Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2023) 31: 1223 – 1236

doi:10.55730/1300-0632.4044

FuzzyCSampling: A Hybrid fuzzy c-means clustering sampling strategy for imbalanced datasets

Abdullah MARAŞ[1*], **Çiğdem SELÇUKCAN EROL**[1,2]
[1]Institute of Science, Division of Informatics, İstanbul University, İstanbul, Turkiye
[2]Informatics Department and Faculty of Science, Departments of Biology, Division of Botany, İstanbul University, İstanbul, Turkiye

Received: 22.09.2022 • **Accepted/Published Online:** 11.08.2023 • **Final Version:** 30.11.2023

Abstract: Classification model with imbalanced datasets is recently one of the most researched areas in machine learning applications since they induce to the emergence of low-performing machine learning models. The imbalanced datasets occur if target variables have an uneven number of examples in a dataset. The most prevalent solutions to imbalanced datasets can be categorized as data preprocessing, ensemble techniques, and cost-sensitive learning. In this article, we propose a new hybrid approach for binary classification, named FuzzyCSampling, which aims to increase model performance by ensembling fuzzy c-means clustering and data sampling solutions. This article compares the proposed approaches' results not only to the base model built on an imbalanced dataset but also to the previously presented state-of-the-art solutions undersampling, SMOTE oversampling, and Borderline Smote Oversampling. The model evaluation metrics for the comparison are accuracy, roc_auc score, precision, recall and F1-score. We evaluated the success of the brand-new proposed method on three different datasets having different imbalanced ratios and for three different machine learning algorithms (k-nearest neighbors algorithm, support vector machines and random forest). According to the experiments, FuzzyCSampling is an effective way to improve the model performance in the case of imbalanced datasets.

Key words: Binary classification, imbalanced datasets, machine learning, sampling, fuzzy c-means

1. Introduction

Advances in hardware technology, device processing capacity, and the resulting increase in the volume of stored raw data allow machine learning algorithms to be employed more effectively. Companies have started a race to adapt machine learning algorithms to their own businesses to remain ahead of their competitors. The widespread use of machine learning across industries has also accelerated the emergence of solutions to real issues that may constitute obstacles to the building of successful models. Classification with imbalanced datasets is one of the study areas that occurs when the number of examples belonging to one of the target variables dominates the other variable.

The main goal of classification is to look for a relationship between the input variables and target variables [1] . There are three types of classification, namely, binary classification, multiclass classification, and multi-label classification. Binary classification refers to the process of assigning items to one of two groups [2]. A wide range of sectors benefit from binary classification to tackle industry-specific problems such as spam detection, disease diagnosis, customer purchase behavior, and more. Multiclass classification arises in the case of multiple class

*Correspondence: abdullah.maras@gmail.com

labels for the target variables [3]. Some applications of multiclass classification are face classification, plant-animal species classification, and network intrusion detection systems. Multi-label classification, in addition to binary and multiclass classification, deals with data issues in which one or more class labels are projected for each data point [4].

The majority of conventional machine learning methods assume a balanced class distribution in the dataset [5]. In the case of imbalanced datasets, the algorithms are unable to accurately capture data distribution characteristics. This results in poor prediction performance of target variables. Working with imbalanced datasets is essential in various real-world machine-learning applications. Fraud detection [6], spam e-mail detection [7], disease diagnosis [8], and text classification [9] are all examples of areas that suffer from imbalanced datasets.

Machine learning algorithms will learn more from the majority class in the scenario of an imbalanced dataset which leads to results influenced by the majority class [10]. The learning process on imbalanced datasets can still provide good accuracy scores. A good accuracy score, on the other hand, is not always indicative of a good model. Since the model learned the pattern mostly from the majority class, the success of the minority class prediction may not be as good as the majority class. Intrinsic or extrinsic reasons for imbalanced datasets can arise [11]. The intrinsic imbalanced datasets emerge based on the nature of the problem. One typical example in the healthcare field is the separation of healthy people and patients with rare diseases. The dataset would have a skewed distribution since it is impossible to collect examples of rare diseases as much as you can collect healthy samples. The project time, improper data collection, and data storage limitations can exemplify the extrinsic imbalanced datasets. In cases where the imbalanced dataset problem is observed, other problems such as small disjuncts, class overlapping, and noisy examples can be seen at the same time [12]. Subproblems accompanying imbalanced datasets may require more complicated solutions.

The difficulty level of challenges in imbalanced datasets differs based on the distribution of the target variable. The Imbalance Ratio (IR) is a statistic that compares the complexity levels of various datasets by dividing the number of negative class instances by the number of positive class examples [13]. Researchers publish their proposed methods with the imbalance ratio of datasets. There are various proposed strategies to cope with imbalanced datasets given the diverse nature of datasets in different industries and the fact this has been a highly studied topic in recent years. Data preprocessing, ensemble techniques, and cost-sensitive learning are the three main divisions to summarize these solutions [14]. Data preprocessing solutions, also known as data-level solutions, strive to balance the distribution of target classes. The fundamental benefit of these methods is that they are unaffected by the chosen classifier [14]. Undersampling and oversampling are the two most prevalent data preprocessing methodologies employed by practitioners and scholars. Undersampling techniques attempt to balance datasets by removing instances, whereas oversampling approaches attempt to balance datasets by replication of existing instances or generation of new instances from current ones. However, both undersampling and oversampling methods have their own set of challenges so researchers come up with hybrid solutions that utilize both undersampling and oversampling techniques [15]. Ensemble techniques might leverage either cost-sensitive learning or data preprocessing approaches to deal with imbalanced datasets [14, 16]. Ensemble methods can also benefit from an ensemble classifier that is a combination of several classifiers to increase model performance. The final set of solutions addressing imbalanced datasets is cost-sensitive learning that increases the model performance by evaluating misclassification costs [17]. Practitioners and scholars mostly pay more attention to minority classes in imbalanced datasets. As a result, the minority class is critical in determining the cost of misclassification and the penalties.

Recently, researchers have been drawn to ensemble strategies in order to improve model performance and mitigate the downsides of undersampling and oversampling approaches. As the demand to work with imbalanced data sets has grown, so has the number of proposed solutions in this field.

The remainder of the work is structured as follows. Section 2 introduces the related work, and Section 3 contains the suggested strategy's technical background. Section 4 presents our FuzzyCSampling methodology, and Section 5 provides experimental results from dataset selection to evaluation results. Finally, Section 6 concludes the article with an overview.

2. Related work

Numerous publications are focused on the imbalanced dataset issue. Scientists are attempting to fix this matter by presenting datasets to algorithms in a balanced manner. The problem first captured great attention in the early 2000s. Many topics from problem definition to solutions are introduced in this decade, resulting in a new era of field progression [18, 19]. Especially there is a considerable amount of work on sampling techniques that can be grouped under data preprocessing solutions. However, among these approaches, the data extracted from the dataset causes information loss in the undersampling techniques, and duplication of data from the minority class generates overfitting in the oversampling approach [20]. Chawla et al. [21] introduced Synthetic Minority Over-sampling Technique (SMOTE) to overcome the disadvantage of the oversampling approach. The SMOTE approach generates samples from the minority class synthetically. Minority class samples are replicated repeatedly in the oversampling approach, whereas fresh samples are created in the SMOTE method. Han et al. proposed borderline-SMOTE1 and borderline-SMOTE2, approaches that only oversample the minority cases around the boundary between the majority target variable and minority target variable [22]. Liu et al. [23] study on the drawbacks of undersampling technique resulted in two new approaches: EasyEnsemble and BalanceCascade. EasyEnsemble takes a subset of the major class, trains a learner with each of them, and then combines their outputs. BalanceCascade is built similarly, although, unlike EasyEnsemble, it excludes from further examination accurately categorized major class instances of trained learners [23].

The hybrid approaches are the subject of more recent studies. Researchers are more likely to combine several solutions for imbalanced data to boost the explanatory power of the model. Ramentol et al. put forth a brand-new hybrid approach that combines the synthetic minority oversampling technique with an editing method based on rough set theory to improve the quality of produced synthetic data [24]. SMOTEBoost [25] is developed to improve SMOTE that generates synthetic instances from the uncommon or minority class by Chawla et al. SMOTEBoost combines the SMOTE and the boosting procedure to modify and update the weights and correct for skewed distributions. Another combination of sampling and the boosting algorithm is RUSBoost [26]. RUSBoost has a similar approach to the SMOTEBoost. Unlike the SMOTEBoost, the RUSBoost benefits from random undersampling to balance the datasets. The comparison between SMOTEBoost and RUSBoost shows that RUSBoost is more straightforward and faster than SMOTEBoost [26]. EHSO is introduced as a result of overlapping areas significance that employs a combination of undersampling and oversampling procedures [27]. EHSO's primary goal is to increase the visibility of the decision border by eliminating pointless majority class samples. EHSO utilizes evolutionary undersampling [28] to determine the best trade-off between classification performance and the replication ratio of random oversampling. S. Garc´ıa and F. Herrera [28] investigate the model performance when different fitness functions are used for evolutionary undersampling to balance the dataset. EUSBoost [29] is also an evolutionary undersampling-based approach to improve RUSBoost performance proposed by Galar et al. They considered employing evolutionary undersampling rather than

random undersampling. Zhang et al. introduce an upgraded version of EUSBoost, called EHSBoost, which makes use of the strengths of both undersampling and oversampling techniques [30].

Ensemble tactics, which combine sampling techniques with a classification algorithm, are another extensively used tool for dealing with imbalanced datasets. To overcome the drawback of undersampling techniques, a rotation forest classifier was employed in conjunction with an undersampling strategy by H. Guo, X. Diao, and H. Liu [31]. They profited from rotation forest since it is more responsive to sampling strategy and generation of individual classifiers is easier. Sun et al. focused on enhancing the number of synthetic samples produced around the minority target variable by combining SMOTE and Adaboost support vector machine [32]. SMOTE was also ensemble with a decision tree by taking into account varied sampling rates to strengthen the variation of the base classifiers [33]. MEBoost is also introduced as a result of being influenced by ideas such as RUSBoost and SMOTEBoost [34]. MEBoost differs from comparable algorithms in that it uses multiple learners, which are decision trees and additional tree classifiers.

Fuzzy approaches are also implemented to increase the model performances. H. Patel and G.S. Thakur proposed a hybrid fuzzy weighted nearest neighbor to deal with imbalanced datasets problem [35]. They also applied adaptive K-nearest neighbor to improve fuzzy K-nearest neighbor classification [36, 37]. H. Patel, D.S. Rajput, O.P. Stan, and L.C. Miclea proposed an updated version of the fuzzy adaptive nearest neighbor algorithm which aims to use optimal weights [38].

The explosive growth of research in the discipline has also resulted in literature reviews to keep academics up to date and track progress. A bibliometric analysis has been conducted to present field progression, collaboration of authors and countries, and the most cited publications [39]. The bibliometric analysis findings show that SMOTE and variations of it are still among the most implemented solutions for imbalanced datasets. In reviews, the effectiveness of the ensemble of sampling and bagging procedures is highlighted [40]. Galar et al. presented an ensemble taxonomy and performed an empirical comparative analysis of the most commonly used ensemble strategies [41]. Liu and Zhou also conducted experiments on different ensemble approaches and compared the results [42]. They also emphasized the lack of research on imbalanced datasets for multiclass classification problems. Patal et al. review imbalanced datasets problem for wireless sensor networks [43].

According to literature surveys, there is a definite tendency toward balancing the imbalanced datasets before training the classification models. It is evident that the distribution of target variables has a substantial impact on model performance. Recent researches have concentrated on ensemble solutions. In this study, we have also presented an ensemble approach that combines fuzzy c-means clustering and sampling approaches to increase model performances by reducing information loss. In the next section, we present the technical background that we benefitted from.

3. Technical background

This section outlines the components of the proposed FuzzyCSampling strategy, as well as the technical background information. Although sampling procedures are effective ways to deal with imbalanced datasets, they also have drawbacks that degrade model performance. The proposed strategy applies fuzzy clustering before the sampling strategies to balance the datasets by reducing these disadvantages. The following subsections summarize the utilized algorithms that benefited and the widely used state-of-the-art solutions for comparison purposes.

3.1. Fuzzy c-means clustering

Clustering algorithms are used to generate groups from a dataset based on similarities. The K-means algorithm, the simplest and widely used clustering algorithm, aims to partition a dataset into k number of groups where k is specified explicitly [44]. Clustering algorithms are divided into two main categories as hard and soft clustering. In hard clustering, each data point is assigned to just one cluster, but in soft clustering, a data point can be assigned to numerous clusters with its probability. The Fuzzy C-means clustering (FCM) used in our suggested technique is simply a soft clustering method [45].

3.2. Sampling methodologies

The sampling strategies are one of the key approaches to dealing with imbalanced datasets at the data level. The undersampling and the oversampling are prominent examples of sampling strategies. Given the fact that the machine learning algorithms are expected to be trained with balanced datasets to observe better results, the imbalanced ratio, and size of the majority and minority class variables also affect the model performance [39].

3.2.1. Undersampling

The undersampling approaches attempt to balance the datasets by removing instances from the majority target variable. The random undersampling is the simplest and fastest implementation that randomly eliminates the examples. However, removing the data from datasets can also result in information loss which is the main drawback of undersampling approaches. To address the shortcomings, several enhanced undersampling procedures, such as the ones shown above, are proposed [26, 29, 46, 47].

3.2.2. Oversampling

The oversampling approach aims to balance the dataset by replicating minority class examples in contrast to undersampling. The random oversampling is also the easiest implementation that duplicates examples in the minority class randomly. Oversampling, like undersampling, has significant disadvantages. The overfitting issue can be observed from the usage of oversampling.

SMOTE is one of the main versions of oversampling to obtain better model performance. SMOTE produces synthetic points in order to enhance the number of examples of minority target variables [21]. Minority target variables and the k-nearest neighbours method are used to produce synthetic data points. In comparison to large datasets, SMOTE can produce good results with smaller datasets. The creation of synthetic data is also affected by its size in terms of time.

Borderline-SMOTE was developed based on SMOTE [22]. There are additionally two Borderline-SMOTE variants, with the primary distinction being the oversampling of the majority target examples. Borderline-SMOTE1 oversamples both the majority and minority target variables in the decision border, whereas Borderline-SMOTE2 only oversamples the minority target variables.

3.3. Classification algorithms

Classification problems are a subfield of supervised machine learning algorithms. Supervised learning algorithms are trained by features of each data point in a dataset and its predefined labels. The three most widely used algorithms are chosen for experimental purposes. K-nearest neighbors (k-NN) assume that data points should

be in the same class if their measured distance is small [48]. k-NN is a simple method that can be used for nonlinear data. However, it can be slow for big datasets because of the computational work of calculating distances for each point. k-NN is also sensitive to imbalanced datasets [49]. Random Forest (RF), an ensemble learner, combines the output of numerous decision trees to produce a single conclusion [50]. Random forest, as opposed to decision trees, reduces the possibility of overfitting. It is also a preferable choice for imbalanced datasets since it penalizes misclassification errors for each tree and produces superior results when combined with sampling strategies. Support vector machines (SVM) search for a hyperplane (line) to differentiate two classes [51]. SVM can work with both linearly separable and nonlinear datasets by choosing the appropriate kernel method. SVM margins favor the majority target variable in the scenario of imbalanced datasets.

3.4. Evaluation metrics

Various evaluation metrics can be used to evaluate the performance of a classification model. Among them, accuracy is a frequently used metric since it is easy to calculate and interpret the results. Accuracy is essentially the ratio of properly predicted values to the number of all dataset instances ((TP + TN) / (TP + TN + FP + FN)) [52]. We benefited from the confusion matrix to define the main evaluation metrics that are summarized in Table 1. The confusion matrix's four primary values are displayed in Table 1 and are described following [52, 53].

Table 1. Confusion matrix for a binary classification problem

	Predicted positive	**Predicted negative**
Actual positive	True Positive (TP)	$False\,Negative\,(FN)$
Actual negative	False Positive (FP)	$True\,Negative\,(TN)$

- **TP**: The number of positive instances that are correctly predicted as positive, True Positive
- **TN**: The number of negative instances that are correctly predicted as negative, True Negative
- **FP**: The number of negative instances that are incorrectly predicted as positive, False Positive
- **FN**: The number of positive instances that are incorrectly predicted as negative, False Negative

Despite its ease of interpretation, accuracy can produce misleading results in the case of imbalanced datasets. Precision, recall, F1-score, and area under the curve are also calculated within the limitations of this study to provide a sound interpretation of the findings [52, 53]. Precision is a measure to calculate the percentage of correctly predicted true positive instances out of all predicted true positive instances (TP / TP + FP). The performance of the model improves as the precision value increases. The recall tries to answer the question of how many true positive points are correctly classified from all positive instances (TP / TP + FN). Recall and precision have the same interpretation. Higher recall values also indicate better classification models. By calculating the harmonic mean of a classifier's precision and recall, the F1-score integrates both into a single metric. Determining the model performance can be particularly challenging when recall and precision show contradictory behavior. In these circumstances, F1-score is advantageous. The better classification models have a higher F1-score, which ranges from 0 to 1.

The recall is also known as true positive rate (TPR) while precision is called false positive rate (FPR). The Receiver Operator Characteristic (ROC) curve is a common method to measure the binary classification models' success. The ROC curve is a plot of the TPR versus FPR at various threshold values. As a summary of the ROC curve, the area under the curve (AUC) explains the model performances. AUC values that are higher also achieve greater model results.

4. FuzzyCSampling

This section introduces our proposed approach which is a data-level strategy to balance the dataset's class distribution, from an algorithmic standpoint by utilizing the previously explained methods. Figure 1 depicts the block diagram of the proposed FuzzyCSampling approach. FuzzyCSampling starts with identifying majority and minority target variables. Then, the majority target class is divided into k number of clusters by fuzzy c-means clustering. The primary objective of using fuzzy c-means clustering is to reduce information loss while balancing the dataset. A data point can be assigned into multiple clusters by fuzzy c-means clustering with a likelihood. We also implement fuzzy c-means clustering since it delivers better results for overlapped boundary regions. The newly generated clusters are undersampled if they have more instances than the minority target variable. After that, a new dataset is generated from undersampled cluster instances and minority target variable instances. While generating the new decreased IR dataset, the minority target variable is left unchanged.

The skewed characteristics of the freshly created dataset have lowered. However, the dataset is still imbalanced. As a second approach, we also implement an oversampling step to balance the dataset obtained from FuzzyCSampling. We oversample the train data resulting from the hold-out method. In section 5, the comparative outcomes of novel methods are shown.

5. Experiments

In this section, we described our experiments with FuzzyCSampling to handle imbalanced datasets. We also presented a comparison of FuzzyCSampling with predominantly employed solutions for imbalanced datasets including such undersampling, SMOTE, and BorderlineSMOTE oversampling. The Python programming language has been used to conduct the experiments in this research (Python 3.7.9). We have utilized scikit learn (0.24.0)[54], pandas (1.3.1)[55], numpy (1.21.5)[56], fcmeans(1.6.3)[57], matplotlib(3.4.2)[58], and imblearn(0.7.0)[59] open source libraries. The experiment was conducted on a computer equipped with Intel Core i7 – 2.70 GHz CPU, 32GB of RAM, and Windows 10 with the 64-bit operating system. The aforementioned classifiers k-NN, RF, and SVM have been implemented without extensive hyperparameter optimization. We have also created a model without sampling strategies for comparative purposes. We utilized five evaluation metrics to assess classification performance which are accuracy, precision, recall, AUC, and F1-score.

5.1. Experimental datasets

We evaluated the effectiveness of the proposed approach on three datasets from different domains with a wide range of IR. Table 2 summarizes the main features of the datasets which are dataset size (size), minority target variable size (#min), majority target variable size (#max), imbalanced ratio (IR), and the number of attributes used in the dataset (#attr).

The values in Table 2 show the dataset characteristics obtained after initial filtering for model input data. Table 2 lists the datasets in order of their IR. The Pima Indians Diabetes Database is a dataset with the least IR that belongs to the health domain [60]. The dataset's goal is to diagnostically forecast whether a patient

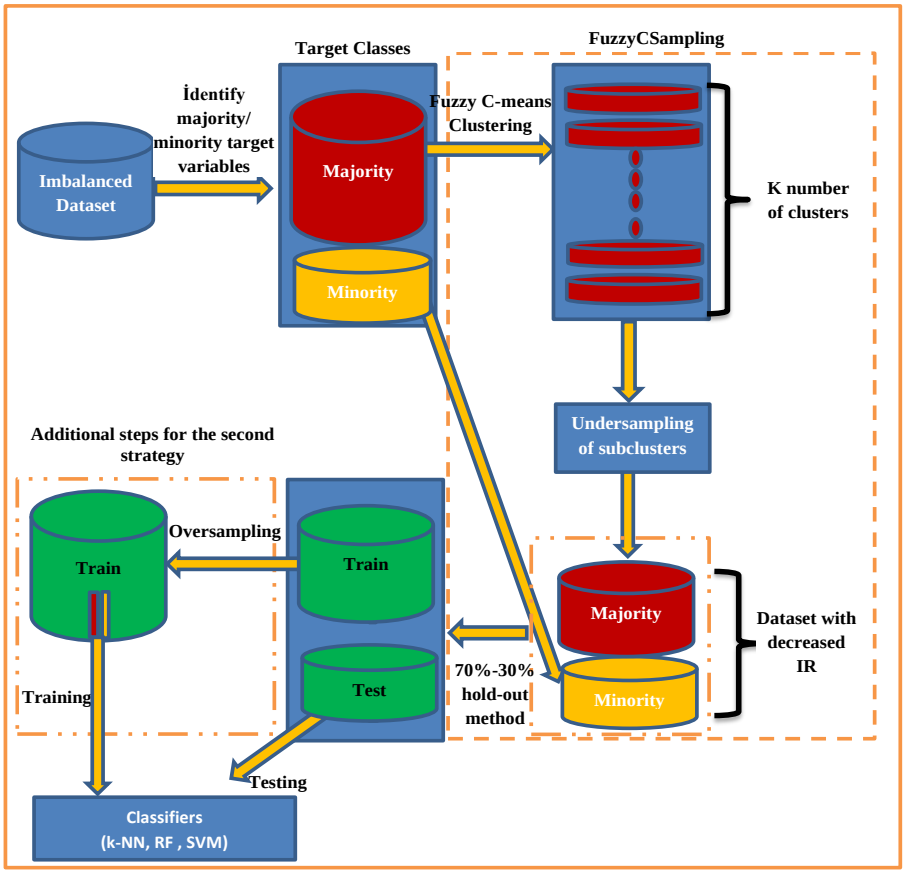

Figure 1. Block diagram of the proposed approach (FuzzyCSampling).

Table 2. The features of the datasets

Name	Size	#min	#max	#attr	IR
Pima Indians Diabetes Database	768	268	500	8	1.87
KDD Cup network intrusion detector Dataset	88831	999	87832	29	88
Credit Card Fraud Detection	275663	473	275190	29	581.8

has diabetes using diagnostic metrics contained in the dataset. The KDD Cup Dataset is an international competition dataset that is originally a multiclass classification problem with the objective of developing a network intrusion detector [61]. We chose the target variables "normal" and "r2l" to construct a new dataset for the binary classification task. The "normal" target variable is our majority target variable with a size of 87,832 and "r2l" is selected for our minority target variable with a size of 999. We have also only used 29 attributes —including our target variable— out of 42 following our fundamental preprocessing. The third dataset is related to detecting credit card fraud. The dataset includes credit card transactions performed by European cardholders in September 2013 with a size of 275,663 samples after preliminary analysis [62]. Twenty-eight variables out of 30 are obtained from PCA analysis. Only 'Time' and 'Amount' features are given in original values. We have utilized the 'Amount' and the 28 PCA features for building the model.

We have adopted the hold-out method for model evaluation. For the sake of simplicity and common practice in the field, all datasets were divided into 70% and 30% training and testing, respectively. The datasets were divided prior to the sampling step in order to prevent data leakage.

5.2. Experimental result and discussion

In this section, we summarize the FucczyCSampling approach's performance in terms of the aforementioned evaluation metrics. Table 3, Table 4, and Table 5 present experiment results of Pima Indians Diabetes, KDD Cup Network Intrusion Detector, and Credit Card Fraud Detection datasets respectively. We discuss the results according to the datasets, algorithms, and sampling strategies. We also shared the experimental results of the algorithms without any balancing strategies that are represented by "I.D."

Table 3 shows that the proposed FuzzyCSampling approach outperforms state-of-the-art approaches in terms of accuracy, AUC, and F1-score for k-NN algorithm. In addition, FuzzyCSampling surpasses I.D. for all evaluation metrics. However, undersampling and BorderlineSMOTE approaches are preferred over FuzzyCSampling in terms of precision and recall respectively. FuzzyCSampling, in particular, improves F1-score significantly. When utilized in conjunction with RF, FuzzyCSampling significantly outperforms the undersampling approach. We avoid making a comparison between sampling techniques for SVM in terms of evaluation metrics since SVM algorithm-based approaches do not give satisfactory results. For the Pima Indians Diabetes Dataset, FuzzyCSampling always outperforms the combination of FuzzyCSampling and BorderlineSMOTE.

Table 3. Experiment results for Pima Indians Diabetes dataset.

	k-NN					RF					SVM				
Sampling Technique	**ACC**	**AUC**	**PRE**	**REC**	**F1**	**ACC**	**AUC**	**PRE**	**REC**	**F1**	**ACC**	**AUC**	**PRE**	**REC**	**F1**
I.D.	85.7	81.6	82.5	70.0	75.9	89.0	85.0	87.0	77.0	82.0	33.0	25.0	2.0	2.0	2.0
Undersampling	80.7	80.9	90.4	69.5	78.6	79.5	79.5	81.8	76.8	79.2	24.8	24.6	29.03	32.9	30.8
SMOTE	86.5	85.4	77.2	82.4	79.7	88.3	86.0	83.0	79.7	81.3	17.0	17.0	9.0	17.5	12.0
BorderlineSMOTE	80.0	81.0	64.5	83.7	72.9	89.1	87.5	82.6	83.7	83.2	17.7	16.9	7.0	14.0	10.0
FuzzyCSampling	88.5	87.2	88.0	81.4	84.6	85.6	84.2	84.0	77.7	80.7	28.7	25.0	8.0	8.0	8.0
FuzzyCSampling+ BorderlineSMOTE	85.1	85.6	77.1	87.6	82.0	84.6	83.4	81.8	77.7	79.7	17.7	16.7	9.0	12.0	10.0

Table 4's analysis of experiment findings using the KDD Cup Dataset shows that FuzzyCSampling considerably outperforms previously proposed sampling strategies in F1-score for k-NN algorithm. While there is a decrease in recall, FuzzyCSampling also provides the highest precision score of any sampling method. When FuzzyCSampling is used in conjunction with the RF algorithm, it produces better results in terms of accuracy, precision, AUC, and F1-score, especially in comparison to I.D. and undersampling. The FuzzyCSampling-based RF model performs similarly or slightly worse than previous approaches in recall. All sampling methods and I.D. typically produce poor performance outcomes for the SVM algorithm. Table 4 demonstrates that FuzzyCSampling produces favorable performance to oversampling techniques. The implementation of BorderlineSMOTE after FuzzyCSampling decreases the model performance according to the comparison of FuzzyCSampling and FuzzyCSampling combined with BorderlineSMOTE. FuzzyCSampling improves the F1-score performance of all three algorithms.

Table 5 presents the experimental results of FuzzyCSampling on the model created using the Credit Card Fraud Detection dataset that has the biggest size and IR among the experimental datasets. FuzzyCSampling performs better than I.D., SMOTE, and BorderlineSMOTE in F1-score for k-NN algorithm. Although FuzzyCSampling produces mildly worse recall results than undersampling, it produces better precision results.

Table 4. Experiment results for KDD Cup Network Intrusion Detector dataset.

	k-NN					RF					SVM				
Sampling Technique	**ACC**	**AUC**	**PRE**	**REC**	**F1**	**ACC**	**AUC**	**PRE**	**REC**	**F1**	**ACC**	**AUC**	**PRE**	**REC**	**F1**
I.D.	99.8	96.3	96.8	92.6	94.7	99.0	98.1	98.2	96.3	97.3	98.4	64.1	29.4	29.0	29.2
Undersampling	98.7	97.7	47.6	96.6	63.8	98.7	97.8	47.3	97.0	63.6	97.7	93.2	32.2	88.6	47.2
SMOTE	99.7	98.7	83.2	97.6	89.8	99.9	98.1	97.6	96.3	96.9	94.2	96.4	16.2	98.6	27.8
BorderlineSMOTE	99.7	98.8	84.0	98.0	90.4	99.9	98.3	98.9	96.6	97.8	94.5	95.2	16.7	96.0	28.4
FuzzyCSampling	99.8	97.0	94.9	94.6	94.8	99.9	98.2	99.3	96.3	97.8	98.5	73.0	36.0	47.0	41.0
FuzzyCSampling+ BorderlineSMOTE	99.3	98.5	65.4	97.6	78.3	99.9	98.3	96.3	96.6	96.5	94.7	95.5	17.2	96.3	29.2

FuzzyCSampling also results in an improvement in all evaluation metrics for RF-based models except the undersampling strategy. In addition, FuzzyCSampling produces better outcomes in terms of AUC and precision when compared to undersampling. SVM-based models on Credit Card Fraud Detection dataset perform better than models created using earlier datasets. Aside from the undersampling strategy, FuzzyCSampling significantly enhances the performance of SVM-based models. Table 5 shows us that the ensemble of FuzzyCSampling and BorderlineSMOTE decreases the performance of the FuzzyCSampling. However, implementing a final oversampling step increases the model performance in terms of recall. Overall, oversampling highly imbalanced datasets degrades model performance in terms of precision, recall, and F1-score when compared to undersampling strategies.

Table 5. Experiment results for Credit Card Fraud Detection dataset.

	k-NN					RF					SVM				
Sampling Technique	**ACC**	**AUC**	**PRE**	**REC**	**F1**	**ACC**	**AUC**	**PRE**	**REC**	**F1**	**ACC**	**AUC**	**PRE**	**REC**	**F1**
I.D.	99.0	85.7	87.5	71.5	78.7	99.0	86.8	94.3	73.7	82.7	99.0	66.3	37.1	32.8	34.0
Undersampling	91.5	91.3	96.7	85.5	90.7	94.0	93.9	96.1	91.3	93.6	88.7	88.6	91.4	84.7	87.9
SMOTE	99.9	89.0	71.8	78.1	74.8	99.9	88.3	92.1	76.6	83.3	98.8	89.9	10.8	81.0	19.1
BorderlineSMOTE	99.9	88.6	71.6	77.3	74.3	99.9	85.7	94.2	71.5	81.3	94.0	84.9	0.2	75.9	4.0
FuzzyCSampling	94.5	91.4	98.3	83.5	90.3	96.0	94.1	97.0	89.2	93.2	85.1	83.0	74.6	77.8	76.2
FuzzyCSampling+ BorderlineSMOTE	90.5	92.0	78.3	95.7	86.1	94.9	94.3	90.9	92.8	91.8	73.3	75.7	54.2	82.1	65.3

6. Conclusion

Imbalanced datasets, defined as a skewed distribution of target variables, have always presented a challenge to machine learning models. Many solutions for mitigating the negative effects of imbalanced datasets on model construction have been proposed. This article presents a novel method, called FuzzyCSampling, for learning from imbalanced datasets which is based on fuzzy c-means clustering and sampling strategies. On various datasets from various industries, we presented the performance of the proposed FuzzyCSampling approach in comparison to formerly proposed approaches. Experiments utilizing datasets with varying degrees of IR demonstrate that FuzzyCSampling significantly improves model performance. The experiments led to the following findings:

- For all three datasets, FuzzyCSampling enhances the performance of the k-NN method more than RF and SVM algorithms.
- Throughout this study, FuzzyCSampling performs better regardless of dataset size or IR. More experiments on diverse datasets are needed to establish a link between FuzzyCSampling performance and dataset size.
- FuzzyCSampling always shows a slight improvement regardless of IR. F1-score and AUC performance have increased noticeably.

- FuzzyCSampling mostly outperforms oversampling strategies in F1-score and AUC. The findings also demonstrate that, depending on the dataset, the proposed solution outperforms the undersampling strategy.
- FuzzyCSampling also reveals that soft clustering techniques can improve model performance by reducing information loss.
- The findings also support the concept that for imbalanced datasets accuracy can result in misleading interpretation of the model [52].

Although the proposed technique has improved the model performance for the specified datasets, more work will be required to fix the shortcomings and thoroughly examine the model performance. We only tested FuzzyCSampling with three classification algorithms in the scope of this article. Future research will focus on testing FuzzyCSampling with more learners. FuzzyCSampling uses a cluster number that is determined heuristically. Future studies will also concentrate on the automatic determination of cluster numbers. Even though the scope of this research is limited to the binary imbalanced classification problem, the proposed strategy can be adapted to multiclass classification problems for future studies. The results of the experiments also highlight the importance of selecting appropriate evaluation metrics in order to pinpoint model performance, as proposed strategies perform differently in terms of evaluation metrics. Finally, more datasets with different degree of IR from various industries will be examined by using FuzzyCSampling to solidify the performance of the proposed approach.

References

[1] Mohamed AE. Comparative study of four supervised machine learning techniques for classification. International Journal of Applied Science and Technology, 2017; 7 (2).

[2] Trifonov R, Gotseva D, Angelov V. Binary classification algorithms. International Journal of Development Research, 2017; 7 (11): 16873-16879.

[3] Aly M. Survey on multiclass classification methods. Neural Netw, 2005; 19(1-9): 2.

[4] de Carvalho AC, Freitas AA. A tutorial on multi-label classification techniques. Foundations of computational intelligence volume, 2009; 5: 177-195.

[5] Chawla NV, Japkowicz N, Kotcz A. Special issue on learning from imbalanced data sets. Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data explorations newsletter, 2004; 6(1): 1-6.

[6] Sohony I, Pratap R, Nambiar U. Ensemble learning for credit card fraud detection. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data;2018. pp. 289-294.

[7] Manek AS, Samhitha MR, Shruthy S, Bhat VH, Shenoy PD et al. RePID-OK: spam detection using repetitive preprocessing. In IEEE 2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies; 2013. pp. 144-149.

[8] Gupta S, Gupta MK. A comprehensive data-level investigation of cancer diagnosis on imbalanced data. Computational Intelligence, 2022; 38 (1): 156-186.

[9] Padurariu C, Breaban ME. Dealing with data imbalance in text classification. Procedia Computer Science, 2019; 159: 736-745.

[10] Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced datasets. Computational intelligence, 2004; 20 (1): 18-36.

[11] He H, Garcia EA. Learning from Imbalanced Data IEEE Transactions on Knowledge and Data Engineering, 2009; 21 (9).

[12] Weiss GM. Mining with Rarity: a Unifying Framework, Association for Computing Machinery's (ACM) Special Interest Group (SIG) on Knowledge Discovery and Data Mining Explorations Newsletter, 2004; 6 (1): 7-19.

[13] Visa S, Ralescu A. Issues in mining imbalanced data sets-a review paper. In Proceedings of the sixteen midwest artificial intelligence and cognitive science conference; 2005. pp. 67-73.

[14] Lópe V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information sciences 2013; 250: 113-141.

[15] Chawla NV. Data mining for imbalanced datasets: An overview. Data mining and knowledge discovery handbook, 2009: 875-886.

[16] Salunkhe UR, Mali SN. Classifier ensemble design for imbalanced data classification: a hybrid approach. Procedia Computer Science, 2016; 85: 725-732.

[17] Lin CX, Sheng VS. Cost-sensitive learning and the class imbalance problem. Encyclopedia of machine learning, 2011: 231-235.

[18] Provost F. Machine learning from imbalanced data sets 101, Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets, Austin, Texas, USA;2000. 68, 1-3.

[19] Chawla NV, Japkowicz N, Koltcz A. Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets, Washington, DC, USA, 2003.

[20] Kotsiantis S, Kanellopoulos D, Pintelas P. Handling imbalanced datasets: A review. GESTS international transactions on computer science and engineering, 2006; 30 (1): 25-36.

[21] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research, 2002; 16: 321-357.

[22] Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning", Advances in Intelligent Computing. ICIC. Lecture Notes in Computer Science, Hefei, China 2005: 878-887.

[23] Liu XY, Wu J, Zhou ZH. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 2008; 39 (2): 539-550.

[24] Ramentol E, Caballero Y, Bello R, Herrera F. SMOTE-RSB*: a hybrid preprocessing approach based on over-sampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowledge and information systems, 2012; 33 (2): 245-265.

[25] Chawla NV, Lazarevic A, Hall LO, Bowyer KW. SMOTEBoost: Improving prediction of the minority class in boosting. In: European conference on principles of data mining and knowledge discovery Springer, Berlin, Heidelberg; 2003. pp. 107-119.

[26] Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2009; 40 (1): 185-197.

[27] Zhu Y, Yan Y, Zhang Y, Zhang Y. EHSO: Evolutionary hybrid sampling in overlapping scenarios for imbalanced learning. Neurocomputing, 2020; 417: 333-346.

[28] García S., Herrera F. Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy, in Evolutionary Computation, 2009; 17 (3): pp. 275-306. doi: 10.1162/evco.2009.17.3.275.

[29] Galar M, Fernández A, Barrenechea E, Herrera F. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern recognition 2013; 46 (12): 3460-3471.

[30] Zhang C, Guo J, Qi C, Jiang ZL, Liao Q, Yao L, Wang X. Ehsboost: Enhancing ensembles for imbalanced data-sets by evolutionary hybrid-sampling. In: IEEE 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC); 2017: 118-123.

[31] Guo H, Diao X, Liu H. Improving undersampling-based ensemble with rotation forest for imbalanced problem. Turkish Journal of Electrical Engineering and Computer Sciences, 2019; 27 (2): 1371-1386.

[32] Sun J, Li H, Fujita H, Fu B, Ai W. Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. Information Fusion, 2020; 54: 128-144.

[33] Sun J, Lang J, Fujita H, Li H. Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Information Sciences, 2018; 425: 76-91.

[34] Rayhan F, Ahmed S, Mahbub A, Jani MR, Shatabda S et al. MEBoost: mixing estimators with boosting for imbalanced data classification. In: IEEE 2017 11th international conference on software, knowledge, information management and applications (SKIMA); 2017. pp. 1-6.

[35] Patel H, Thakur GS. Classification of imbalanced data using a modified fuzzy-neighbor weighted approach. International Journal of Intelligent Engineering and Systems 2017; 10 (1): 56-64.

[36] Patel H, Thakur GS. An improved fuzzy k-nearest neighbor algorithm for imbalanced data using adaptive approach. IETE Journal of Research 2019; 65 (6): 780-789.

[37] Patel H, Thakur GS. Improved fuzzy-optimally weighted nearest neighbor strategy to classify imbalanced data. Int J Intell Eng Syst, 2017, 10, pp.156-162.

[38] Patel H, Rajput DS, Stan OP, Miclea LC. A new fuzzy adaptive algorithm to classify imbalanced data. CMC-Computers Materials Continua 2022; 70 (1): 73-89.

[39] Maraş A, Çiğdem EROL. Emerging Trends in Classification with Imbalanced Datasets: A Bibliometric Analysis of Progression. Bilişim Teknolojileri Dergisi 2022; 15 (3): 275-288.

[40] Ali H, Salleh MNM, Saedudin R, Hussain K, Mushtaq MF. Imbalance class problems in data mining: a review. Indonesian Journal of Electrical Engineering and Computer Science, 2019; 14 (3): 1560-1571.

[41] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F.A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 2011; 42 (4): 463-484.

[42] Liu XY, Zhou ZH. Ensemble methods for class imbalance learning. Imbalanced Learning: Foundations, Algorithms, and Applications, 2013: 61-82.

[43] Patel H, Rajput SD, Reddy TG, Iwendi C, Bashir et al. A review on classification of imbalanced data for wireless sensor networks. International Journal of Distributed Sensor Networks, 2020; 16 (4).

[44] Dubey A, Choubey APDA. A systematic review on k-means clustering techniques. Int J Sci Res Eng Technol (IJSRET, ISSN 2278–0882), 2017; 6 (6).

[45] Uçar T, Karahoca A. Benchmarking data mining approaches for traveler segmentation. International Journal of Electrical and Computer Engineering (IJECE), 2021; 11 (1): 409-415.

[46] Shelke MS, Deshmukh PR, Shandilya VK. A review on imbalanced data handling using undersampling and oversampling technique. Int. J. Recent Trends Eng. Res 2017; 3 (4): 444-449.

[47] Devi D, Purkayastha B. Redundancy-driven modified tomek-link based undersampling: A solution to class imbalance. Pattern Recognition Letters, 2017; 93: 3-12.

[48] Sun B, Du J, Gao T. Study on the improvement of k-nearest-neighbor algorithm. In: IEEE 2009 International Conference on Artificial Intelligence and Computational Intelligence; 2009: 4. pp. 390-393.

[49] Liu C, Cao L, Philip SY. A hybrid coupled k-nearest neighbor algorithm on imbalance data. In: IEEE 2014 International Joint Conference on Neural Networks (IJCNN) 2014; 2011-2018.

[50] Boulesteix AL, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012; 2 (6): 493-507.

[51] Dibike YB, Velickov S, Solomatine D. Support vector machines: Review and applications in civil engineering. In: Proceedings of the 2nd Joint Workshop on Application of AI in Civil Engineering 2000; 215-218.

[52] Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. Pattern recognition letters, 2009; 30 (1): 27-38.

[53] Japkowicz N. Assessment metrics for imbalanced learning. Imbalanced learning: Foundations, algorithms, and applications 2013: 187-206.

[54] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 2011; 12: 2825-2830

[55] The pandas development team. pandas-dev/pandas: Pandas. Zenodo, 2022. https://doi.org/10.5281/zenodo.7093122

[56] Harris CR, Millman KJ, van der Walt SJ et al. Array programming with NumPy. Nature, 2020; 585: 357–362. https://doi.org/10.1038/s41586-020-2649-2

[57] Madson LDD. fuzzy-c-means: An implementation of Fuzzy C-means clustering algorithm. Zenodo, 2019. doi: 10.5281/zenodo.3066222

[58] Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 2007; 9 (3): 90-95. doi: 10.1109/MCSE.2007.55

[59] Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research, 2017; 18 (1): 559-563.

[60] Zehra A, Asmawaty T, Aznan MA. A comparative study on the pre-processing and mining of Pima Indian Diabetes Dataset. In: 3rd International Conference on Software Engineering & Computer Systems 2014; 1-10.

[61] Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. In: 2009 IEEE symposium on computational intelligence for security and defense applications; 2009; 1-6.

[62] Dal Pozzolo A, Caelen O, Johnson RA, Bontempi G. Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE symposium series on computational intelligence; 2015; 159-166.