Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 11, No. 2, June 2023, pp. 503~514
ISSN: 2089-3272, DOI: 10.52549/ijeei.v11i2. 4471

Handling Imbalanced Data through Re-sampling: Systematic Review

Razan Yasir[1], Abdelrahman Elsharif Karrar[2], Waleed Ibrahim Osman[3], Moez Mutasim[4]
[1,3,4] College of Computer Studies, The National Ribat University, Sudan
[2]College of Computer Science and Engineering, Taibah University, Saudi Arabia

Article Info

Article history:

Received Jan 28, 2023
Revised Apr 29, 2023
Accepted Jun 17, 2023

Keyword:

Data Mining
Imbalance Data
Re-sampling
Over-sampling
Under-sampling
Hybrid Sampling
SMOTE

ABSTRACT

Handling imbalanced data is an important issue that can affect the validity and reliability of the results. One common approach to addressing this issue is through re-sampling the data. Re-sampling is a technique that allows researchers to balance the class distribution of their dataset by either over-sampling the minority class or under-sampling the majority class. Over-sampling involves adding more copies of the minority class examples to the dataset in order to balance out the class distribution. On the other hand, under-sampling involves removing some of the majority class examples from the dataset in order to balance out the class distribution. It's also common to combine both techniques, usually called hybrid sampling. It is important to note that re-sampling techniques can have an impact on the model's performance, and it is essential to evaluate the model using different evaluation metrics and to consider other techniques such as cost-sensitive learning and anomaly detection. In addition, it is important to keep in mind that increasing the sample size is always a good idea to improve the performance of the model. In this systematic review, we aim to provide an overview of existing methods for re-sampling imbalanced data. We will focus on methods that have been proposed in the literature and evaluate their effectiveness through a thorough examination of experimental results. The goal of this review is to provide practitioners with a comprehensive understanding of the different re-sampling methods available, as well as their strengths and weaknesses, to help them make informed decisions when dealing with imbalanced data.

Corresponding Author:

Razan Yasir Eltayeb,
Department of Information Technology,
National Ribat University
Nile Street, buri, Khartoum, Sudan.
Email: razanyasir946@gmail.com

1. INTRODUCTION

Knowledge Base (KB) is field which refers to a specially designed resource for gathering and processing knowledge in logical statement formats that define the relationship between graphical entities [1], and the field of machine learning which refers to the capability of a machine to imitate intelligent human behavior, those are the common fields in artificial Intelligence and both of them can have a class imbalanced data problem. The class imbalance problem become greatest issue in data mining and the imbalanced data appears in daily application [2] specially on the medical apps as [2] discussed and in academic apps like using the citation analysis by main path analysis (MPA) [3], this means that most of the fields are challenged by the imbalanced data problem, which comes from having too much or too little data in one class relative to another. This might also lead to complications like over-fitting, where a model becomes highly dependent on the training data and struggles to generalize new data. There are several strategies to deal with the imbalanced data problem, including altering data collection process, preprocess the data and impute the missing values to achieve better results [4], use employing new algorithms that are less sensitive to imbalance, and over-sampling or under-sampling data.

Numerous academics have made significant contributions in this field due to how challenging the problem of imbalanced data. In this section, we highlight relevant contributions made by various research studies. Various academics have investigated recurring issues, imbalanced class distributions in a variety of professions, and assessments of earlier studies based on the benchmark datasets they've employed. Additionally, they have discussed a comparison of various approaches and learning algorithms related to imbalanced data distributions issue. As illustrated in Figure 1, the imbalanced data problem can manifest in a variety of ways, such as having too much data in one class and too little in another.

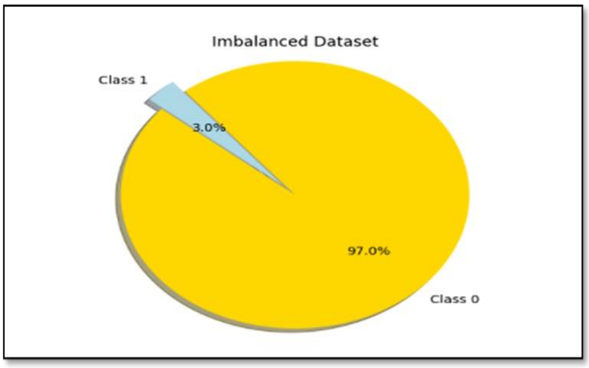

Figure 1. Illustration of Imbalanced Data [5]

2. METHODS FOR IMBALANCE HANDLING AT THE DATA LEVEL

A dataset is definitely imbalanced when there is a significant imbalance in the class distribution, such as a ratio of 1:100 or 1:1000 samples from the minority class to the majority class.This bias in the training dataset could have an impact on many machine learning algorithms, while other algorithms could totally ignore the minority class. This is significant, because the minority class generally relies more heavily on prediction.

2.1. Re-sampling techniques

In many real-world supervised learning problems, an available dataset is imbalanced because it contains a significantly lower number of instances of one class than another. In general, it is necessary to correctly predict the minor class, but due to the limited information available, this can be challenging. One solution to this class imbalance issue is to re-sample the training dataset. The two main approaches for re-sampling an imbalanced dataset are under-sampling, which means removing examples from the majority class, and over-sampling, which requires duplicating examples from the minority class. However, there are several re-sampling techniques, making it challenging to pick the best one. As depicted in Figure 2, the two main approaches for re-sampling an imbalanced dataset are under-sampling and over-sampling.

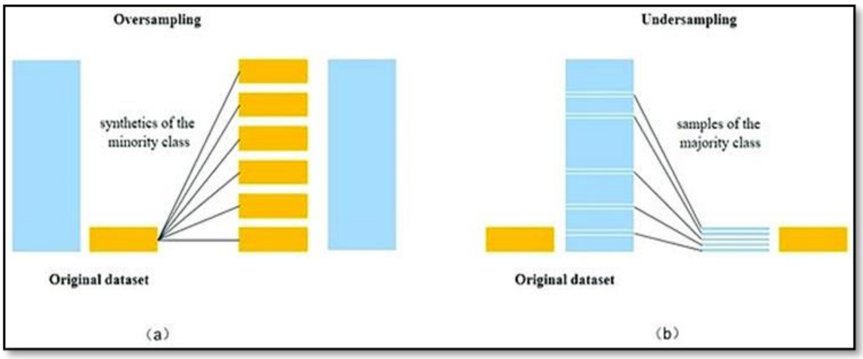

Figure 2. Illustration of Over-sampling and Under-sampling [6]

2.1.1. Re-sampling techniques classification based

Under-sampling: When the volume of the majority class is lowered, information is lost and potentially overbroad rules are generated. Despite its low computational cost, the under-sampling strategy is commonly utilized alongside clustering algorithms to tackle this problem. K-means is a partitional approach that is well known and widely used, and it offers a general way to properly cluster data. K-means was employed in [7] to tackle the problem of a regular clustering center that was susceptible to disruption and falls into the optimal solution locally. In a separate article, K-means [8] has been used to introduce an alternate method of assigning a value to a cluster, enhancing the quality of unbalanced data in the algorithm's outputs. The DBSCAN algorithm is another well-known algorithm to balance the training set, this algorithm selects the majority class samples that are most appropriate and completely ignores other majority class samples. It was utilized in [9] to provide an information-theoretic-based unsupervised gene selection technique. The majority class examples that are appropriate for the offered novel on [10] are picked, and unwanted majority class samples are eliminated to balance the training set. All of the data points were divided into different levels based on their local density, and on [11], the DBSCAN identified the lawmakers among all of the data points and constructed initial clusters based on the identified representatives for further analysis of the clusters' properties. This led to the suggestion of a unified clustering framework without imbalanced data. An application of the fuzzy principle is made possible by the fuzzy algorithms used in the cluster field to determine sample values during the undersampling phase. Furthermore, [12] provides many parallelization solutions for clustering algorithms for both Intel and NVidia-based architectures by focusing on three frequently used fuzzy clustering approaches, including FCM, GK-FCM, and FM. Lastly, based on a fuzzy concept, [13] discussed a clustering technique that allows for simplicity in data balancing.

Some techniques, such as the nearest neighbor balanced dataset, concentrate on circumstances near the decision boundary. that was adapted by [22] to present a quantum algorithm for data classification that breaks up larger groups of data belonging to the same class into smaller groups with sub-labels to help create boundaries between data with different labels and then creates a quantum circuit for classification that has multiple control gates. The under-sampling method is combined with the C4.5 classification model in [23].

Over-sampling: To achieve a balanced class distribution, approaches that generate fictional data are more flexible than those that modify the machine learning model. Any predictor can be used with class-imbalanced datasets according to these over-sampling techniques, which also alter the training data. For this purpose, a number of algorithms have been presented, although most of them are complex and frequently create additional noise. The over-sampling algorithm K-Means is applied for class-imbalanced data. It helps grouping by generating minority-class samples in appropriate and significant areas of the input space. The method avoids generating noise while actually solving imbalances between and within classes. When compared to standard K-means and other K-means-like algorithms, the cosine similarity measure in [14] is applied to assign data points to clusters and enhance the accuracy for certain datasets. When the dataset is uneven, the K-cosine-means algorithm fails. One further survey [15] mentioned a few evolutionary algorithms and dividing calculations. According to the correlation of those algorithms, the primary expert of k-implies is less expensive to evaluate, but a con is that it is more responsive to noisy data and outliers than fuzzy k-means and k-medoids. According to the distribution of the particulate matter values, the K-means clustering algorithm in [16] separates the area of the center of the fine dust distribution and then determines the coordinates of the optimal point. With regard to the DBSCAN algorithm, [17] offers a new grid-based algorithm, GBCN, that uses input parameters in clusters. The first parameter is determined automatically and computes the distance between each element of a dataset and its k nearest neighbors. The second parameter sets the minimum number of elements in clusters at five to provide high efficiency. A novel density-based clustering algorithm called ADBSCAN, where "A" stands for "adaptive," was developed to balance the dataset [18]. This algorithm offers a new way to identify the dataset's local high-density samples, and it can effectively adapt to datasets with significant variations in density without using any user-inputted variables. By clustering the dataset, granting each element a membership value, and finding the initial cluster center, the fuzzy C-means clustering algorithm described in [19] reduces the time complexity. The goal of the researchers in [20] and [21] is to offer several clustering algorithms and explain how they behave with imbalanced data, are frequently applied to particular tasks, and are not general. Table 1 presents a summary of previous studies that have employed cluster-based re-sampling techniques, the table includes information on the specific clustering method used, the dataset(s) analyzed, and the main aim of the studies.

In an attempt to balance out the medical data in 2021, [24] paired the c4.5 algorithm with a cluster-based over-sampling technique. High classification accuracy, quick computation times, and understandable classification rules are all benefits of the C4.5 decision tree algorithm. An unique method for generating artificial data points while keeping in mind the distribution of the data and the k nearest neighbors is the K Nearest Neighbor Oversampling (KNNOR) algorithm. The KNNOR method has exceeded the most powerful augmentation methods by enabling classifiers to obtain significantly higher accuracy after adding artificial minority data points to imbalanced datasets. A method to enhance the performance of kNN on imbalanced data has been proposed in [25] and has shown promise in its ability to handle imbalanced datasets without compromising model size. This method uses GA to optimize both feature weights and class weights. On [26], a different methodology is offered that makes use of CNN to automatically extract useful features of various fault classes from datasets with highly skewed data and then adopts data augmentation techniques to solve the problem. Imbalanced data classification using the Random Forest Classification (RFC) technique has become extremely popular in the modern application period since the beginning of 2020. Doing so enables [27] to use random forest (RF) as a classifier for imbalanced data after attempting to reduce outliers using isolation forest (iForest) and increasing the number of instances in the dataset in a balanced way. According to Table 2, many previous studies have utilized re-sampling techniques in combination with classification-based approaches to improve the performance of their algorithms.

Table 1. Re-sampling Techniques Cluster-Based

Author(s)	Aim of study	Imbalanced data	Applied on	Methods
Rahmanian, M. and Mansoori, E.G. [9]	propose an unsupervised gene selection scheme based on information theoretic measures	yes	Bench mark microarray gene expression dataset	A similarity-based algorithm, density-based clustering methods
Wang, Y. and Wang, D. [11]	propose an effective variational clustering algorithm called VDPC	No	20 benchmark datasets	DBSCAN and DPC
Oh, Yoosoo and Min, Seonghee. [16]	performed the K-means clustering algorithm to cluster feature datasets	No	air pollution stations in the korean dataset	K-means
Kume, A. and Walker, S.G. [8]	introduce an alternative way to assign a value to a cluster	No	sequential manner dataset	k–means
Starczewski, A. and Scherer, M.M. [15]	proposed a new grid-based algorithm GBCN	Yes	several 2-dimensional datasets	DBSCAN algorithm, new grid-based algorithm
Rohan Mittal [19]	explains the fuzzy C-means clustering algorithm	No	image segmentation, geo cover image capture dataset	fuzzy C-means
Khan, M.K. and Ahmed, S.M. [14]	using the cosine similarity measure to assign data points to clusters	Yes	homogeneous datasets, namely the Iris and Seeds dataset	standard K-means and other K-means-like algorithms
Cebrian, J.M. and Imbernón, B. [12]	introduces several parallelization strategies for clustering algorithms for both Intel and Nvidia-based architectures	Yes	different types of datasets	fuzzy clustering techniques such as FCM, GK-FCM, and FM
Lukauskas, M. and Ruzgas, T. [20]	introduce different clustering algorithms and present how these clustering algorithms work	Yes	distinguishing datasets	UNIC, k-Medoids (PAM), Gaussian Mixture, TCLUST, Trimmed k-means, Spectral Clustering, Density-Based Spatial Clustering, MULIC, DENCLUE, SOMs (NeuralNet), SVM, HIERDENC, deeply embedded clustering, etc
Patibandla, R.L. and Veeranjaneyulu, N. [15]	outlined a few dividing calculations and Evolutionary Algorithms	Yes	different info datasets	k-implies, k-medoids, Fuzzy k-means, Expectation Maximization, etc
Jian, S., Li, D. and Yu, Y. [7]	proposed improved k-means clustering algorithm	Yes	extensive data of New York City taxis	improved K-means
Lukauskas, Mantas & Ruzgas, Tomas. [22]	introduce different clustering algorithms and present how these clustering algorithms work	Yes	spherical in shape datasets	UNIC, k-Medoids (PAM), Gaussian Mixture, TCLUST, Trimmed k-means, Spectral Clustering, Density-Based Spatial Clustering, MULIC, DENCLUE, SOMs (NeuralNet), SVM, HIERDENC, Deep embedded clustering, etc.
Pratiwi, N.B.I. and Saputro, D.R.S. [13]	discussed the clustering algorithm with an approximation based on a fuzzy concept, which grants simplicity in the process of constructing clusters on datasets	No	non-linear dataset	fuzzy c-means and Fuzzy C-Shells
Mirzaei, B., Nikpour, B. and Nezamabadi-Pour, H. [10]	presented a novel and effective under-sampling technique is presented to select the suitable samples of the majority class	Yes	over fifteen imbalanced datasets	DBSCAN algorithm
Li, H., Liu, X., Li, T. and Gan, R. [18]	Propose a novel density-based clustering algorithm. After using the density estimator to filter noise samples	No	several artificial and real-world datasets	ADBSCAN algorithm

Table 2. Re-sampling Techniques Classification-Based

Author(s)	Aim of study	Imbalanced data	Applied on	Methods
Li, J. and Kais, S. [22]	presented a quantum algorithm for data classification based on the nearest-neighbor learning algorithm	No	trained experimental dataset	nearest-neighbor
Xu, Z. and Shen, D. [24]	proposed a cluster-based oversampling algorithm (KNSMOTE)	Yes	imbalanced medical dataset	Synthetic minority oversampling technique (SMOTE) and k-means algorithm
Ijaz, M.F., Attique, M. and Son, Y. [27]	proposes a cervical cancer prediction model (CCPM) that offers an early prediction of cervical cancer using risk factors as inputs	Yes	foremost prevailing cancer in females dataset	random forest (RF)
Saqlain, M., Abbas, Q. and Lee, J.Y. [26]	proposed a deep learning-based CNN-WDI model to classify wafer map defects in the semiconductor fabrication process	No	real wafer map dataset	CNN algorithm
Nugraha, W., Maulana, M.S. and Sasongko, A. [23]	using the undersampling strategy with clustering techniques and classification model	Yes	ten imbalance datasets from KEEL-repository	undersampling and c4.5
Shih, Y.H. and Ting, C.K. [25]	proposed a method to improve the performance of kNN by developing a new algorithm on imbalanced datasets and using it to optimize both feature and class weights	Yes	up-sampling dataset	FCWkNN and KNN algorithm

2.1.2. Synthetic Minority Over-sampling Technique (SMOTE)

One solution for imbalanced data is to over-sample the minority class. The simplest approach is to replicate examples from the minority class, although these examples don't add any better knowledge to the model. Instead, by combining the current examples, new ones can be created. For the minority class, data augmentation techniques like SMOTE are used. To further illustrate this technique, refer to Figure 3 which visually demonstrates the SMOTE process, and tell us how SMOTE generates new synthetic examples of the minority class by interpolating between existing examples in the minority class.

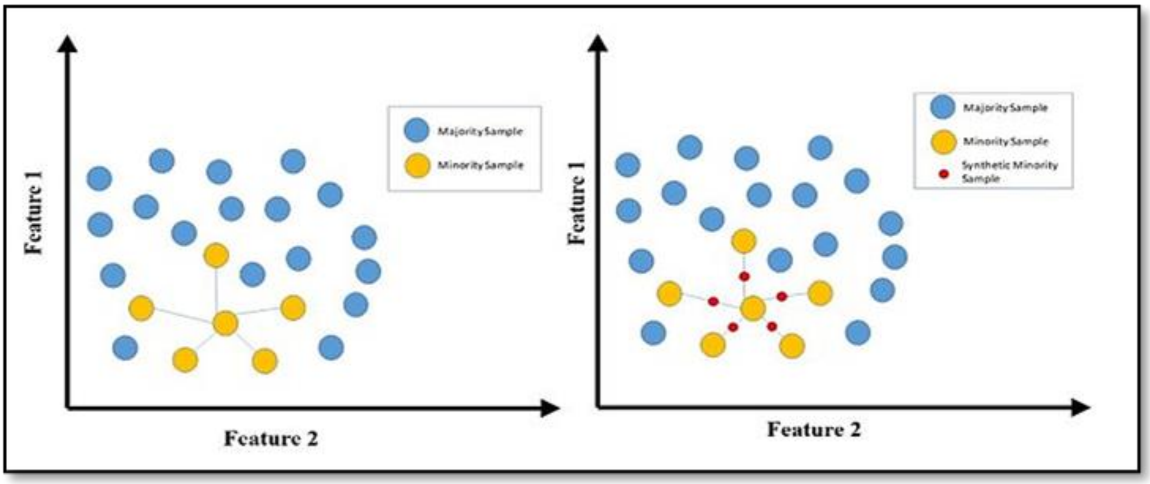

Figure 3. SMOTE Illustration [28]

There are various studies, especially on SMOTE and extreme learning machine algorithms, including [29] and [30]. The [31] develop a Majority Weighted Minority Over-sampling Technique (MOT), which creates synthetic samples from the weighted information samples using the clustering method. Utilizing SMOTE as a superior over-sampling technique with K-means to overcome the imbalanced learning problem [32]. K-means SMOTE successfully addresses data imbalance while enhancing the quality of freshly produced artificial data and reducing the production of noisy data. Table 3 presents a summary of previous studies that have applied re-sampling techniques using SMOTE method, and it includes information about authors, aim of study and the dataset used.

Table 3. SMOTE

Author(s)	Aim of study	Imbalanced data	Applied on	Technique	Methods
Goyal, A., Rathore, L. and Kumar, S. [29]	Using SMOTE to ensemble learning algorithms is a more efficient classifier in classifying imbalances	Yes	highly imbalanced datasets	Classification	SMOTE method for over-sampling
Tian, C., Zhou, L., Zhang, S., and Zhao, Y. [31]	present a new weighted minority oversampling technique for the classification of imbalanced datasets	Yes	generate synthetic samples dataset	Classification	Weighted Minority Oversampling Technique (MWMOTE)
Saah, D. and Tenneson, K. [32]	addressing both (between and within)-class, while effectively overcoming data imbalance	Yes	seven remote sensing benchmark datasets	Clustering	K-means and SMOTE algorithm
Last, F., Douzas, G. and Bacao, F. [30]	presents a simple and effective oversampling method	Yes	90 benchmark datasets	Clustering	K-means and SMOTE algorithm

2.2. Hybrid-Sampling:

Hybrid-sampling is a technique that combines multiple over-sampling and/or under-sampling methods to address class imbalance in a dataset. The idea is to use the strengths of different methods to overcome the limitations of a single approach. Figure 4 illustrates how hybrid-sampling is used to combine multiple over-sampling and under-sampling methods.

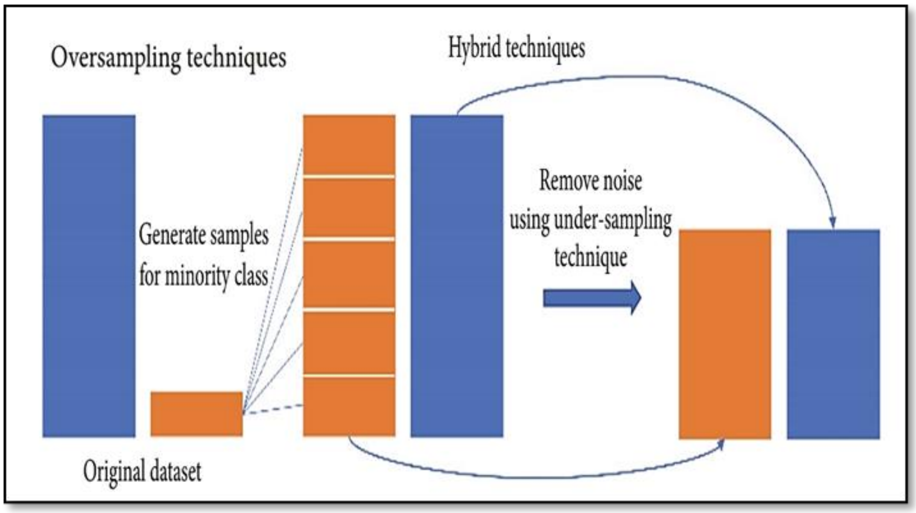
Oversampling techniques
Hybrid techniques
Generate samples for minority class
Remove noise using under-sampling technique
Original dataset

Figure 4. Illustration of hybrid-sampling [33]

Hybrid methods, on the other hand, include both under-sampling and over-sampling strategies. A novel cluster ensemble approach (CES) was developed by [34], which generates a single partition in the majority function and may cause compatibility problems with the cluster ensemble architecture's functionality. In order to improve predictions in binary imbalanced classification issues by the end of 2020, the proposed study [35] attempted to create a distribution-free super ensemble classifier, which is a hybridization of HDDT and RBFN models.

Researchers in [36] uses ensemble classifiers and the K-means clustering algorithm to deal with various under-sampling or over-sampling approaches. From a side perspective, researchers [37] provide a method that based on the K-means clustering algorithm to help in the selection of the dataset's important features, and the K-nearest neighbor (KNN) is then used to assess the classification performance in the experiments. A year earlier, [38] optimized the K-means clustering algorithm on the Hadoop platform and developed a first cluster center selection technique based on the basic ideas of the Hash algorithm.

DBSCN and K-mean cluster analysis are currently the two most popular and simple-to-use algorithms for clustering and classification-based analysis of data [39] because they can handle the majority of cases effectively when the data has a lot of randomness and no obvious set to use as a parameter, as in the case of linear or logistic regression algorithms.

Researchers in [40] and introduce a model to enhanced hybrid bag-boost model with a proposed re-sampling technique. This model includes a proposed re-sampling method that uses edited nearest neighbor (ENN) under-sampling to reduce noise and K-Means SMOTE as an over-sampling method.

By combining distance with a mean-based re-sampling strategy, random forest was employed by [41] to suggest an over-sampling method and deploy synthetic samples for the minority class.

Table 4 provides a comprehensive overview of previous studies that have used the hybrid-sampling technique, and it highlights the effectiveness of this method in addressing class imbalance and improving the performance of classifiers.

Table 4. hybrid sampling

Author(s)	Aim of study	Imbalanced data	Applied on	Technique	Methods
Bajal, E. And Katara, V. [39]	Used easily implementable algorithm for clustering	Yes	a renal adenocarcinoma dataset	Clustering	DBSCAN and K-mean
Puri, A. and Kumar Gupta, M. [40]	improved hybrid bag boost with the proposed re-sampling technique model	Yes	binary imbalanced dataset	Clustering	K-Means SMOTE (Synthetic Minority Oversampling Technique), edited the nearest neighbor (ENN)
Khan, T. and Tian, W. [34]	proposed a new cluster ensemble method (CES)	Yes	ten real-world benchmark datasets	Clustering	cluster ensemble method (CES)
GÜLDAL, S. [41]	reduce the imbalanced ratio	Yes	Different collected datasets	Classification	Random Forest (RF) and Support Vector Machine (SVM) algorithms
Shahabadi, M.S.E. and Tabrizchi, H. [36]	propose a novel clustering-based undersampling method to create a balanced dataset	Yes	44 benchmark datasets from the KEEL repository	Clustering	k-means
Chakraborty, T. and Chakraborty, A.K. [35]	proposed a novel distribution-free super ensemble classifier which is a hybrid model for improving predictions in binary imbalanced classification problems	Yes	small or medium-sized datasets	Classification	hybridization of HDDT and RBFN models
Moslehi, F. and Haeri, A. [37]	present a novel and integrated technique to select practical features in the dataset and remove the non-relevant features	No	non-relevant features dataset	Clustering	K-means
Hou, X. [38]	optimized the K-means clustering algorithm on the Hadoop platform	No	Hadoop dataset	Clustering	K-means algorithm

2.3. Combined Techniques for Computational Ensemble

The previous section discussed how to handle imbalanced data by re-sampling the original data to create classes that are balanced. The combination of techniques that manage imbalanced datasets and improve performance would be investigated in this section.

The Markov Clustering (MCL2) approach was determined to be the best suitable one for the dataset in order to reproduce local identical-by-descent (IBD) graphs. [42] applies the Mean Shift Clustering algorithm in a framework for detecting malicious actions and attack patterns in cyberspace. According to an adaptive threshold, the new overlap-based under-sampling (OBU) approaches were introduced on [43] to exclude negative instances from the overlapping region. Over fifty of them are available right away through the widely utilized Fundamental Clustering Problems Suite (FCPS) on [44]. By decreasing and balancing energy consumption, the study showed energy-efficient clustering method presented by [45] sought to increase the energy efficiency of WSNs. Individuals are grouped using the distributed clustering algorithm in [46] based on their social relationships and COVID-19 risk levels for serious disease. Lastly, [47] assessed, using a bagging approach, how the use of smaller pixels impacts lesion detection performance in general oncologic PET imaging.

Table 5 is a useful resource for researchers looking to improve the performance of their classifiers by using combined techniques for computational ensemble. It provides an overview of the various methods used in previous studies and the results obtained from their application.

Table 5. combined methods

Author(s)	Aim of study	Imbalanced data	Applied on	Methods
Shemirani, R. and Belbin, G.M. [48]	demonstrated the shortcomings of LFR the standard benchmark algorithm, in simulating local identical-by-descent (IBD) graphs	No	graphs dataset	Markov Clustering algorithm
Thrun, M.C. and Stier, Q. [44]	presented immediate access to over fifty fundamental clustering algorithms	No	arbitrary sample size dataset	CRAN in R, python
Insausti, X. and Zárraga-Rodríguez, M. [46]	presented a distributed clustering algorithm that groups individuals according to their social contacts and the risk level of severe illness	No	severe illness from the COVID-19 dataset	distributed consensus algorithm
Niranjana, R., Kumar, V.A. and Sheen, S. [42]	monitoring framework for deducing malicious activities and attack patterns in the cyberspace	No	Darknet traffic dataset	Mean Shift clustering algorithm
Vuttipittayamongkol, P. and Elyan, E. [43]	proposed new Overlap-based Undersampling (OBU) methods	Yes	well-known imbalanced datasets	Overlap-based Undersampling method
Lin, D. and Wang, Q. [45]	proposed a novel energy-efficient clustering algorithm that aimed at improving the energy efficiency of WSNs via reducing and balancing energy consumption	Yes	energy overhead dataset	dual-cluster-head mechanism
Morey, A.M., Noo, F. and Kadrmas, D.J. [47]	evaluated how the use of smaller pixels affects lesion-detection performance in general oncologic PET imaging	No	general PET cancer imaging dataset	

3. DISCUSSION

Re-sampling the training dataset is a common solution to address class imbalance in a dataset. There are three main techniques for re-sampling an imbalanced dataset, which are under-sampling, over-sampling, and a combination of both (hybrid approach). All the previous studies in this article have also used these techniques as well as other methods, as illustrated in figure 5.

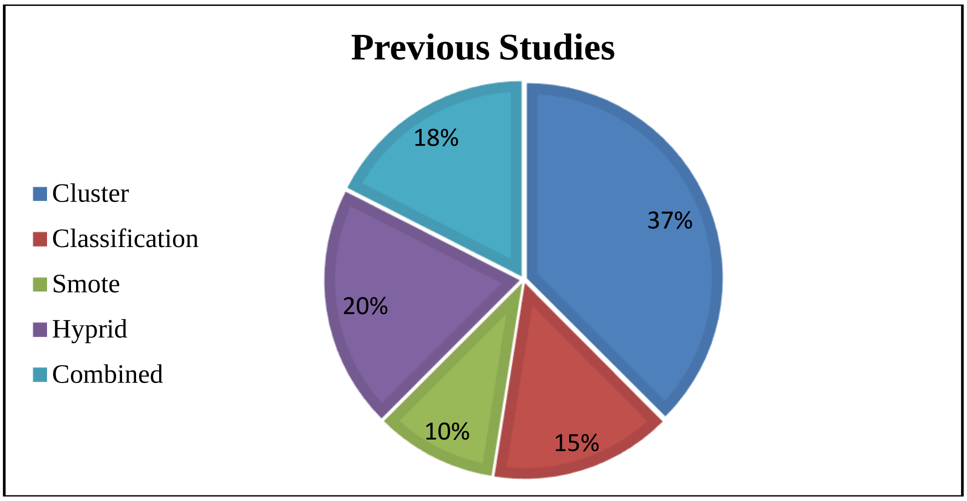

Figure 5. Complete papers review

From the cluster-based previous studies, results shows the analysis of cluster-based re-sampling studies has two main advantages over the use of generalized estimating equations. Firstly, there is no need to create a correlation structure as within-cluster re-sampling takes care of it implicitly. Secondly, generalized estimating equations may not be effective when dealing with non-ignorable clusters, whereas within-cluster based re-sampling is still appropriate. Additionally, the use of these techniques results in more accurate predictions as they create homogeneous groups with similar prognostic characteristics and corresponding survival patterns. Figure 6 provides an illustration of the cluster-based re-sampling techniques discussed in the paper.

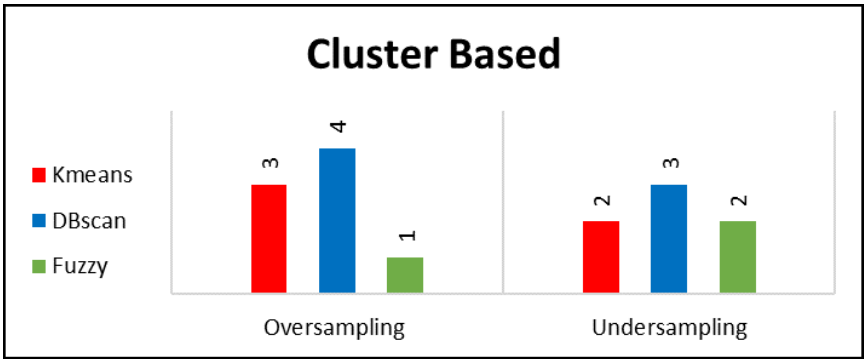

Figure 6. Cluster-Based Technique

Data imbalance can be addressed by adjusting the proportion of majority and minority classes in the data set through techniques such as undersampling the majority class or oversampling the minority class. Many researchers have attempted to address this issue through methods such as cost-sensitive classification and resampling techniques. These methods aim to improve the accuracy of identifying minority data by comparing the performance of different classification algorithms. As shown in figure 7, these techniques can effectively distinguish minority data from majority data.

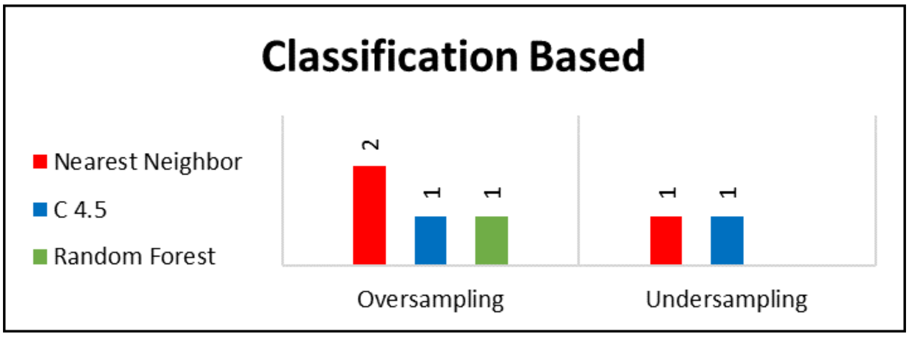

Figure 7. Classification Based Technique

Previous research suggests that using SMOTE (Synthetic Minority Over-sampling Technique) can increase recall for minority class predictions, but at the cost of decreased precision. This means that more minority class predictions will be made, but some may not be accurate. Studies have also found that SMOTE can lead to more class overlap and noise due to not considering neighboring samples from different classes. Additionally, SMOTE may not be as effective for data with high dimensions. However, it can reduce overfitting and there is no information loss when using it. It is also easy to understand and manipulate. Figure 8 illustrates the results of previous studies on SMOTE.

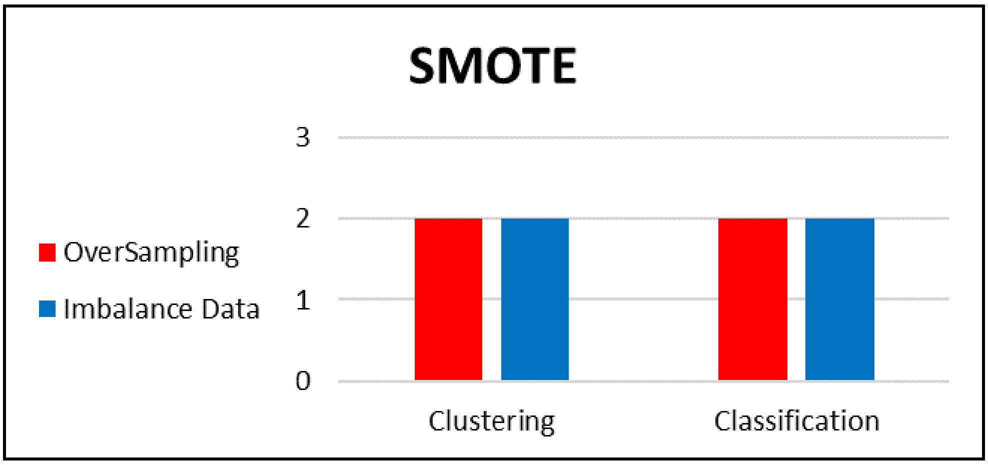

Figure 8. SMOTE Technique

Finally, figure 9 illustrate the hybrid approach. Taken together, all the studies presented previously in this section suggest a hybrid sampling approach that combines two sampling techniques to create a balanced dataset. Combining several sampling strategies enhances each method's strengths while reducing its drawbacks. As a result, the hybrid-sampling technique regularly outperforms the individual sampling techniques.

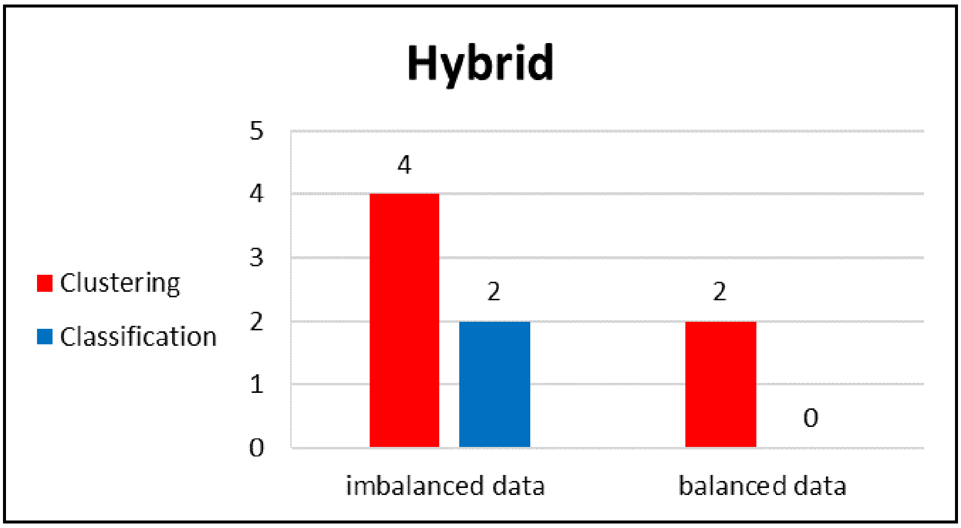

Figure 9. Hybrid-sampling Technique

4. CONCLUSION

Re-sampling is the suitable way for handling imbalanced data, according to all of the prior studies in this paper. The re-sampling and combined approaches can accomplish much more because they can provide the learning algorithm with new information or get rid of unnecessary information. On small, imbalanced datasets, the minority class appears to be poorly represented by an excessively reduced number of examples that may not be sufficient for learning, especially when a high degree of class overlap exists and the class is further categorized into subclusters. The above relationship between training set size and improper classification or clustering performance for imbalanced datasets appears to be causal. The impact of these complicated factors appears to be diminished for larger datasets as the minority class is better represented by a greater number of instances.

REFERENCES:

[1] A. E. Karrar, "A Proposed Model for Improving the Performance of Knowledge Bases in Real-World Applications by Extracting Semantic Information," International Journal of Advanced Computer Science and Applications, vol. 13, no. 2, 2022, doi: 10.14569/ijacsa.2022.0130214.

[2] A. E. Karrar, "Investigate the Ensemble Model by Intelligence Analysis to Improve the Accuracy of the Classification Data in the Diagnostic and Treatment Interventions for Prostate Cancer," International Journal of Advanced Computer Science and Applications, vol. 13, no. 1, 2022, doi: 10.14569/ijacsa.2022.0130122.

[3] M. Umair et al., "Main Path Analysis to Filter Unbiased Literature," Intelligent Automation & Soft Computing, vol. 32, no. 2, pp. 1179–1194, 2022, doi: 10.32604/iasc.2022.018952.

[4] A. E. Karrar, "The Effect of Using Data Pre-Processing by Imputations in Handling Missing Values," Indonesian Journal of Electrical Engineering and Informatics (IJEEI), vol. 10, no. 2, Apr. 2022, doi: 10.52549/ijeei.v10i2.3730.

[5] https://www.urbanstat.com/handlingimbalanceddatasets/?doing_wp_cron=1659547994.1132130622863769531250

[6] Xia, Wei & Ma, Caihong & Liu, Jianbo & Liu, Shibin & Chen, Fu & Zhi, Yang & Duan, Jianbo. (2019). High-Resolution Remote Sensing Imagery Classification of Imbalanced Data Using Multistage Sampling Method and Deep Neural Networks. Remote Sensing. 11. 2523. 10.3390/rs11212523.

[7] Jian, S., Li, D. and Yu, Y., 2021. 'Research on Taxi Operation Characteristics by Improved DBSCAN Density Clustering Algorithm and K-means Clustering Algorithm'. In Journal of Physics: Conference Series (Vol. 1952, No. 4, p. 042103). IOP Publishing.

[8] Kume, A. and Walker, S.G., 2021. 'The utility of clusters and a Hungarian clustering algorithm'. Plos one, 16(8), p.e0255174.

[9] Rahmanian, M. and Mansoori, E.G., 2022. 'An unsupervised gene selection method based on multivariate normalized mutual information of genes '. Chemometrics and Intelligent Laboratory Systems, 222, p.104512.

[10] Mirzaei, B., Nikpour, B. and Nezamabadi-Pour, H., 2020. 'An under-sampling technique for imbalanced data classification based on DBSCAN algorithm'. In 2020 8th Iranian Joint Congress on Fuzzy and intelligent Systems (CFIS) (pp. 21-26). IEEE.

[11] Wang, Y., Wang, D., Zhou, Y., Zhang, X. and Quek, C., 2021. 'VDPC: Variational Density Peak Clustering Algorithm'. arXiv preprint arXiv:2201.00641.

[12] Cebrian, J.M., Imbernón, B., Soto, J. and Cecilia, J.M., 2021. 'Evaluation of Clustering Algorithms on HPC Platforms'. Mathematics, 9(17), p.2156.

[13] Pratiwi, N.B.I. and Saputro, D.R.S., 2020, August. 'Fuzzy c-shells clustering algorithm'. In Journal of Physics: Conference Series (Vol. 1613, No. 1, p. 012006). IOP Publishing.

[14] M.K., Ahmed, S.M., Sarker, S. and Khan, M.H., 2021. 'K-Cosine-Medoids Clustering Algorithm'. In 2021 5th International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-5). IEEE.

[15] Patibandla, R.L. and Veeranjaneyulu, N., 2021. 'Clustering Algorithms: An Exploratory Review'.

[16] Oh, Yoosoo & Min, Seonghee. (2021). 'Practical Application Using the Clustering Algorithm'. 10.5772/intechopen. 99314.

[17] Starczewski, A., Scherer, M.M., Ksiazek, W., Debski, M. and Wang, L., 2021. 'A novel grid-based clustering algorithm'. Journal of Artificial Intelligence and Soft Computing Research, 11.

[18] Li, H., Liu, X., Li, T. and Gan, R., 2020. 'A novel density-based clustering algorithm using nearest neighbor graph'. Pattern Recognition, 102, p.107206.

[19] Mittal, Rohan. (2021). 'Fuzzy C-Means Clustering Algorithm'.

[20] Lukauskas, M. and Ruzgas, T., 2021. 'Analysis of clustering methods performance across multiple datasets'. In DAMSS 2021: 12th conference on data analysis methods for software systems, Druskininkai, Lithuania, December 2–4, 2021 (pp. 45-46). Vilnius university press.

[21] Lukauskas, Mantas & Ruzgas, Tomas. (2021). 'Comparative analysis of clustering algorithms for synthetic and real data'.

[22] Li, J. and Kais, S., 2021. 'A universal quantum circuit design for periodical functions'. New Journal of Physics, 23(10), p.103022.

[23] Nugraha, W., Maulana, M.S. and Sasongko, A., 2020. 'Clustering Based Undersampling for Handling Class Imbalance in C4.5 Classification Algorithm'. In Journal of Physics: Conference Series (Vol. 1641, No. 1, p. 012014). IOP Publishing.

[24] Xu, Z., Shen, D., Nie, T., Kou, Y., Yin, N. and Han, X., 2021. 'A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data'. Information Sciences, 572, pp.574-589.

[25] Shih, Y.H. and Ting, C.K., 2019. 'Evolutionary optimization on k-nearest neighbors classifier for imbalanced datasets'. In 2019 IEEE Congress on Evolutionary Computation (CEC) (pp. 3348-3355). IEEE.

[26] Saqlain, M., Abbas, Q. and Lee, J.Y., 2020. 'A deep convolutional neural network for wafer defect identification on an imbalanced dataset in semiconductor manufacturing processes'. IEEE Transactions on Semiconductor Manufacturing, 33(3), pp.436-444.

[27] Ijaz, M.F., Attique, M. and Son, Y., 2020. ‘Data-driven cervical cancer prediction model with outlier detection and over-sampling methods’. Sensors, 20(10), p.2809.

[28] Vijayvargiya, Ankit & Prakash, Chandra & Kumar, Rajesh & Bansal, Sanjeev & Tavares, Joao. (2021). Human Knee Abnormality Detection from Imbalanced sEMG Data. Biomedical Signal Processing and Control. 66. 10.1016/j.bspc.2021.102406

[29] Goyal, A., Rathore, L. and Kumar, S., 2021. ‘A Survey on Solution of Imbalanced Data Classification Problem Using SMOTE and Extreme Learning Machine’. In Communication and Intelligent Systems (pp. 31-44). Springer, Singapore.

[30] Last, F., Douzas, G. and Bacao, F., 2017. ‘Oversampling for imbalanced learning based on k-means and smote’. arXiv preprint arXiv:1711.00837.

[31] Tian, C., Zhou, L., Zhang, S. and Zhao, Y., 2020. ‘A new majority weighted minority oversampling technique for classification of imbalanced datasets’. In 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) (pp. 154-157). IEEE.

[32] Saah, D., Tenneson, K., Matin, M., Uddin, K., Cutter, P., Poortinga, A., Nguyen, Q.H., Patterson, M., Johnson, G., Markert, K. and Flores, A., 2019. ‘Land cover mapping in data scarce environments: challenges and opportunities’. Frontiers in Environmental Science, 7, p.150.

[33] Tuong Le, Minh Thanh Vo, Bay Vo, Mi Young Lee, Sung Wook Baik, "A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction", Complexity, vol. 2019, Article ID 8460934, 12 pages, 2019. https://doi.org/10.1155/2019/8460934

[34] Khan, T., Tian, W., Kadhim, M.R. and Buyya, R., 2021. ‘A Novel Cluster Ensemble based on a Single Clustering Algorithm’. In 2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS) (pp. 127-135). IEEE.

[35] Chakraborty, T. and Chakraborty, A.K., 2020. ‘Superensemble classifier for improving predictions in imbalanced datasets’. Communications in Statistics: Case Studies, Data Analysis and Applications, 6(2), pp.123-141.

[36] Shahabadi, M.S.E., Tabrizchi, H., Rafsanjani, M.K., Gupta, B.B. and Palmieri, F., 2021. ‘A combination of clustering-based under-sampling with ensemble methods for solving imbalanced class problem in intelligent systems’. Technological Forecasting and Social Change, 169, p.120796.

[37] Moslehi, F. and Haeri, A., 2020. ‘An evolutionary computation-based approach for feature selection’. Journal of Ambient Intelligence and Humanized Computing, 11(9), pp.3757-3769.

[38] Hou, X., 2019. ‘An improved k-means clustering algorithm based on hadoop platform’. In The International Conference on Cyber Security Intelligence and Analytics (pp. 1101-1109). Springer, Cham.

[39] Bajal, E., Katara, V., Bhatia, M. and Hooda, M., 2022. ‘A Review of Clustering Algorithms: Comparison of DBSCAN and K-mean with Oversampling and t-SNE’. Recent Patents on Engineering, 16(2), pp.17-31.

[40] Puri, A. and Kumar Gupta, M., 2022. ‘Improved hybrid bag-boost ensemble with K-means-SMOTE–ENN technique for handling noisy class imbalanced data.’ The Computer Journal, 65(1), pp.124-138.

[41] GÜLDAL, S., 2021. ‘Improving Machine Learning Performance of Imbalanced Data by Resampling: DBSCAN and Weighted Arithmetic Mean’. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi, 10(4), pp.1563-1574.

[42] Niranjana, R., Kumar, V.A. and Sheen, S., 2020. ‘Darknet traffic analysis and classification using numerical agm and mean shift clustering algorithm’. SN Computer Science, 1(1), pp.1-10.

[43] Vuttipittayamongkol, P. and Elyan, E., 2020. ‘Neighbourhood-based undersampling approach for handling imbalanced and overlapped data’. Information Sciences, 509, pp.47-70.

[44] Thrun, M.C. and Stier, Q., 2021. ‘Fundamental clustering algorithms suite’. SoftwareX, 13, p.100642.

[45] Lin, D. and Wang, Q., 2019. ‘An energy-efficient clustering algorithm combined game theory and dual-cluster-head mechanism for WSNs’. IEEE Access, 7, pp.49894-49905.

[46] Insausti, X., Zárraga-Rodríguez, M., Nolasco-Ferencikova, C. and Gutierrez-Gutierrez, J., 2021. ‘Distributed clustering algorithm for adaptive pandemic control’. IEEE Access, 9, pp.160688-160696.

[47] Morey, A.M., Noo, F. and Kadrmas, D.J., 2016. ‘Effect of using 2 mm voxels on observer performance for PET lesion detection’. IEEE transactions on nuclear science, 63(3), pp.1359-1366.

[48] Shemirani, R., Belbin, G.M., Burghardt, K., Lerman, K., Avery, C.L., Kenny, E.E., Gignoux, C.R. and Ambite, J., 2021. ‘Selecting Clustering Algorithms for IBD Mapping’. bioRxiv.