Journal of Education and Science (ISSN 1812-125X)

www.edusj.mosuljournals.com

A Hybrid Undersampling-SMOTE Method for Imbalanced Big Data Classification

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

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

Article information

Article history:
Received: September 27, 2023
Accepted: November 03, 2023
Available online: December 01, 2023

Keywords:
Big data
Classification
Imbalanced Problem
Resampling, Clustering.

Correspondence:
Shaymaa A.Razoqi
shymaa.raazoqi@uomosul.edu.iq

Abstract

Imbalanced data is an important issues and challenges faced in data classification. This will lead to poor performance of binary classifiers, this is due to bias in classification in favour of the majority class and lack of understanding of the influence of the minority class, while the minority class is usually the most important in the classification process. In order to find a compromise between the information loss and balance the data set before applying the classification, the research proposed a hybrid algorithm based on the use of clustering methods to divide the majority class into subgroups in the first phase, and using a method to encode the majority class. The Algorithm used the code to group samples that are similar to each other and reduce the majority class count. At the same time, the Synthetic Minority Oversampling Technique (SMOTE) was used to increase the number of minority class samples in the next phase. The study examined the impact of the proposed algorithm on five classifiers based on the AUC and F-score post-classification performance parameters using benchmark datasets with different sizes and imbalance factors. The results showed that the proposed algorithm significantly improved the performance of the classifiers when applied to the resampled data.

DOI: 10.33899/edusj.2023.143612.1393,

1. Introduction

There is an imbalance problem in the real data. This problem occurs when the data set for one of the classes, usually called the minority class, contains fewer samples than for the other class, usually called the majority class. This is a class imbalance problem that causes the minority class to get poor grades, which is very important. Therefore, it is difficult for a binary classifier to effectively distinguish minority classes from majority classes, especially when class imbalance is severe [1]. The classification operation involves two steps. The model construction represents the first step where each sample is considered to belong to a predefined class as specified by the class label. A training set is the collection of samples used for model construction. The next step in the classification operation is, the model built in the first step is used for classifying forthcoming or unknown samples. The evaluation results for the classification model are done when the known label of a test sample is compared with the classification result for the same test sample. The accuracy rate is the ratio of samples in the test set that are rightly classified by the classification model. It is important to note that training examples should not be manipulated to estimate model accuracy. To compose the test set, multiple independent samples must be used[2].

Class imbalance affects both traditional data and big data. However, due to the excessive class imbalance in big data, the impact can be much more serious. Granted that big data is nevertheless data and the foundations are the same, where small and big data manipulation lead to the use of the same domains, such as probability theory, mathematical statistics, computer science, and visualization. [3]. In addition, given the fact that the amount of data is massive, multiple efficient strategies have to be operated to process the data to construct notifications and decisions quickly [4]. Different types of techniques can be considered when it comes to handling classification for imbalanced problems. Through the defined solutions, the resampling-

based method is employed as a pre-processing to balance class distribution, which is considered as widely utilised of all methods. The advantages of data resampling methods are evident since it allows various suggestions to be applied to the same or different classifiers, to specify the approach that sufficiently adapts to the input data. There are two kinds of pre-processing strategies, On the one hand, the oversampling methods replicate instances of the minority class. In contrast, resampling methods eliminate examples from the majority class. Each method has its capabilities, as the oversampling method allows for maintaining the proper information of the problem, it is strengthening the borderline regions for the sets of the minority class. Also, the undersampling allows implicit refining of possible squawking data and supports the treatment of class overlapping [5].

The clustering algorithm is an essential process that can be used to improve imbalanced solutions and the distribution of large data sets. Clustering is the process of splitting data samples into different groups. The K-mean classifier is denoted as the most commonly used, unsupervised learning classifier [6]. A Mini-batch k-means algorithm is proposed as an improved copy of the K-means method. This algorithm selects a subset of samples from the dataset randomly each time and in general decreases the convergence time [7]. In the current research, a model of Hybrid Undersampling-SMOTE (HU-SMOTE) methods was designed to solve the imbalance problem in the classification model. This solution benefited from the advantages provided by the clustering algorithms. The proposed algorithm is a combination of clustering the majority class with sampling coding as the undersampling method and using the standard SMOTE method as the oversampling method. The effect of the proposed algorithm on the performance of five classifiers: Decision Tree (DT), Gradients Boost Tree (GBT), Gaussian Naïve-Bayes (GNB), Logistic Regression (LR), and Random Forest(RF), is studied based on performance measures AUC, and F-score. This was applying the classification using benchmark data sets with different dimensions and imbalance ratios from the Kaggle repository.

The rest of this research is organized as Related Works in the next section, while sections 3-5 dealt with Imbalanced in Big data classification, Synthetic Minority Over-Sampling Technique, and Clustering Algorithms. Sections 6 present the Suggested Methodology and Sections 7 present Results and Discussion, while the conclusion appears in the last section.

2. RELATED WORKS

Imbalanced data is one of the most influential performance problems of the classifier, especially on large datasets. This is because the classification results are biased in favour of the majority class and the results of the minority class are ignored in the classification, which is often an important class when applying the classification. Over the past years, several studies have presented innovative solutions to address the problem of data imbalance in a hybrid manner that integrates more than one method of intelligent and statistical methods. At first, in 2012, Ramentol and others[8], improved SMOTE method by adding the properties of the Rouph set. This method applies SMOTE, and after that keeps only the synthetic samples that belong to the minimal approximation of the minority class, it does this according to the Rough Set Theory. Later, in 2018, Douzas in his paper presented with two partners, proposed a method based on dividing training data into clusters. Then, the nature of the imbalance of these clusters is studied, and then the clusters whose imbalance ratio is less than a predetermined threshold are chosen to be resampled using the SMOTE method [9]. Miah et al. suggested using clustering with a random resampling method, and the model was implemented using a Random Forest to get rid of the problem of overfitting. The research was presented at the 1st International Conference on Advances in Science in 2019, to improve the accuracy of the intrusion detection process [10].

In 2020, a hybrid method was presented to deal with the imbalance problem in big_data by Ubaya and Juairiah. The method combines the use of RUS and SMOTE and applies that to the classification of Twitter Spam Data Using Random Forest. The hybrid method is incompetent with extremely imbalanced big data [11]. The Smtmk method used to increase classification in imbalanced data was proposed by bin Alias et al. in 2021. The method is a hybrid incorporating SMOTE with Tomek-Link (T-link) to obtain balanced training data. A step was added based on deleting the majority samples randomly, in proportion to the amount of increase in the minority class. It is taken that this method is carried out only on numeric attributes [12]. In addition to the above, Xu and others proposed a KNSMOTE algorithm based on the principle of clustering to oversample minority classes. The proposed method combines both SMOTE and K-means which are used to build new synthetic samples, and the percentage of large in the minority samples is determined according to the percentage of imbalance of the original data. The results of this proposed method are represented as the deletion of noise and boundary samples, as well as the formation of new boundary samples and the preservation of important samples [13].

In 2022 Swana et al. presented a study to address a machine fault classification when data is out of balance. The proposed model used SMOTE as an oversampling method that replicates the minority classes, and T-link is an undersampling method, it served as post-processing cleaning data. The given methodology studies the SVM, KNN, and NBC model performances based on simulated and empirical for condition monitoring of merged numerous signals as imbalanced data [14]. S. Liu noticed that oversample methods lead to an increase in the number of artificial noise samples. So, he worked on proposing a solution to address data imbalance while ensuring that no noise boundary samples are formed. He suggested the use of the

local means-based KNN (LMKNN) to filter samples in addition to adoption in describing the characteristics of the original data. Then, SMOTE is used based on the results of LMKNN to produce new not noisy central minority samples [15].
Later, in 2023, Dai proposed a new algorithm to force the SMOTE method to generate minority samples within the minority class area. This is to address the problem of the SMOTE method in generating minority samples in majority-class areas, which leads to poor performance of classifiers. The method adopted is the use of the distance-based arranging oversampling (DAO) technique, which further filters the synthesized instances from noise [16]. Few of the suggested SMOTE forms can eliminate noise problems. R. Liu suggested a method that adopts antialiasing for class boundaries, and the formation of new samples at the boundaries in the original data, where synthetic minority samples are formed based on relative and absolute densities is proposed. This method used a novel filter based on relative density with SMOTE (SMOTE-RD) to remove noise and sparsity and create boundary weights. Then, normalized weights based on absolute and sparse weights are used to generate more synthetic minority samples in the class boundary and sparse regions [17]. Islam and Mustafa suggested a Multi-Layer Hybrid (MLH) approach to address data imbalance. The proposed method adopted a two-layer model to reduce the number of majority samples. The model used three methods: ADASYN, SVM-SMOTE, and SMOTE+ ENN. Then, hybrid them on one model. As a result, gives a distributed, noise-free output. Also, for highly imbalanced datasets produced data is much more appropriate for RF and Artificial Neural Networks (ANN) to accomplish outcomes with more heightened accuracy [18].

3. IMBALANCED IN BIG DATA CLASSIFICATION

Large data is inherently complex data that requires numerous resources to process. big data is generally classified according to five dimensions, referred to as the 5 Vs: Volume, Velocity, Variety, Veracity, and Variability. Volume is a Massive amount of data that is overwhelming firms, Velocity means that huge amounts of data are collected at very high speeds, Variety means big data can be structured and unstructured, Veracity refers to the candour of the data, which indicates that the quality of the data can vary hardly which hampers correct anomaly detection, and in the last, Value means that accessing and holding big data in storage is a sufficient manner, but unless turning it into the value it is ineffective[19-20].
Imbalance is one of the significant troubles in classification and the challenge additionally evolves acute when the data has an enormous number of features or samples, different solutions there are, it can be split generally into two categories based on the nature of the data and structure of the model. In data-based methods, an attempt is constructed on the expected balance via decreasing the majority class sample data or via generating new minority class sample data as a redistribution process. In model-based methods, an endeavour is produced to construct an improved model that is sensitive to the incorrectly classifying cost of minority class sample data [5,21-22].
The data-based methods solutions incorporate numerous different resampling forms, such solutions as oversampling methods, which increase the minority samples and lead to an increase in the size of training data. This operation increases the time needed for training the model. Intelligent methods were developed for the minority oversampling operation to relieve the problem of overfitting, such as ROS, and SMOTE. The other data-based methods solutions are Under-sampling methods that make a training set balanced by discarding samples from the majority class, this must lead to a loss of information. Undersampling is an inferential cleaning of potentially noisy data and assists in the treatment of class overlap [5,21]. To balance overfitting and information loss, the aforementioned tasks have suggested a combination of intelligent sampling methods[23]. As T-link method can be an undersampling method that removes the discovered majority samples that are closest to the minority class by applying the nearest neighbours algorithm for selecting samples[24]. The T-link method also can remove samples from the minority due to the tribulation in determining the well-defined boundaries between classes. Hybrid approaches must incorporate a collection of resampling and algorithm methods. This method includes using preprocessing methods before data training and adjusting the original imbalanced data [25], whereas the data resampling method is designed to treat the problems caused by imbalance classes, the hybrid method can upgrade the performance of the model [24].

4. Synthetic Minority Oversampling Technique (SMOTE)

The SMOTE method is a process of increasing samples in minority classes. This method creates new artificial samples based on the KNN algorithm and use the original samples [1,21,23-24,26]. Various modifications were proposed to the original SMOTE method, such as the use of statistical methods, optimization methods, clustering methods, and fuzzy logic [27]. Similarly, there were proposed variants of SMOTE such as cluster SMOTE, or borderline SMOTE were found to perform sufficiently for datasets with an imbalance ratio between low to moderate [28]. In addition, the rowdy generation of methods that are based on SMOTE would additionally problematize the training process [29].

5. Clustering Algorithm

Clustering is the process of splitting data samples into different groups by classifying sets of data into a sequence of subsets denominate as clusters [30]. The clustering algorithm is an essential field to study and analyze data based on grouping data

samples. Clustering is one of the unsupervised machine-learning mechanisms. Clustering aims to find similarities in the set of data points under analysis and group similar attributed data points under a common cluster. Clustering methods subcategorize are distance-based and probability-based. Distance-based methods use a distance metric to measure the similarity between data points (including partitioning clustering as K-means and Mini-batch k-means, the two methods used in this research) whereas probability-based methods work using probability distributions to cluster the data [31-32].

The K-mean classifier is denoted as the most commonly used, unsupervised learning classifier [6]. K-means clustering aims to categorise N samples into K clusters in which each sample is put into the cluster which has the nearest mean value to it, serving as a prototype of the cluster.[30,32] The K-means clustering is based on using the Euclidean distance algorithm which aims to determine identical subgroups (clusters) of features within the data sample[31]. A Mini-batch k-means algorithm is proposed as an improved copy of the K-means method. It is different with K-means. This algorithm selects a subset of samples from the dataset randomly each time, not using all the data samples in the dataset, and therefore greatly reduces the time for clustering, and in general, decreases the convergence time[7]. Mini-batch is operated in large datasets to decline the computation time. It also endeavours to optimize the consequence of the clustering operation. The Mini-batch k-means is supposed faster than the K-mean algorithm and is commonly employed in big data sets.[32]. Mini-batch k-means, disperses the data into numerous parts or sets, thereby sidestepping significant loss in clustering performance via the concept of Mini-batch optimization[31]. The algorithm is established by extracting a small subset of samples randomly and iteratively figuring the centre of each cluster according to the contained in the subset until the centre is stable [6,33]. This strategy decreases the number of calculated distances per iteration in front of the lower cost of cluster quality.

6. Suggested Methodology

The training data needs to be balanced for the classification model to work well. Since most of the datasets in the recent problems are imbalanced, this requires a solution to balance the data before performing the classification process. To solve the problem of data imbalance using the resampling methods, the action is using either undersample or oversample techniques. The proposed algorithm in this research uses hybridization of the two techniques (undersampling and oversampling) to take benefit of the advantages of each and try to reduce their disadvantages.

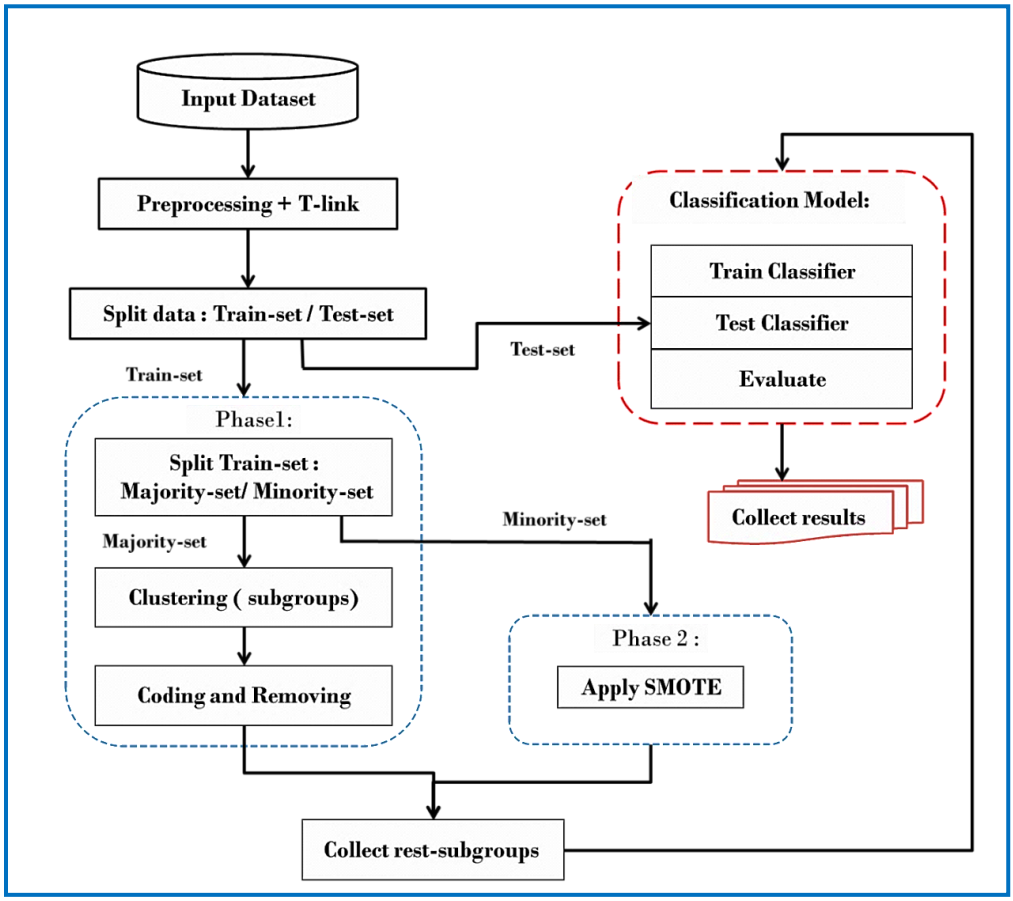

Figure 1. The proposed HU-SMOTE algorithm

The proposed HU-SMOTE algorithm is applied to the training data only before using it at the training stage of the classification model, and then test data is used in the test stage as it is natural without any modification. Preliminary data processing operations are performed to get rid of the overlapping border area samples using the T-Link method. The HU-SMOTE algorithm embraces two consecutive phases, as shown in Figure 1. The first phase is used as the undersampling method for the majority class. The undersampling method starts by splitting the training dataset into the majority set and the minority set, then using one of the clustering methods to divide the majority set into several subgroups. The methods K-means and Mini-batch k-means were applied in different experimentations to implement clustering to split the majority set. Each of the aforementioned clustering methods gave different results and a different effect on the performance of classifiers. The number of groups (*R*) here is equal to the ratio between the number of majority samples to the number of minority samples. Then, for each subgroup (*G*), give each sample (*S*) in the subgroup a Hash_Code (*HC*) that distinguishes it from other samples.

A method is developed to encode the data samples based on the centre of the subgroup, calculate the Euclidean distance (*E*), and then use the hash function with a threshold value denoted by (*TRn*) to ensure that samples close to each other had similar code. The process of reducing samples in each subgroup is done by reducing the samples which have a similar Hash_Code and those which are close to each other, so that does not significantly affect the information loss.

The ratio of similarity of samples in one group (the code that distinguishes them) is controlled by a parameter that represents a threshold degree(*TRn*). The undraped samples represent the rest of the subgroup (*restG*r). Figure 2. Shows the steps of the first phase of the algorithm.

The second phase of the HU-SMOTE algorithm begins with adding the minority set to each subgroup results from the previous phase, provided that the size of the minority set is less than the size of that subgroup. Then the SMOTE method is used to oversample the minority class by the same amount that was reduced from that majority subgroup, as shown in Figure 3. At the end of the second phase, the new training set is assembled and is ready to participate in the training phase of the classification model.

Input : *Training set,* ***TRn*** *{ Threshold value* ***}***
Output : restGr{ *Reduced Majority set* ***}***
Start :
Find ***R*** *(number of subgroups) =Majority count / Minority count*
Start
Cluster majority dataset into ***R*** *subgroups using K-means or MiniBach K-means*
For each subgroup G_r ***(*** *where* ***r = 1… R*** *) do*
  Find the number of samples (***M*** *) in* G_r
  Find a center C_r *of* G_r
  For each sample Sm (where m = 1... M) in Gr do
    Find $E(S_m)$
    Create ***HC*** *[* $E(S_m)$,***TRn****]*
    Collect S_m *with other similar samples depend on* ***HC***
    Select randomly one from each similar samples set and drop the rest
    Collect selected samples to $restG_r$
    Find the number of samples (***N****) in* $restG_r$
  END
END
End

Figure 2. The first phase of the algorithm

Input : restGr *,* ***MinSet*** *{ Minority set* ***},*** *Majority count , Minority count*
Output : NewDS { *near balanced dataset used in training* ***}***
Start :
Count dropped majority ***DMj*** *as Majority count -* ***restGr*** *count*
Find New Minority count ***NMn*** *as Minority count +* ***DMj***
Set SMOTE random strategy parameter ***RS*** *as { min :* ***NMn****, maj:* ***restGr*** *count}*
*Run SMOTE with (****MinSet*** *+* ***restGr*** *,* ***RS*** *) and Get new dataset* ***NewDS***

Figure 3. The second phase of the algorithm

7. Results and Discussion

To examine the proposed HU-SMOTE algorithm, a group of benchmark big data sets collected from website (Kaggle) was used, and these data sets had different imbalance ratios(IR), in addition to the different sizes and dimensions of those data sets, Table 1. displays the characteristics of those datasets used. The performance of the proposed algorithm and its impact on the classification methods were measured using two well-known performance metrics for imbalanced data (F-score and AUC). This is because the data used is imbalanced, so it is not possible to rely on accuracy as a basic measure. F-score is a famous evaluation metric when dealing with imbalance problems. Perhaps the most common metric is the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) analysis, and it assesses overall classification performance. The binary classifiers: Decision Tree (DT), Gradients Boost Tree (GBT), Gaussian Naïve-Bayes (GNB), Logistic Regression (LR), and Random Forest(RF) were adopted in this research and collected results at each of the benchmark data sets.

Table 1. Description of the used datasets

Datasets	Features	Samples	IR
Yeast_dataset	8	1484	0.395
Creditcard _dataset	30	284807	0.0017
Collision_dataset	69	286424	0.144
Augtrain_dataset	10	382154	0.195
Susy4_dataset	18	2711734	0.249
Higgs16_dataset	28	4954754	0.062

The HU-SMOTE algorithm use one of the two cluster methods (K-means, Mini-batch k-means). Each of the two cluster methods produced different subgroups of the majority class and this affected the creation of different sample codes in each subgroup, and then different results. The hybrid algorithm is implemented on the training dataset before classification, while the test dataset does not change. The classifiers trained on the resampled training dataset. Then, the testing result is obtained using the test

Table 2.AUC results of the used datasets in five classifiers

Datasets	DT			GNB			GBT			LR			RF		
	Clusters method			Clusters method			Clusters method			Clusters method			Clusters method		
	no cluster	k-mean	mini-batch k-mean	no cluster	k-mean	mini-batch k-mean	no cluster	k-mean	mini-batch k-mean	no cluster	k-mean	mini-batch k-mean	no cluster	k-mean	mini-batch k-mean
Augtrain	0.6664	**0.6822**	0.6811	0.6824	0.7951	**0.8009**	0.6348	**0.8245**	0.8241	0.5000	0.5577	**0.6974**	0.6571	0.7494	**0.7514**
Collision	**0.6877**	0.6827	0.6850	0.6073	0.6552	**0.6553**	0.6413	**0.7275**	0.7192	0.5000	**0.6290**	0.6229	0.6474	0.6735	**0.6740**
Creditcard	**0.8894**	0.8735	0.8785	0.8090	**0.8632**	0.8608	0.8197	0.9361	**0.9376**	0.8747	0.9356	**0.9338**	0.9007	0.9068	**0.9068**
Yeast1	0.6395	0.6366	**0.6600**	0.5264	0.5264	0.5264	0.6602	0.6941	**0.7012**	0.6061	**0.6373**	0.6340	0.6740	**0.7101**	0.7068
Susy4	0.7156	0.7213	**0.7525**	0.7502	**0.7615**	0.7520	**0.7707**	0.7517	0.7442	0.7198	0.7439	**0.7625**	**0.7670**	0.7632	0.7464
Higgs	0.5743	**0.6149**	0.6143	0.5036	0.5637	**0.5660**	0.5087	**0.5832**	0.5809	0.5007	0.5058	**0.5062**	0.5245	**0.6015**	0.5976

Table 3. F-score results of the used datasets in five classifiers

Datasets	DT			GNB			GBT			LR			RF		
	Clusters method			Clusters method			Clusters method			Clusters method			Clusters method		
	no cluster	k-mean	mini-batch k-mean	no cluster	k-mean	mini-batch k-mean	no cluster	k-mean	mini-batch k-mean	no cluster	k-mean	mini-batch k-mean	no cluster	k-mean	mini-batch k-mean
Augtrain	0.4404	**0.4600**	0.4585	0.4598	0.5439	**0.5450**	0.4048	**0.5970**	0.5967	0.0000	0.2385	**0.4438**	0.4397	0.5481	**0.5496**
Collision	**0.4478**	0.4397	0.4426	0.3511	0.3621	**0.3871**	0.4340	0.4602	**0.4896**	0.0000	0.3025	**0.3043**	0.4429	0.4730	**0.4739**
Creditcard	**0.7586**	0.7342	0.7538	**0.2099**	0.2006	0.2005	**0.6824**	0.2602	0.2794	**0.7108**	0.1198	0.1264	0.8689	0.8702	**0.8691**
Yeast1	0.4943	0.4995	**0.5305**	0.4767	0.4767	0.4767	0.5110	0.5718	**0.5822**	0.3708	**0.4534**	0.4467	0.5333	**0.5951**	0.5902
Susy4	0.5548	0.5853	**0.6360**	0.6298	**0.6426**	0.6187	**0.6534**	0.6183	0.6103	0.5313	0.6099	**0.6434**	**0.6465**	0.6436	0.6132
Higgs	0.1944	0.2041	**0.2045**	0.0200	0.1290	**0.1302**	0.0350	**0.2420**	0.2386	0.0032	0.0251	**0.0269**	0.0937	**0.2916**	0.2857

dataset. To obtain greater accuracy in the results for each experiment carried out on one of the classifiers, the classification experiments were re-applied three times, and then the average of the results that were implemented for each dataset was taken. The average results collected from the experiments when applying all classifiers at all datasets three times are shown in Table 2 and Table 3. The results are arranged depending on the classifiers at AUC, and F-score measures. The leftmost column represents the datasets and the results of each classifier are shown in three columns. The first one represents the classifier results for the original dataset, while the second and third columns represent the results of the proposed algorithm with clustering algorithms (K-means, Mini-batch k-means) respectively.

From Table 2, the results obtained when applying the DT to different datasets show that the performance of the classifier improved well in four of the six datasets concerning the AUC measure, while the results when using GNB and LR were perfect for all datasets in general, as it led to an increase in performance for the same measure. It is also noted that when classifying GBT and RF, the performance of the data sets has improved, except for one dataset, which did not give a noticeable improvement.

If we move to the F-score performance results of the classifiers in Table 3, the results showed a clear improvement in the classifier DT as well as GNB for all data sets, while the results of increased performance were in each of RF (achieved an increase in performance in five of the six data sets), and LR and GBT (an increase in performance was achieved in Four of the six data sets) were good to a median at the F-score.

When moving towards studying the effect of clustering methods on performance, from the data shown in the previous figures, it was found that the use of Mini-batch k-means gave better results than K-means on all scales with the DT classifier and for all data sets. Likewise, the results of the LR classifier in general were better when Mini-batch k-means, on the contrary, the results of the RF classifier were better when the K-means method was used. It is also noted that no significant differences appeared in performance when using GNB and GBT classifiers with the aforementioned clustering methods.

8. Conclusion

The SMOTE method is one of the most widely used methods to solve the data imbalance problem, especially for big data sets, but it also has many weaknesses. In this study, a HU-SMOTE algorithm combining two phases is developed. The first phase is based on the use of clustering the majority into subgroups. After that, an encoder is created for each sample based on the centre of the subgroup to which it belongs. Then, apply the undersampling operation depending on the code similarity. The second phase applies oversampling operation on minority class using the (SMOTE) method. The proposed algorithm aims to improve the performance of classifiers when dealing with big imbalanced data and to get rid of the weaknesses that appear when using SMOTE, as well as to preserve the information of the majority class when performing the sampling process.

The proposed algorithm has resulted in the size of the samples after balancing being preserved as much as possible and not increasing excessively, since it stipulates that the number of samples artificially added to the minority class in Phase 2 must be equal to the number of samples that were removed from the majority class in Phase 1. Practical experiments showed that the proposed algorithm led to an increase in the performance of classifiers at different degrees of imbalance ratios, as well as different data sizes. The increase in performance is observed at two measures, AUC and F-score. Experiments also showed that there are no significant differences in performance between the use of the two clustering methods (K-means and Mini-batch k-means), but the differences appear clear in the speed of implementation in using the latter.

9. ACKNOWLEDGEMENT

We extend our thanks to the University of Mosul, as well as our Colleges, the College of Computer Science and Mathematics and the College of Education for Pure Science.

10. REFERENCES

[1] F. R. Torres, J. F. Trinidad, and J. A. Ochoa, "An Oversampling Method for Class Imbalance Problems on Large Datasets," Applied Sciences, vol. 12.7, pp. 3424, 2022. DOI.org/10.3390/app12073424.

[2] A. B. Desai, Distributed AdaBoost Extensions for Cost-sensitive Classification Problems. PhD Thesis, Ahmedabad, Indian: Ahmedabad University, 2020. DOI.org/10.5120/ijca2019919531

[3] A. M. AbouTabl, Big Data Analytics for Complex Systems. PhD Thesis, Windsor, Ontario, Canada: University of Windsor, 2019.

[4] J. Wang, C. Xu, J. Zhang, and R. Zhong, "Big data analytics for intelligent manufacturing systems: A review," Journal of Manufacturing Systems, vol. 62, pp. 738-752, 2022. DIO.org/10.1016/j.jmsy.2021.03.005

[5] S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, "FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification," Pattern Recognition, vol. 124, pp. 108511, 2022. DOI.org/10.1016/j.patcog.2021.108511

[6] T. Iqbal, A. Elahi, W. Wijns, and A. Shahzad, "Exploring unsupervised machine learning classification methods for physiological stress detection," Frontiers in Medical Technology, vol. 4, 2022. DOI.org/10.3389/fmedt.2022.782756

[7] K. Peng, V. C. Leung, and Q. Huang, "Clustering approach based on mini batch kmeans for intrusion detection system over big data," IEEE Access, vol. 6, pp. 11897-11906, 2018. DOI.org/10.1109/ACCESS.2018.2810267

[8] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, "SMOTE-RSB*: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory," Knowl. Inf. Syst., vol. 33, pp. 245–265, 2012. DOI.org/10.1007/s10115-011-0465-6

[9] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," Inf. Sci., vol. 465, pp. 1–20, 2018. DOI.org/10.1016/j.ins.2018.06.056

[10] Md. O. Miah, S. S. Khan, S. Shatabda, and Md. D. Farid, "Improving detection accuracy for imbalanced network intrusion classification using cluster-based under-sampling with random forests," IEEE ICASERT, In 1st international conference on advances in science, engineering and robotics technology, United States, pp. 1-5, May 2019. DOI.org/10.1109/ICASERT.2019.8934495

[11] H. Ubaya, and R. S. Juairiah, "Performance of RUS and SMOTE Method on Twitter Spam Data Using Random Forest," In Journal of Physics: Conference Series. IOP Publishing, South Sumatera, Indonesia, pp. 012130, October 2019. DOI.org/10.1088/1742-6596/1500/1/012130

[12] M. S. Bin ALias, N. Binti Ibrahim, and Z. Bin MohdZin, "Improved sampling data Workflow using Smtmk to increase the classification accuracy of imbalanced dataset," European Journal of Molecular & Clinical Medicine, vol. 8.02, pp. 91-99, 2021.

[13] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, "A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data," Information Sciences, vol. 572, pp. 574-589, 2021. DOI.org/10.1016/j.ins.2021.02.056

[14] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset," Sensors, vol. 22.9, pp. 3246, 2022. DOI.org/10.3390/s22093246

[15] S. Liu, " Smote-lmknn: A synthetic minority oversampling technique based on local means-based k-nearest neighbor," International Journal of Pattern Recognition and Artificial Intelligence, vol. 36.05, pp. 2250019, 2022. DOI.org/10.1142/S0218001422500197

[16] Q. Dai, J. Liu, J. Zhao, "Distance-based arranging oversampling technique for imbalanced data," Neural Computing and Applications, vol. 35.2, pp. 1323-1342, 2023. DOI.org/10.1007/s00521-022-07828-8

[17] R. Liu, "A novel synthetic minority oversampling technique based on relative and absolute densities for imbalanced classification," Applied Intelligence, vol. 53.1, pp. 786-803, 2023. DOI.org/10.1007/s10489-022-03512-5

[18] M. T. Islam, and H. A. Mustafa, "Multi-Layer Hybrid (MLH) balancing technique: A combined approach to remove data imbalance," Data & Knowledge Engineering, vol. 143, pp. 102105, 2023. DOI.org/10.1016/j.datak.2022.102105

[19] C. K. Maurya, ANOMALY DETECTION IN BIG DATA. PhD Thesis, Roorkee, India: Department of Computer Science and Engineering Indian Institute of Technology Roorkee, 2016.

[20] R. Pereira, and M. Pereira, "Challenges, Open Research issues and Tools in Big Data Analytics Covid-19," International Journal for Research in Applied Science & Engineering Technology, vol. 10.4, Apr 2022. DOI.org/10.22214/ijraset.2022.41820

[21] M. Fattahi, M. H. Moattar, and Y. Forghani, "Improved cost-sensitive representation of data for solving the imbalanced big data classification problem," Journal of Big Data, vol. 9.1, pp. 1-24, 2022. DOI.org/10.1186/s40537-022-00617-z

[22] J. Leevy, Machine Learning Algorithms for Predicting Botnet Attacks in IoT Networks. PhD Thesis, Boca Raton, FL: Florida Atlantic University, 2022.

[23] Y. Shao, Imbalance Learning and Its Application on Medical Datasets. PhD Thesis, Henan, China: Georg-August-University Göttingen, 2021.

[24] K. M. Hasib, M. Iqbal, F. M. Shah, J. A. Mahmud, M. H. Popel, M. Showrov, et al., "A survey of methods for managing the classification and solution of data imbalance problem," Journal of Computer Science, vol. 16.11, 2020.

[25] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of imbalanced data: a review of methods and applications," in IOP Conference Series: Materials Science and Engineering, 2021. DOI. org/10.1088/1757-899X/1099/1/012077

[26] S. BEJ, Improved imbalanced classification through convex space learning. PhD Thesis, West Bengal, Indian: University Rostock, 2021.

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

[28] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," Information, vol. 14.1, pp. 54, 2023. DOI.org/10.3390/info14010054
[29] J. Zhang, T. Wang, W.Y. Ng, and W. Pedrycz, "Ensembling perturbation-based oversamplers for imbalanced datasets," Neurocomputing, vol. 479, pp. 1-11, 2022. DOI.org/10.1016/j.neucom.2022.01.049
[30] M. M. Chavan, A. Patil, L. Dalvi, and A. Patil, "Mini batch k-means clustering on large dataset," International Journal of Scientific Engineering and Technology Research, vol. 4.7, pp. 1356-1358, 2015.
[31] D.Tan, M. Suvarna, Y. S. Tan, J. Li, and X. Wang, "A three-step machine learning framework for energy profiling, activity state prediction and production estimation in smart process manufacturing," Applied Energy, vol. 291, pp. 116808, 2021. DOI.org/10.1016/j.apenergy.2021.116808
[32] M. A. Ahmed, H. Baharin, P. NE. Nohuddin, "Mini-Batch k-Means versus k-Means to Cluster English Tafseer Text: View of Al-Baqarah Chapter," Journal of Quranic Sciences and Research, vol. 2.2, pp. 48-53, 2021. DOI.org/10.30880/jqsr.2021.02.02.006
[33] Z. He, W. Qin, C. Duan, "Chemical composition analysis of ancient glass products based on decision tree," Highlights in Science, Engineering and Technology, vol. 42, pp. 211-219, 2023. DOI.org/10.54097/hset.v42i.7097
[34] Kaggle Machine Learning and Data-Science community. Online Available: https://www.kaggle.com/dataset

طريقة هجينة (Undersampling-SMOTE) لتصنيف البيانات الضخمة غير المتوازنة

شيماء احمد رزوقي[1]،غيداء عبدالعزيز الطالب [2]

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

المستخلص:

يعد اختلال توازن البيانات من أهم المشاكل والتحديات التي تواجه تصنيف البيانات. وهذا من شأنه أن يؤدي إلى ضعف أداء المصنفات الثنائية، وذلك بسبب انحياز التصنيف نحو فئة الأغلبية وتجاهل تأثير فئة الأقلية، في حين أن فئة الأقلية غالبا ما تكون هي الأكثر أهمية عند التصنيف. من أجل إيجاد حل وسط بين فقدان المعلومات وموازنة مجموعة البيانات قبل تطبيق التصنيف، اقترح البحث خوارزمية هجينة تعتمد على استخدام طرق التجميع لتقسيم فئة الأغلبية إلى مجموعات فرعية في المرحلة الأولى، واستخدام أسلوب ترميز فئة الأغلبية. تستخدم الخوارزمية الرموز لتجميع العينات المتشابهة مع بعضها البعض وتقليل عدد عينات فئة الأغلبية. بينما تم استخدام تقنية الإفراط في أخذ عينات الأقليات الاصطناعية (SMOTE) لزيادة عدد عينات فئة الأقليات في المرحلة التالية. يدرس البحث تأثيرات الخوارزمية المقترحة على خمسة مصنفات اعتمادا على مقاييس الأداء AUC وF−score بعد تطبيق التصنيف باستخدام مجموعات البيانات المعيارية ذات أبعاد ونسب عدم توازن مختلفة. أظهرت النتائج أن الخوارزمية المقترحة أعطت نتائج جيدة في تحسين أداء المصنفات عند تطبيقها على البيانات بعد اعادة التوزيع.