Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-022-00987-w

ORIGINAL RESEARCH

Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (SBS)

Hadeel Ahmad[1] · Bassam Kasasbeh[1] · Balqees Aldabaybah[1] · Enas Rawashdeh[2]

Received: 14 February 2022 / Accepted: 2 May 2022

Abstract Credit card fraud is a growing problem nowadays and it has escalated during COVID-19 due to the authorities in many countries requiring people to use cashless transactions. Every year, billions of Euros are lost due to credit card fraud transactions, therefore, fraud detection systems are essential for financial institutions. As the classes' distribution is not equally represented in the credit card dataset, the machine learning trains the model according to the majority class which leads to inaccurate fraud predictions. For that, in this research, we mainly focus on processing unbalanced data by using an under-sampling technique to get more accurate and better results with different machine learning algorithms. We propose a framework that is based on clustering the dataset using fuzzy C-means and selecting similar fraud and normal instances that have the same features, which guarantees the integrity between the data features.

Keywords Under-sampling technique · Fuzzy C-means · Credit card fraud detection · Machine learning · Unbalanced dataset

1 Introduction

In the beginning of March 2020 many countries have decided to perform a lockdown to reduce the spread of COVID-19 virus, shops temporarily closed but purchasing over internet has continued and the number of online purchasing transactions has increased dramatically. Amazon has announced on its website that during the crisis customer's purchases have dramatically increased like never before [1, 2]. The COVID-19 pandemic forced everyone to adapt and shift towards online shopping, this kept everyone's health safe during the pandemic. Online shopping has become the most suitable and reliable method of purchasing since it limits physical interactions and keeps customers' health safe. According to [3], A survey had been conducted in 9 countries including about 3700 consumers, finding that the crisis has forever changed online shopping behaviors and concluding that traditional purchasing will never be the same after COVID-19. The COVID-19 pandemic is likely to have a permanent effect on shopping and it has increased the share of e-commerce in total retail [4].

Fraudsters are taking advantage of the COVID-19 pandemic by developing their strategies and finding new loopholes, they have been very quick to adapt new techniques to create fraudulent transactions. According to the European Central Bank [5] every year billions of Euros are lost due to credit card fraud transactions. These statistics are an indication that financial institutions' loss due to fraud is a major problem. Fraud detection systems (FDS) play a crucial role in securing financial institutions and minimizing the risk of financial loss. Whatever the implemented FDS is, fraudsters will keep finding out a new loophole [6]. Thus, keeping enhancing and investing in

Publisher's Note Springer International Publishing.

✉ Bassam Kasasbeh
b_kasasbeh@asu.edu.jo

[1] Department of Computer Science, Applied Science Private University, Amman 11931, Jordan

[2] Department of Management Information Systems, Albalqa' Applied University, Amman 11931, Jordan

Published online: 21 June 2022

FDS is a must, that is the challenge for all financial institutions.

To this end, it is important to identify the factors that affect the performance of credit card FDSs and find a strategy to improve and enhance the detection process. One of the crucial issues that needs more study in FDS is the dataset class imbalance issue. The Dataset is considered to be unbalanced when the classes are not equally represented in the dataset. In credit card datasets, the number of fraudulent transactions is much less than the number of normal transactions [7–9]. This is a common issue in the fraud detection process because the common case is a legitimate transaction and the less to happen is a fraudulent transaction.

The problem of the unbalanced distribution ratio of the two classes resides in that the standard algorithms such as logistic regression perform very well according to the majority class [10], as a result of that the minority class gets neglected and the algorithm treat it as a noise [11]. This means that in fraud detection process the machine learning (ML) trains the model according to the majority class. To overcome this issue, many techniques and algorithms have been implemented to reduce the gap between the two classes. Some methods were implemented to increase the number of instances in the minority class by randomly replicating the instances, which is called oversampling [7, 11, 12] and the other methods were used to reduce the instances in the majority class which is called under-sampling [7].

The under-sampling technique can be implemented using the Random Under-Sampling (RUS) approach. This approach works by randomly removing transactions from the majority class to delete from the training set [6]. Since transactions are deleted randomly, the probability of removing critical or important transactions is high, resulting in a loss in the performance of the detection process. To tackle this problem, we propose a framework based on clustering the dataset first and then resampling it using our proposed Similarity-Based Selection (SBS) process to help improve the performance and accuracy of the detection process.

The primary contribution in this paper involves building an intelligent model to overcome the class imbalance issue and the problems of RUS. We propose a framework that makes an improvement in the sampling technique to improve the performance and the accuracy of the detection process. Our framework is based on clustering the dataset using fuzzy C-means and then selecting similar fraud and normal instances that have the same features. Thus, our framework will not only solve the problems of RUS, but also guarantees the similarity and the integrity of the data features.

The rest of the paper is organized as follows: Sect. 2 describes the most related work to the problem, Sect. 3 illustrates our methodology, Sect. 4 demonstrates our experiment and results, and Sect. 5 concludes the paper.

2 Related work

Imbalanced dataset is a very critical problem in ML because it degrades the classifier performance and the minority class guessed as noise or outliers. The techniques to deal with imbalanced datasets can be done on the Data Level Technique (DLT), Algorithm Level Technique (ALT), or on Ensemble Learning-based Technique (ELT) [13, 14]. In DLT, the resampling process is performed to balance between the minority and majority class before training on a classifier. In ALT, traditional classification algorithms are modified to deal with unbalanced datasets either by modifying cost or weights. Finally, the ELT that combines the performances of multiple classifiers to make predictions [15, 16]. This section briefly reviews a few recent works that deal with imbalanced datasets.

DLT can be mainly divided into two different types: under-sampling and oversampling. under-sampling techniques aim to obtaining a balanced dataset by eliminating the instances of the majority class from the training set [7]. In other hand, Oversampling techniques aim to obtaining a balanced dataset by increasing the number of minority class by producing new synthetic samples [10].

There are many studies in the literature that used the DLT to deal with imbalanced datasets, for instance, RUS where balance is achieved though random elimination of the majority class [17], and Under-Sampling Based on Clustering (SBC) Where balance is achieved by dividing the dataset into several clusters, then, randomly selects a sufficient number of samples of the majority class from each cluster [18]. Another Under-Sampling technique called Condensed Nearest Neighbor (CNN), this technique eliminates the majority class samples that are sufficiently far from decision boundary because these are considered to be less relevant to learning [19]. Similarly, the Tomek links (TL) have also been employed to eliminate the majority class samples since, if two samples form a TL, then either one of these is noise or both instances are borderline [20]. There are many techniques that used under-sampling, other than those mentioned above like One-sided selection (OSS) [21], Neighborhood Cleaning Rule (NCL) [22], and Distance-based Under-Sampling (DUS) [23]. More recently, G. Rekha and A. Tyagi in [24] proposed a Cluster-Based Under-Sampling Using Farthest Neighbor technique (CBUFN), this technique eliminates the majority class by calculated the average distance of minority class instances and then selected the majority class based on the Euclidean distance. They adopted K-means for clustering with k value varies between 3 and 5. In another study, Guo and Wei in [25] proposed a hybrid technique based on clustering and logistic regression for imbalanced learning. In the proposed technique, clustering was used to partition samples of the majority class into clusters. Vuttipittayamongkol and Elyan in [26] present Under-Sampling method based on Recursive Neighborhood Searching (URNS), this method work to maximize the visibility of the minor class by reducing the bias using an overlap-based under-sampling method. The main sampling method was done recursively by identifying the overlapped negative instances depending on their k nearest neighbors.

On the other hand, there have been attempts to deal with this problem by using oversampling techniques, such as the Synthetic Minority Oversampling Technique (SMOTE) [27] which considered one of the most widely used and effective methods. SMOTE generates synthetic data, according to the similarities between the minority class samples by using k Nearest Neighbors (kNN) of each of the minority samples. Although the main disadvantage of this technique is the synthetic data samples may overlap with the majority class samples. To deal with this point, there are a lot of researchers worked to extend the original version of SMOTE. For instance, SMOTE-TL [17], SMOTE-ENN [28], Borderline SMOTE [29], MSMOTE [30], VFC-SMOTE [31] and etc.

In other studies, the two techniques were used together to deal with this problem. For instance, researchers in [32] enhanced the performance of Borderline-SMOTE technique. The (DBMUTE) proposed algorithm was a new under-sampling algorithm that removes major class instances with low distance motivated by the Borderline-SMOTE. Distance here is defined withing a density function that identify the shortest path between each major class instance and the centroid of minority class cluster. Hence, the insignificant instances that obscure the classification boundaries are removed. The algorithm aimed to delete the major class instances in dense minor regions.

In ALT, the classification algorithms are developed to learn from the minority class [33]. For instance, Kernel-Boundary Alignment (KBA) [34], Confusion Matrix based Kernel LOGistic Regression (CM-KLOGR) [35], and the author in [36] proposed a class rectification loss (CRL) to avoid the dominant effect of majority class by discovering small number of sampled boundaries of minority class.

One of the most popular techniques used to deal with imbalance dataset, is the ELT which work by learning from many ML classifier to produce improved results. There are three types used to define this technique: Bagging, Boosting, and Random Forest. Bagging methods aim to enhance classification accuracy and mitigate the variance by selecting and re-using data [37, 38]. Boosting methods aim to enhance the overall accuracy and prediction power of learning algorithms by following a sequential re-classification process [39]. Random forest is an ELT that is an extension of bagging technique of decision tree can be used both for regression and classification [40, 41].

In this study, we proposed new Under-Sampling technique based on similarity between the minority and majority classes using Fuzzy C-Means clustering methodology to extract a useful set of samples from the majority class. This technique was applied to a high imbalance dataset of credit card fraud detection in order to select the similar fraud and normal instances that have the same features. This will improve the process of selecting instances to get rid of the majority class other than Under-sampling techniques.

3 Methodology

To effectively tackle the problem of RUS that may result in loss of the performance of the detection process, in this paper, we propose a framework based on clustering and selecting similar instances. This framework will not only solve the problems of RUS, but also guarantees the similarity of the data features, which is challenging to the algorithm to learn from. In other words, the framework we propose makes an improvement in the sampling technique to improve the performance and the accuracy of the detection process. The proposed framework carries out the problem of class imbalance by using Fuzzy C-means clustering approach which provides a robust set for the sampling step. The clustering step guarantees the grouping of the instances based on similarity of their features. After clustering phase is done, we use our own sampling technique that is based on selecting and combining the instances that have similar features according to the desired ratios. Eventually, this helps in reducing the removing of relevant and important data that is caused when using the RUS approach. The proposed SBS framework is described in detail in Fig. 1.

After the dataset is balanced, the dataset is divided into training set and testing set. This framework uses four ML algorithms namely: ANN, LR, KNN, NB along with python as the implementation language.

In this Section, Sect. 3.1 describes the dataset that is used in this paper. Section 3.2 describes the dataset preprocessing and features selection. Section 3.3 presents an overview of the clustering algorithm that was used in this paper. Section 3.4 presents our proposed SBS technique. Finally, Section 3.5 describes the training and testing sets for the ML models. Figure 2 explains the flow diagram of SBS.

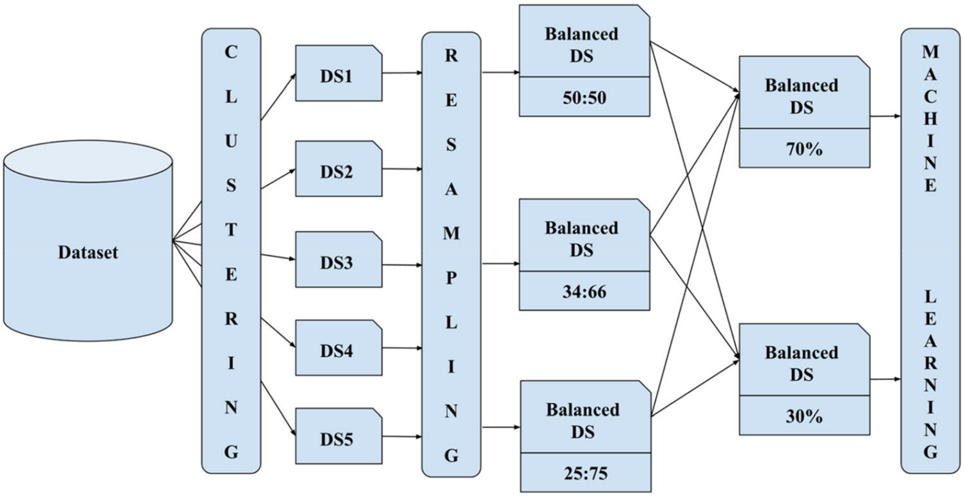

Fig. 1 The proposed similarity-based selection framework

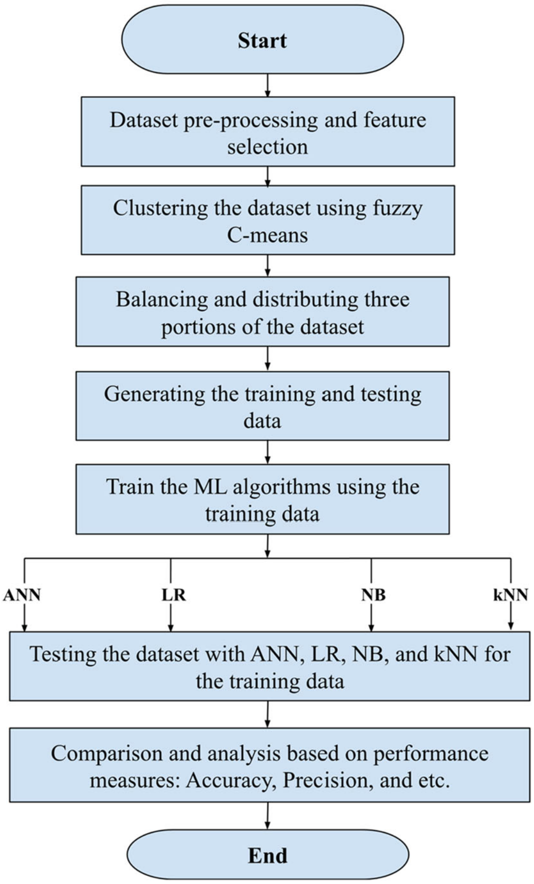
Start
Dataset pre-processing and feature selection
Clustering the dataset using fuzzy C-means
Balancing and distributing three portions of the dataset
Generating the training and testing data
Train the ML algorithms using the training data
ANN
LR
NB
kNN
Testing the dataset with ANN, LR, NB, and kNN for the training data
Comparison and analysis based on performance measures: Accuracy, Precision, and etc.
End

Fig. 2 Flow diagram of SBS

3.1 Dataset description

The dataset that is used in this paper is acquired from Kaggle [7]. The dataset contains real transactions collected from European cardholders in September 2013. Table 1 shows the information of the dataset. As shown in Table 1, the dataset is highly unbalanced.

3.2 Dataset pre-processing and feature selection

In the original dataset, there are 31 features in total. According to the relation between the features and to ensure that SBS is performed on features that do really affect the classification process, some features have been removed from the model. ‘Time’, ‘Class’ and ‘Amount’ have been removed because we do not want these features to affect the classification process. The number of features is reduced to 28 features and the dataset is ready for the clustering phase.

3.3 Fuzzy C-means clustering

Clustering is an approach of grouping the data according to certain conditions [42]. Fuzzy C-means clustering is a ML clustering method that divides the dataset into two or more clusters. The aim of this process is to group the dataset according to transactions similarities, where each cluster contains the transactions that have similar features. In this clustering method each transaction is assigned to a cluster based on the similarity which represents the distance between the instance and the cluster center [43]. The transactions that are as similar as possible are combined in the same cluster while the transactions that belong to other clusters are as dissimilar as possible. This increases the similarity among all the transactions within the same cluster and reduces the dissimilarity. In this paper, the number of generated clusters is being affected by the total number of instances, according to the following equation:

$$\begin{aligned} &Number\ of\ clusters = k * (number\ of\ instances), \\ &\quad where\ k\ is\ the\ desired\ ratio. \end{aligned} \tag{1}$$

First, the minor class instances are denoted as C_{minor} and the major class instances are denoted as C_{major}. Since the

Table 1 Description of dataset

No. of instances	No. of normal cases	No. of fraudulent cases	Normal (%)	Frauds (%)
284,807	284,315	492	99.828%	0.172%

minor class is the class of interest, the number of clusters is calculated based on instances in C$_{minor}$. See (2)

$$\text{Thus, } Number\ of\ clusters = k * C_{minor}. \quad (2)$$

We apply this equation to C_{minor}, there are 492 instances in C_{minor} and we choose to reach 10% of C$_{minor}$ in each cluster. This means we get five clusters each with 10% of the C_{minor} denoted as DS1, DS2, DS3, DS4, DS5.

3.4 Dataset resampling

After the clustering phase is done, in which similar data instances are grouped in one cluster. It is now the time to get the balanced dataset from the clusters, to be used as an input for the different classification models. This step describes in detail the proposed SBS technique that is based on instances similarity selection. In this paper, the data instances are distributed in three portions: 50:50, 34:66, 25:75 as (fraud: normal).

Since each cluster contains the instances that have the same features, we apply SBS technique that works by selecting a fraud instance and its nearest similar normal instance, starting from the center of the cluster, and doing the same towards the boundaries. We select the instances starting from cluster 1 till cluster 5, and this is repeated until we get the desired ratio. For example, in ratio 50:50, the selection technique works by selecting instances from C$_{major}$ and C$_{minor}$ as 1:1, a fraud and a normal instance, that have the same features, until we get the desired ratio.

Finally, we get three balanced portions of the dataset with different ratios (50:50, 34:66, 25:75), where each portion contains a representative portion of every cluster guaranteeing the similarity of the features between the transactions and guaranteeing the avoidance of dissimilar transactions.

After the dataset is balanced, the balanced dataset is divided into two portions: 70% for the training set, and 30% for the testing set, and then the ML algorithms namely: ANN, LR, KNN, NB are used to train the models.

4 Results and discussion

This section will show the discussion of the experimental results applying our methodology to the mentioned dataset [7]. The experimental results have been implemented using Python. They have also been executed using Intel (R) Core i5 CPU with 8.0 GB of RAM and 2.40 GHz processor with Windows 10 64-bits Operating System. In this paper, four ML algorithms are tested on the original dataset, and on the balanced datasets obtained by our SBS technique. To benchmark SBS performance, we compared the obtained results among RUS [6], and CBUFN [24].

To ensure the fairness of comparison, the dataset has been distributed in three proportions, which are taken as follows: Class A: (50% fraud, 50% normal), Class B:(34% fraud, 66% normal), and Class C:(25% fraud, 75% normal) (to ease the comparison with [6]).

In this paper, the performance measure of our method (SBS) is investigated on six evaluations metric including: Accuracy (ACC), Precision (P), F-Measure (F), Sensitivity (SEN), Specificity (SPE), and Area Under the ROC curve (AUC).

4.1 Experiment results for balancing techniques

In this research a comparison will be applied between the original dataset (without sampling), RUS, and CBUFN using LR, KNN, and NB. Table 2 shows that the original dataset is superior in accuracy compared with all the mentioned under-sampling techniques, but this thing is misleading because the used dataset is considered very highly imbalance. The high accuracy is normal because the number of non-fraud cases is very large [44]. Looking at other measures, such as Precision and F-Measure, as shown in Table 2, the original dataset was very low compared with the other under-sampling technique. On the other hand, we notice from Table 2 in Class A that the accuracy is close to all balancing methods using LA, and we note using NB shows relative superiority compared to the others. Also, we note the distinction of our method from the rest of the methods using the kNN, where the accuracy is improved by 30.9% when using RUS, and 14.41% when using CBUFN.

Table 3 confirms that our method is superior to the rest of methods, as we note sensitivity increased by 28.82%, 24.65%, and 14.64% over RUS, and CBUFN, respectively. One of the explanations for KNN superiority is that KNN uses similarity measures to select the nearest instances [45, 46] and SBS selects the similar instances in the dataset, which helps the ML algorithms to learn better. Finally, our method clearly outperformed using ANN, where the accuracy reached 0.943.

As shown in Table 3, when we applied SBS technique to the balanced dataset with 34% fraud: 66% normal, we noticed a significant improvement in the results compared

Table 2 Calculated accuracy, precision, and F-measure

Dataset distribution	Under-sampling method	LR			NB			kNN			ANN		
		ACC	P	F	ACC	P	F	ACC	P	F	ACC	P	F
Original dataset	Without sampling	0.999	0.785	0.810	0.998	0.557	0.695	0.998	0.620	0.700	0.999	0.763	0.822
Class A	RUS	0.912	0.951	0.913	0.854	0.959	0.846	0.679	0.701	0.694	0.909	1.000	0.889
	CBUFN	0.922	0.925	0.922	0.862	0.942	0.848	0.777	0.820	0.761	0.912	0.930	0.910
	SBS	0.929	0.944	0.928	0.892	0.975	0.882	0.889	0.891	0.888	0.943	0.958	0.942
Class B	RUS	0.923	1.000	0.875	0.902	1.000	0.836	0.681	0.544	0.508	0.933	0.905	0.902
	CBUFN	0.924	0.968	0.878	0.924	0.902	0.887	0.816	0.754	0.716	0.938	0.969	0.903
	SBS	0.950	0.985	0.921	0.926	0.975	0.882	0.922	0.896	0.884	0.952	0.957	0.927
Class C	RUS	0.959	0.991	0.909	0.915	0.979	0.789	0.832	0.642	0.692	0.949	0.899	0.899
	CBUFN	0.961	0.937	0.921	0.929	0.827	0.865	0.816	0.601	0.68	0.951	0.954	0.896
	SBS	0.963	0.957	0.923	0.941	0.945	0.873	0.937	0.894	0.872	0.966	0.964	0.930

Table 3 Calculated Sensitivity, Specificity and Area Under the ROC

Dataset distribution	Under-sampling method	LR			NB			kNN			ANN		
		SEN	SPE	AUC	SEN	SPE	AUC	SEN	SPE	AUC	SEN	SPE	AUC
Original dataset	Without sampling	0.838	0.999	0.919	0.926	0.999	0.962	0.804	0.999	0.902	0.892	0.999	0.946
Class A	RUS	0.878	0.949	0.914	0.757	0.964	0.86	0.687	0.669	0.678	0.818	1.000	0.909
	Cluster based	0.919	0.926	0.922	0.770	0.953	0.862	0.710	0.845	0.777	0.892	0.932	0.912
	SBS	0.912	0.946	0.929	0.804	0.96	0.892	0.885	0.892	0.889	0.926	0.96	0.943
Class B	RUS	0.777	1.000	0.888	0.718	1.000	0.859	0.477	0.789	0.633	0.899	0.951	0.925
	Cluster based	0.804	0.986	0.895	0.872	0.951	0.911	0.682	0.885	0.784	0.845	0.986	0.915
	SBS	0.865	0.993	0.929	0.804	0.990	0.897	0.872	0.948	0.910	0.899	0.979	0.939
Class C	RUS	0.839	0.997	0.918	0.664	0.995	0.829	0.405	0.861	0.663	0.899	0.966	0.932
	Cluster based	0.905	0.980	0.943	0.905	0.937	0.921	0.750	0.860	0.805	0.845	0.986	0.916
	SBS	0.892	0.986	0.939	0.811	0.984	0.898	0.851	0.966	0.909	0.899	0.989	0.944

to the other methods, especially in the sensitivity, which increased significantly by 82.8%, 27.86%, and 6.6% when we applied kNN over RUS, and CBUFN respectively, which indicates that SBS is very efficient at predicting the true positives correctly. We notice from the graphical representation in Figs. 3, 4 and 5, that our proposed model (SBS) achieved the best performance in the three different ratios. As we notice from these figures a high convergence in the results in SBS in terms of accuracy, F-measure, and Sensitivity in the different ML algorithms. In other hand, we note a divergence between the results of different algorithms in CBUFN, and RUS.

In Class C, we balanced the dataset by increasing the percentage of normal cases to 75%, as we note from Table 2 and 3, the results continued to rise in most of the performance measures in which we tested our method in comparison with the other methods. It is also noticeable that SBS method made the results close to the different ML algorithms as shown in Fig. 5.

4.2 Discussion of results for SBS

In this part, we will further explain the results of SBS technique by comparing the results between different ML algorithms that were used in this paper. We note from Table 2 that ANN algorithm has outperformed the rest of the ML algorithms, where the classification accuracies were 0.943, 0.9517, and 0.966 as compared to 0.929, 0.95, and 0.963 for LR, for Class A, B, and C, respectively. This is because ANN contains a large number of free parameters, weights and biases between interconnected neurons and other variables, which gives them flexibility to fit very complex data [44]. In addition, the decision limits of ANN are very flexible, and the powerful representation of hidden

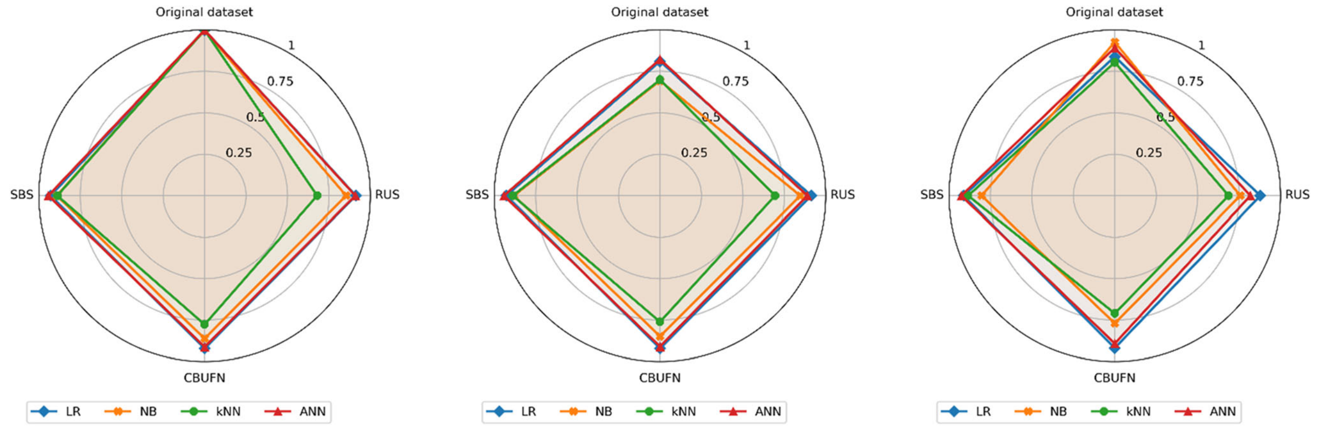

Fig. 3 Class B Results. **a** Accuracy. **b** F-measure. **c** Sensitivity

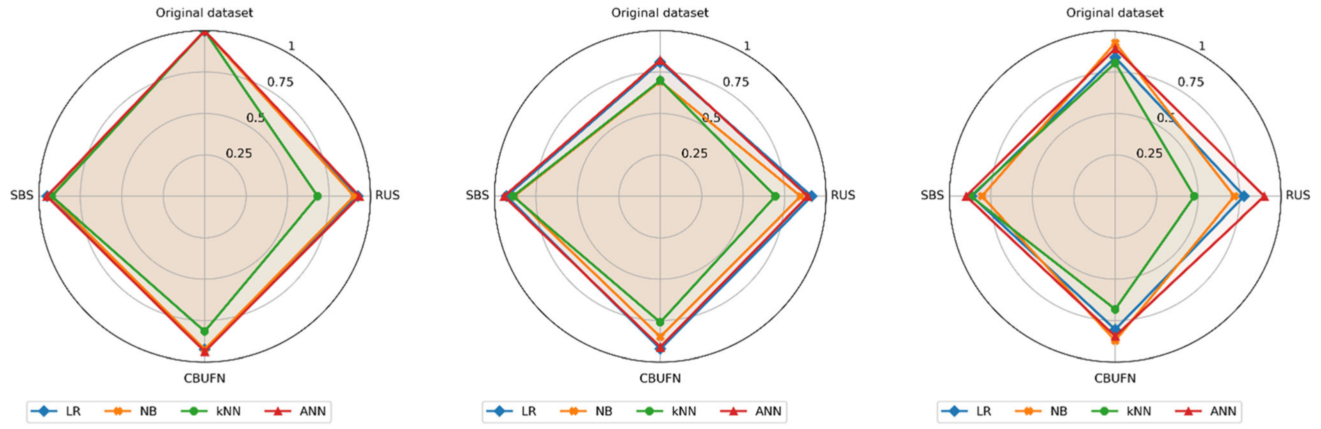

Fig. 4 Class A Results. **a** Accuracy. **b** F-measure. **c** Sensitivity

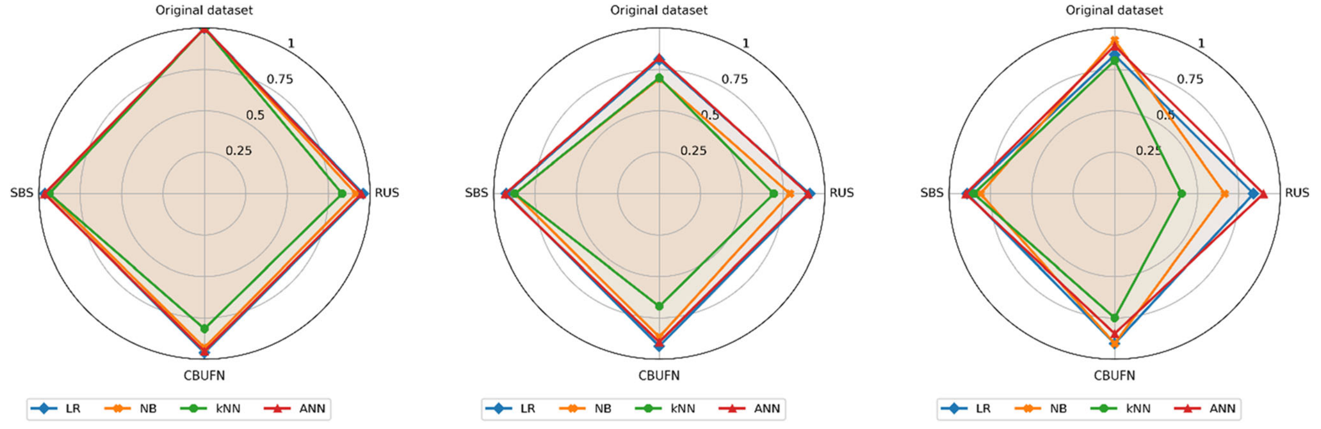

Fig. 5 Class C Results. **a** Accuracy. **b** F-measure. **c** Sensitivity

layers helps to make the right decision. For NB and kNN algorithms, they have lower accuracy compared to ANN and LR, but not too far from them.

Table 2 Shows precision results which answer how many of those records in the balanced dataset labelled by the system as fraud are actually fraud. It appears that kNN got less precision with 0.891, 0.896, and 0.894. Then LR with 0.944, 0.985, and 0.957. Then ANN with 0.958, 0.9568, and 0.964. Finally, NB got higher precision in class A with 0.975, 0.975 in Class B, and 0.945 in class C. From this we conclude that the results of the ANN and kNN were more consistent across the different classes.

Table 3 especially in the specificity indicates the quality that determines how good the test is in avoiding false alarms and shows that all algorithms that we used were able to avoid this with high efficiency, especially in Class C. Where the ANN was the highest with 0.989, 0.986 for LR, 0.984 for NB, and 0.966 for kNN.

5 Conclusion

In this paper, we investigated the problem of imbalanced credit card dataset, the study work was carried out with the purpose of finding the best under-sampling technique that gets rid of the RUS problem and guarantees better results. We propose a framework (SBS) to solve the problem of imbalanced class distribution by using Fuzzy C-means. The performance of the experiment is compared with other methods to emphasize the power of SBS technique and to prove it's superiority. SBS aims to solve the problem of RUS and guarantees the similarity and integrity of the instances' features. In the future work, we will continue to research on the basis of enhancing the proposed framework to achieve better and optimal results, which can give better performance in dealing with imbalanced credit card dataset.

References

1. Bhutani KBH (2021) COVID-19-the inflexion point for E-commerce. Indian J Econ Bus 20(2),PP 649-659.
2. Abdelrhim M, Elsayed A (2020) The effect of COVID-19 spread on the E-commerce market: the case of the 5 Largest E-Commerce Companies in the World. In: Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3621166, Jun. 2020. https://doi.org/10.2139/ssrn.3621166.
3. UNCTAD (2020) COVID-19 and e-commerce: Findings from a survey of online consumers in 9 countries. [Online accessed 6-March-2022]. https://unctad.org/system/files/official-document/dtlstictinf2020d1_en.pdf
4. Guthrie C, Fosso-Wamba S, Arnaud JB (2021) Online consumer resilience during a pandemic: an exploratory study of e-commerce behavior before, during and after a COVID-19 lockdown. J Retail Consum Serv 61:102570. https://doi.org/10.1016/j.jretconser.2021.102570
5. Wang H, Zhu P, Zou X, Qin S (2018) An ensemble learning framework for credit card fraud detection based on training set partitioning and clustering. In: 2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2018, pp 94–98. https://doi.org/10.1109/SmartWorld.2018.00051.
6. Itoo F, Meenakshi, Singh S (2021) Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. Int J Inf Technol 13(4):1503–1511. https://doi.org/10.1007/s41870-020-00430-y
7. Pozzolo AD, Caelen O, Johnson RA, Bontempi G (2015) Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE Symposium series on computational intelligence, 2015, pp 159–166. https://doi.org/10.1109/SSCI.2015.33
8. Boutaher N, Elomri A, Abghour N, Moussaid K, Rida M (2020) A review of Credit card fraud detection using machine learning techniques. In: 2020 5th International Conference on cloud computing and artificial intelligence: technologies and applications (CloudTech), 2020, pp 1–5. https://doi.org/10.1109/CloudTech49835.2020.9365916.
9. Sisodia DS, Reddy NK, Bhandari S (2017) Performance evaluation of class balancing techniques for credit card fraud detection. In: 2017 IEEE International Conference on power, control, signals and instrumentation engineering (ICPCSI), 2017, pp 2747–2752. https://doi.org/10.1109/ICPCSI.2017.8392219
10. Santoso SHH, Wibowo W (2019) Integration of synthetic minority oversampling technique for imbalanced class. Indones J Electr Eng Comput Sci 13(1):102–108. https://doi.org/10.11591/ijeecs.v13.i1.pp102-108
11. Mishra A, Ghorpade C (2018) Credit card fraud detection on the skewed data using various classification and ensemble techniques. In: 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2018, pp 1–5. https://doi.org/10.1109/SCEECS.2018.8546939
12. Prasetiyo B, Alamsyah A, Muslim MA, Baroroh N (2021) Evaluation performance recall and F2 score of credit card fraud detection unbalanced dataset using SMOTE oversampling technique. J Phys Conf Ser 1918(4):042002. https://doi.org/10.1088/1742-6596/1918/4/042002
13. Lebichot B, Le Borgne Y-A, He-Guelton L, Oblé F, Bontempi G (2020) Deep-learning domain adaptation techniques for credit cards fraud detection. In: Recent advances in big data and deep learning. Cham, pp 78–88
14. Barua S, Islam MdM, Yao X, Murase K (2014) MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans Knowl Data Eng 26(2):405–425. https://doi.org/10.1109/TKDE.2012.232
15. Rekha G, Tyagi AK, Reddy VK (2019) A wide scale classification of class imbalance problem and its solutions: a systematic literature review. J Comput Sci 15(7):886–929
16. Nanni L, Fantozzi C, Lazzarini N (2015) Coupling different methods for overcoming the class imbalance problem. Neurocomputing 158:48–61. https://doi.org/10.1016/j.neucom.2015.01.068
17. Batista GE, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl 6(1):20–29
18. Yen S-J, Lee Y-S (2006) Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In: Intelligent control and automation. Springer, pp 731–740
19. Hart P (1968) The condensed nearest neighbor rule (corresp.). IEEE Trans Inf Theory 14(3):515–516
20. Tomek I (1976) Two modifications of CNN. IEEE Trans Syst Man Cybern 6:769–772
21. Kubat M, Matwin S et al (1997) Addressing the curse of imbalanced training sets: one-sided selection. Icml 97(1):179
22. Laurikkala J (2001) Improving identification of difficult small classes by balancing class distribution. In: Conference on artificial intelligence in medicine in Europe, 2001,(LNAI,volume 2101) pp 63–66
23. Li H, Zou P, Wang X, Xia R (2013) A new combination sampling method for imbalanced data. In: Proceedings of 2013 Chinese Intelligent Automation Conference, 2013, pp 547–554

24. Rekha G, Tyagi AK (2021) Cluster-based under-sampling using farthest neighbour technique for imbalanced datasets. In: Innovations in Bio-Inspired computing and applications. Cham, pp 35–44
25. Guo H, Wei T (2019) Logistic regression for imbalanced learning based on clustering. Int J Comput Sci Eng 18(1):54–64
26. Vuttipittayamongkol P, Elyan E (2020) Overlap-based undersampling method for classification of imbalanced medical datasets. In: Artificial intelligence applications and innovations. Cham, pp 358–369
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
28. Dhalaria M, Gandotra E (2021) CSForest: an approach for imbalanced family classification of android malicious applications. Int J Inf Technol 13(3):1059–1071. https://doi.org/10.1007/s41870-021-00661-7
29. Han H, Wang W-Y, Mao B-H (2005) Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International Conference on intelligent computing, 2005,(LNTCS,volume 3644) pp 878–887
30. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C (2009) Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Pacific-Asia Conference on knowledge discovery and data mining, 2009 (LNAI,volume 5476) pp 475–482
31. Bernardo A, Della Valle E (2021) VFC-SMOTE: very fast continuous synthetic minority oversampling for evolving data streams. Data Min Knowl Discov 35(6):2679–2713
32. Bunkhumpornpat C, Sinapiromsaran K (2017) DBMUTE: density-based majority under-sampling technique. Knowl Inf Syst 50(3):827–850. https://doi.org/10.1007/s10115-016-0957-5
33. Maldonado S, Weber R, Famili F (2014) Feature selection for high-dimensional class-imbalanced data sets using support vector machines. Inf Sci 286:228–246
34. Wu G, Chang EY (2005) KBA: Kernel boundary alignment considering imbalanced data distribution. IEEE Trans Knowl Data Eng 17(6):786–795
35. Ohsaki M, Wang P, Matsuda K, Katagiri S, Watanabe H, Ralescu A (2017) Confusion-matrix-based kernel logistic regression for imbalanced data classification. IEEE Trans Knowl Data Eng 29(9):1806–1819
36. Dong Q, Gong S, Zhu X (2018) Imbalanced deep learning by minority class incremental rectification. IEEE Trans Pattern Anal Mach Intell 41(6):1367–1381
37. Sun J, Lang J, Fujita H, Li H (2018) Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Inf Sci 425:76–91
38. Zyblewski P, Sabourin R, Woźniak M (2021) Preprocessed dynamic classifier ensemble selection for highly imbalanced drifted data streams. Inf Fusion 66:138–154
39. Sabzevari M, Martínez-Muñoz G, Suárez A (2018) Vote-boosting ensembles. Pattern Recognit 83:119–133
40. Zhu X, Du X, Kerich M, Lohoff FW, Momenan R (2018) Random forest based classification of alcohol dependence patients and healthy controls using resting state MRI. Neurosci Lett 676:27–33
41. Kaur S, Singh KD, Singh P, Kaur R (2021) Ensemble model to predict credit card fraud detection using random forest and generative adversarial networks. In: Emerging technologies in data mining and information security. Springer, pp 87–97
42. Nayak J, Naik B, Behera HS (2015) Fuzzy C-Means (FCM) clustering algorithm: a decade review from 2000 to 2014. In: Computational Intelligence in Data Mining—Volume 2, New Delhi, 2015, pp 133–149. https://doi.org/10.1007/978-81-322-2208-8_14
43. Suganya R, Shanthi R (2012) Fuzzy c-means algorithm—a review. Int J Sci Res Publ 2(11):440–442
44. Kasasbeh B, Aldabaybah B, Ahmad H (2022) Multilayer perceptron artificial neural networks based model for credit card fraud detection. Indones J Electr Eng Comput Sci 26(1):1. https://doi.org/10.11591/ijeecs.v26.i1.pp%25p
45. Alqwadri A, Azzeh M, Almasalha F (2021) Application of machine learning for online reputation systems. Int J Autom Comput 18(3):492–502. https://doi.org/10.1007/s11633-020-1275-7
46. Nassif AB, Mahdi O, Nasir Q, Talib MA, Azzeh M (2018) Machine learning classifications of coronary artery disease. In: 2018 International Joint Symposium on artificial intelligence and natural language processing (iSAI-NLP), Nov. 2018, pp 1–6. https://doi.org/10.1109/iSAI-NLP.2018.8692942

Publisher's Note Springer International Publishing.