borderline SMOTE determines borderline minority instances due to oversampling, then Bacteria-Foraging Optimization applies with SA, which is a process based on probability that avoids falling into a local optimum during the search process. Disadvantages: a lot of parameters have to be tuned, which takes more time.
2021	HSDP: "A Hybrid Sampling Method for Imbalanced Big Data Based on Data Partition" [24]	HSDP	Decomposed dataset into two regions: samples in the noise minority region were removed, and samples in the boundary minority region were weighted oversampled. The sample weight is computed using the ratio of the number of the majority of neighboring samples of this sample over the number of all its neighbors. Disadvantages: greatly affected by the nature of the distribution and overlapping of classes, especially when determining the noise samples.

2021	"Research on Expansion and Classification of Imbalanced Data Based on SMOTE Algorithm" [25]	SMOTE with Normal distribution RF-classifier	the Improving of SMOTE in Normal distribution leads to a distribution of the new samples near the minority class center, with a large possibility to avoid the examples on the edge of the dataset being marginalized. Disadvantages: production of new samples at the minority class center only.
2022	Towards an Effective Intrusion Detection Model Using Focal Loss Variational Autoencoder for Internet of Things[46]	Distributed Deep learning	Introduced an adaptive version of the focal loss function, the data available at each local node assumed that there were backup nodes for each local node that failed. Two training phases were used: pre-training using an autoencoder, then re-training with a class label. Disadvantages: expensive processing.
2022	VBLSH: Volume-balancing locality-sensitive hashing algorithm for K-nearest neighbors search [47]	VBLSH	Building a large-scale, high-dimensional data index structure with the volume-balancing search algorithm based on LSH, which is used to solve the KNN problem. Disadvantages: significant time and processing during the test phase.
2022	A Density-Based Random Forest for Imbalanced Data Classification [48]	DBRF	Density-based random forest (DBRF) solves the problem of boundary minority samples by calculating the density of samples in space and then augmenting them. Two RF were generated, and the final output was determined using a bagging technique. Disadvantages: Overlapping between classes leads to the removal of many samples and information loss.
2022	An Oversampling Method for Class Imbalance Problems on Large Datasets[49]	Fast SMOTE, Fast NDO, CART, NB	Accelerate the production of artificial elements based on a mathematical approach that uses the median values of the attributes and a variable coefficient to produce samples close to the center of the minority class in oversample methods. Disadvantages: random value for a variable gives inaccurate results.
2022	A Spark-Based Artificial Bee Colony Algorithm for Unbalanced Large Data Classification[50]	ABCC	Designing a parallel model of the ABC algorithm for data classification by reducing the weight of misclassified samples across the set of parallel processing points and improving the fitness function using F-score, and G-mean. Disadvantages: Not good with too many features.
2022	Statistical Analysis of the Performance of Four Apache Spark ML Algorithms[51]	RF, SVM, NB, MLP	Parallel processing is used via Spark ML to improve the execution time and performance of classification algorithms with high-dimensional data, comparing performance relative to the number of attributes and number of processing elements. Disadvantages: No practical method for feature selection has been defined.

KNN is one of the easiest machine learning methods used in classification, as it is free from complex arithmetic operations. It is widely used to develop or improve data resampling methods. It is also used to assist in the pre-processing of the rest of the classification technique.

5. Types of Datasets used in Related Works

Previous studies relied on different sources of imbalanced data; some of these studies relied on dataset sources provided by sites such as the UCI Machine Learning Repository [52], the KEEL Dataset Repository [53], and the Kaggle Machine Learning and Data Science Community [54]. From the previous reviews, nine of the studies relied on datasets from UCI only to estimate the proposed approaches' efficiency [25–28, 42–45, 48], two studies [38, 39, 49] used datasets from KEEL only, and studies [24] and [37] used datasets from both KEEL and UCI. The study [32] used datasets from both UCI and Kaggle, while three other studies used different datasets from the Kaggle website [5, 35, 37]. Previous studies dealing with imbalanced intrusion detection data solutions [30] and [40] used datasets from the KDD website. [41] used CMS Medicare Data, and [31] used the PubChem dataset library. While other studies devoted themselves to research dealing with a specific type of data, as in [33], the histopathological dataset was used and the fault dataset in [3] was used as classification training samples. Some researchers resorted to developing an artificial dataset that fits their proposed method [34]. Also, the dataset used in [50] was from the SEER repository. In [51], the high-dimensional dataset for cancer prediction is used.

6. Metrics used in Related Works

Through the study of previous articles, notes were made on the metrics used to measure the efficiency of the classification process. Previous studies used one or more of the measures mentioned in the previous paragraph to show the efficiency of the classification process when dealing with imbalanced data. Table4 show the metrics used in each study by putting χ signs in the corresponding cells.

Table 4: Metrics used in references (references sorted by year)

References	Recall	ROC-AUC	Precision	G-mean	F-measure	Specificity	Accuracy
[26] 2015				χ	χ		χ
[27] 2016	X	χ	χ	χ	χ		χ
[28] 2017	X	χ	χ				
[33] 2018		χ		χ	χ		
[34] 2018				χ	χ		
[35] 2019	X			χ	χ	χ	
[38] 2019		χ					χ
[39] 2019		χ					
[29] 2020			χ				
[41] 2020		χ		χ			
[42] 2020		χ		χ			
[43] 2020		χ					
[5] 2021		χ					
[24] 2021	X		χ	χ	χ		
[25] 2021	X	χ	χ	χ	χ		
[30] 2021	X		χ		χ		χ
[31] 2021	X			χ		χ	χ
[32] 2021				χ			χ
[37] 2021		χ		χ			χ
[44] 2021		χ					χ
[45] 2021	X	χ	χ				
[3] 2022	X		χ	χ	χ		
[36] 2022							χ
[46] 2022							χ
[47] 2022	X		χ				
[48] 2022		χ					
[49] 2022		χ					
[50] 2022				χ	χ		
[51] 2022	X	χ	χ		χ	χ	χ

The ROC-AUC is the most widely used measure for all types of data and is preferred by researchers. The AUC value indicates how the model performs by distinguishing between the two classes, and when the AUC value increases, it indicates better performance and also more separability [55], so it is not significantly affected by bias and is usually used alone [1, 5, 12, 39, 43, 48, 49].

In the second degree comes the use of recall, precision, and G-mean. Notice that many researchers prefer to use a number of metrics rather than rely on a single metric to compare the proposed classification algorithms. Depending on the type of applications and the degree of data imbalance, the metrics that can be adopted are selected.

Intrusion detection systems have a slightly different data type; the detection rate is used as a sensitivity measure to measure the efficiency of the proposed workbook [56]. Miah [40] presented in his research the use of a "Detection Rate," in addition to the use of a false-positive measure.

Neural networks as a classifier differ from other classifiers in that they have special training properties. While the AUC scale is actively used to measure classification accuracy, neural networks use "training time" as a basic determinant of the efficiency of the classification process [19, 36, 41, 46]. However, researchers prefer to use other measures to enhance the process [35, 44].

7. Conclusion

Classification approaches give better performance whenever the available dataset is large, but this may lead to the emergence of a problem of data imbalance. During the previous years, many studies were presented that dealt with solutions to this problem. This review deals with a group of studies that presented solutions to the problem of data imbalance to improve the results of classification approaches.

Through the study of related works, we can conclude that the best solutions were produced when using hybrid approaches that integrated solutions at the data level and algorithms. In general, the evidence that distinguishes hybrid approaches was lacking. The ensemble approaches are preeminent for learning from imbalanced data, and there must be extra experiments before drawing such determinations. In conclusion, we cannot clearly compare the discussed approaches in our review since they experiment on different datasets with different imbalance ratios and those results were obtained using different performance measurements. Extra indicates that the performance of the suggested algorithm is strongly dependent on the labeling model of classes, the complexity of the problem, and the results' performance metrics. and thus, extra analysis is needed to test the performances of machine learning applications for classification in imbalanced domains.

8. Acknowledgements

First, we want to give our thanks to the University of Mosul, and we are very grateful and render thanks to our colleagues for helping to complete this research.

References

[1] Q. Dai, J.-w. Liu, and Y. Liu, "Multi-granularity relabeled under-sampling algorithm for imbalanced data," *Applied Soft Computing*, vol. 124, p. 109083, 2022. Available: https://doi.org/10.1016/j.asoc.2022.109083

[2] YAP, Bee Wah, et al, "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets," *in Proceedings of the first international conference on advanced* *data and information engineering (DaEng-2013)*, vol. 285: Springer Science & Business Media, 2013.

[3] F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors*, vol. 22, p. 3246, 2022.

[4] A. Fernández, S. del Río, N. V. Chawla, and F. Herrera, "An insight into imbalanced big data classification: outcomes and challenges," *Complex & Intelligent Systems*, vol. 3, pp. 105-120, 2017.

[5] M. S. A. bin Alias, N. B. Ibrahim, and Z. B. M. Zin, "Improved sampling data Workflow using Smtmk to increase the classification accuracy of imbalanced dataset," *European Journal of Molecular & Clinical Medicine*, vol. 8, p. 2021, 2021.

[6] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of imbalanced data: review of methods and applications," *in IOP Conference Series: Materials Science and Engineering*, 2021.

[7] K. M. Hasib, M. Iqbal, F. M. Shah, J. A. Mahmud, M. H. Popel, M. Showrov, et al., "A survey of methods for managing the classification and solution of data imbalance problem," *Journal of Computer Science*, vol. 16 no. 11, 2020.Available: https://doi.org/10.3844/jcssp.2020.1546.1557

[8] H. Patel, D. Singh Rajput, G. Thippa Reddy, C. Iwendi, A. Kashif Bashir, and O. Jo, "A review on classification of imbalanced data for wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 16, p. 1550147720916404, 2020.

[9] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International journal of pattern recognition and artificial intelligence*, vol. 23, pp. 687-719, 2009.

[10] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, pp. 42-47, 2012.

[11] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM Sigkdd Explorations Newsletter*, vol. 6, pp. 40-49, 2004.

[12] N. Verbiest, E. Ramentol, C. Cornelis, and F. Herrera, "Improving SMOTE with fuzzy rough prototype selection to detect noise in imbalanced classification data," *in Ibero-american conference on artificial intelligence, Lecture Notes in Computer Science*, vol 7637, Springer, Berlin, Heidelberg. 2012. Available: https://doi.org/10.1007/978-3-642-34654-5_18

[13] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *in 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), Hong Kong*, 2008.Available: doi: 10.1109/IJCNN.2008.4633969.

[14] T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy, and R. A. Bauder, "Severely imbalanced big data challenges: investigating data sampling approaches," *Journal of Big Data*, vol. 6, pp. 1-25, 2019.

[15] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," *International Journal of Computer Science and Network (IJCSN)*, vol. 2, no. 1, 2013.

[16] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Computing Surveys (CSUR)*, vol. 49, pp. 1-50, 2016.

[17] W. C. Sleeman IV and B. Krawczyk, "Bagging using instance-level difficulty for multi-class imbalanced big data classification on spark," *in IEEE international conference on big data (big data)*, 2019, pp. 2484-2493.Available: doi: 10.1109/BigData47090.2019.9006058.

[18] M. S. Santos, P. H. Abreu, N. Japkowicz, A. Fernández, C. Soares, S. Wilk, et al., "On the joint-effect of class imbalance and overlap: a critical review," *Artificial Intelligence Review*, pp. 1-69, 2022. Available: https://doi.org/10.1007/s10462-022-10150-3.

[19] E. Alfaro, M. Gamez, and N. Garcia, "Adabag: An R package for classification with boosting and bagging," *Journal of Statistical Software*, vol. 54, pp. 1-35, 2013.

[20] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," *in European conference on principles of data mining and knowledge discovery*, 2003.

[21] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, pp. 185-197, 2009.

[22] S. Ahmed, F. Rayhan, A. Mahbub, R. Jani, S. Shatabda, and D. M. Farid, "LIUBoost: locality informed under-boosting for imbalanced data classification," *in Emerging Technologies in Data Mining and Information Security, ed: Springer*, 2019.

[23] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS international transactions on computer science and engineering*, vol. 30, pp. 25-36, 2006.

[24] L. Chen, J. Jiang, and Y. Zhang, "HSDP: A Hybrid Sampling Method for Imbalanced Big Data Based on Data Partition," *Complexity*, vol. 2021, 2021.

[25] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Scientific Reports*, vol. 11, pp. 1-11, 2021.

[26] Y. Hu, D. Guo, Z. Fan, C. Dong, Q. Huang, S. Xie, et al., "An improved algorithm for imbalanced data and small sample size classification," *Journal of Data Analysis and Information Processing*, vol. 3, p. 27, 2015.

[27] T. Elhassan and M. Aljurf, "Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method," *Global J Technol Optim S*, vol. 1, 2016.

[28] M. Imran, V. S. Rao, T. Amarasimha, and S. Z. Quadri, "A novel technique on class imbalance big data using analogous over sampling approach," *International Journal of Computational Intelligence Research*, vol. 13, pp. 2407-2417, 2017.

[29] Z. Shi, "Improving k-nearest neighbors algorithm for imbalanced data classification," *in IOP Conference Series: Materials Science and Engineering*, 2020.

[30] S. Priya and A. Uthra, "RETRACTED ARTICLE: Comprehensive analysis for class imbalance data with concept drift using ensemble based classification," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 4943–4956, 2021.

[31] S. K. Hussin, S. M. Abdelmageid, A. Alkhalil, Y. M. Omar, M. I. Marie, and R. A. Ramadan, "Handling imbalance classification virtual screening big data using machine learning algorithms," *Complexity*, vol. 2021, 2021.

[32] M. J. Basgall, M. Naiouf, and A. Fernández, "FDR2-BD: A fast data reduction recommendation tool for tabular big data classification problems," *Electronics*, vol. 10, p. 1757, 2021.

[33] M. Koziarski, B. Kwolek, and B. Cyganek, "Convolutional neural network-based classification of histopathological images affected by data imbalance," *in Video Analytics. Face and Facial Expression Recognition, ed: Springer*, 2018.

[34] F. Wang, S. Liu, W. Ni, Z. Xu, Z. Qiu, Z. Wan, et al., "Imbalanced data classification algorithm with support vector machine kernel extensions," *Evolutionary Intelligence*, vol. 12, pp. 341-347, 2019.

[35] J. Jiang, H. Zhang, D. Pi, and C. Dai, "A novel multi-module neural network system for imbalanced heartbeats classification," *Expert Systems with Applications*, vol. 1, p. 100003, 2019.

[36] U. Aggarwal, A. Popescu, and C. Hudelot, "Minority Class Oriented Active Learning for Imbalanced Datasets," *in 25th International Conference on Pattern Recognition (ICPR), 2021*.

[37] A. Islam, S. B. Belhaouari, A. U. Rehman, and H. Bensmail, "KNNOR: An oversampling technique for imbalanced datasets," *Applied Soft Computing*, vol. 115, p. 108288, 2022.

[38] C. K. Aridas, S. Karlos, V. G. Kanas, N. Fazakis, and S. B. Kotsiantis, "Uncertainty based under-sampling for learning naive bayes classifiers under imbalanced data sets," *IEEE Access*, vol. 8, pp. 2122-2133, 2019.

[39] T. Das, A. Khan, and G. Saha, "Classification of imbalanced big data using SMOTE with rough random forest," *Int. J. Eng. Adv. Technol*, vol. 9, pp. 5174-5184, 2019.

[40] M. O. Miah, S. S. Khan, S. Shatabda, and D. M. Farid, "Improving detection accuracy for imbalanced network intrusion classification using cluster-based under-sampling with random forests," *in 1st international conference on advances in science, engineering and robotics technology (ICASERT)*, 2019.

[41] J. M. Johnson and T. M. Khoshgoftaar, "The effects of data sampling with deep learning and highly imbalanced big data," *Information Systems Frontiers*, vol. 22, pp. 1113-1131, 2020.

[42] UBAYA, Huda, JUAIRIAH, Ria Siti, "Performance of RUS and SMOTE Method on Twitter Spam Data Using Random Forest," *Journal of Physics: Conference Series. IOP Publishing*, p. 012130, 2020.

[43] E. Hassib, A. El-Desouky, L. Labib, and E.-S. M. El-Kenawy, "WOA+ BRNN: An imbalanced big data classification framework using Whale optimization and deep neural network," *Soft Computing*, vol. 24, pp. 5573-5592, 2020.

[44] V. Babar, "Classification of Imbalanced Data of Medical Diagnosis using Sampling Techniques," *Communications on Applied Electronics (CAE)*, vol. 7, no. 36, 2021.

[45] C.-Y. Lee, et al., "Hybrid Algorithm Based on Simulated Annealing and Bacterial Foraging Optimization for Mining Imbalanced Data," *Sensors and Materials*, vol. 33, pp. 1297-1312, 2021.

[46] KHANAM, Shapla, et al. "Towards an Effective Intrusion Detection Model Using Focal Loss Variational Autoencoder for Internet of Things (IoT)," *Sensors*, vol. 22, no. 15, 2022.

[47] S. Zhang, H. Lai, W. Chen, L. Zhang, X. Lin, and R. Xiao, "VBLSH: Volume-balancing locality-sensitive hashing algorithm for K-nearest neighbors search," *Information Sciences*, vol. 587, pp. 774-793, 2022.

[48] J. Dong and Q. Qian, "A Density-Based Random Forest for Imbalanced Data Classification," *Future Internet*, vol. 14, p. 90, 2022.

[49] RODRÍGUEZ-TORRES, Fredy; MARTÍNEZ-TRINIDAD, José F. CARRASCO-OCHOA, Jesús A., "An Oversampling Method for Class Imbalance Problems on Large Datasets," *Applied Sciences*, vol. 12, no. 7, 2022.

[50] AL-SAWWA, Jamil. ALMSEIDIN, Mohammad, "A Spark-Based Artificial Bee Colony Algorithm for Unbalanced Large Data Classification," *Information*, vol. 13, no. 11, 2022.

[51] CAMELE, Genaro, et al., "Statistical analysis of the performance of four Apache Spark ML Algorithms," *Journal of Computer Science & Technology*, vol. 22, no. 2, 2022.

[52] UCI Machine Learning Repository. [Online]

[53] Available: https://archive.ics.uci.edu/ml/datasets.php.

[54] KEEL-dataset repository. [Online] Available: https://sci2s.ugr.es/keel/datasets.php.

[55] Kaggle Machine Learning and data science community. [Online]

[56] Available: https://www.kaggle.com/datasets.

[57] M. M. Salih, M. A. Ahmed, B. Al-Bander, K. F. Hasan, M. L. Shuwandy, and Z. Al-Qaysi, “Benchmarking Framework for COVID-19 Classification Machine Learning Method Based on Fuzzy Decision by Opinion Score Method”, *Iraqi Journal of Science*, vol. 64, no. 2, pp. 922–943, Feb. 2023.

[58] M. J. Gatea and S. M. Hameed, “An Internet of Things Botnet Detection Model Using Regression Analysis and Linear Discrimination Analysis”, *Iraqi Journal of Science*, vol. 63, no. 10, pp. 4534–4546, Oct. 2022.