SN Computer Science _####################_
https://doi.org/10.1007/s42979-021-00850-y



ORIGINAL RESEARCH

# Anovel HEOMGA Approach for Class Imbalance Problem in the Application of Customer Churn Prediction

Ibrahim AlShourbaji[1,2] · Na Helian[1] · Yi Sun[1] · Mohammed Alhameed[3]

Received: 15 January 2021 / Accepted: 24 August 2021


**Abstract**
Making class balance is essential when learning from highly skewed datasets; otherwise, a learner may classify all instances to a negative class, resulting in a high false-negative rate. As a result, a precise balancing strategy is required. Many researchers have investigated class imbalance using Machine Learning (ML) methods due to their powerful generalization performance and interpreting capabilities, comparing with random sampling techniques, to handle the problem of class imbalance in the preprocessing phase to facilitate learning process and improve performance results of learners. In this research, an effective method called HEOMGA is presented by combining Heterogeneous Euclidean-Overlap Metric (HEOM) and Genetic Algorithm (GA) for oversampling minority class. The HEOM is employed to define a fitness function for the GA. To assess the performance of the proposed HEOMGA method, three benchmark datasets from UCI repository in the domain of customer churn prediction are examined using three different ML learners and evaluated with three performance metrics. The experiment results show the effectiveness of the proposed method compared to some popular oversample methods, such as SMOTE, ADASYN, G SMOTE, and Gaussian oversampling methods. The HEOMGA method significantly outperformed the other oversampling methods in terms of recall, G mean, and AUC when the Wilcoxon signed-rank test is used.

**Keywords** Class imbalance problem · Genetic algorithm · HEOM · Oversampling · Classification



## Introduction

The Telecom industry is evolving rapidly over time. In the same vein, the industry is facing severe revenue losses, because customers tend to leave a company and move to a competitor in the Telecom market (i.e., customer churn). The data of customers stored in such as Customer Relationship Management (CRM) systems could be transformed into valuable information with data mining and Machine Learning (ML) techniques. These techniques aid Telecom companies to formulate new policies, develop campaigns for existing clients, and figure out the main reasons behind customer churn. In this way, companies can easily observe their customer's behavior over time and manage them effectively. However, training learners with datasets which suffer from class imbalance distribution is an important and challenging problem in data mining and ML.

In recent years, the problem of imbalance class has been widely studied in the areas of ML. Typically, this problem occurs when the classes in a given dataset are unequally distributed between the minority and majority classes. Without consideration of this problem, effective learning process by classification algorithms will be a challenge, since the main goal is the detection of minority classes [1]. Addressing this problem has attracted increased attention from the research community due to its importance in different applications; examples include malware detection [2], medical diagnosis domain [3], financial crisis prediction [4], and churn prediction [5]. Several studies carried out comparisons on random

✉ Ibrahim AlShourbaji
alshourbajiibrahim@gmail.com

Na Helian
n.helian@herts.ac.uk

Yi Sun
y.2.sun@herts.ac.uk

Mohammed Alhameed
malhameed@jazanu.edu.sa

1 School of Computer Science, University of Hertfordshire, Hatfield, UK

2 Department of Computer and Network Engineering, Jazan University, Jazan, Saudi Arabia

3 Department of Computer Science, Jazan University, Jazan, Saudi Arabia



| Journal : Large 42979 | Article No : 850 | Pages : 12 | MS Code : 850 | Dispatch : 14-9-2021 |
|---|---|---|---|---|

sampling techniques to handle the class imbalance problem in the preprocessing phase. The results from these efforts highlighted that these methods were useful before applying classification algorithms [6, 7]. This is also confirmed by the work of [8], when 26 datasets were used to investigate the influence of class imbalance before and after balancing the datasets. On the other hand, it was reported that random sampling methods for class imbalance were shown not to be useful in improving the performance of prediction results [9, 10].

Balancing class is necessary when learning from highly skewed datasets, because an imbalanced dataset could result in classifying all the instances as negative, and hence leads the learner to have a high false-negative rate [11, 12]. Therefore, a balancing strategy having better interpreting capability is essential in the preprocessing phase to specify churn customers. The cost is usually high when a learner misclassifies the positive class instances, especially in churn prediction. In this work, we propose a novel method based on Heterogeneous Euclidean-Overlap Metric (HEOM) and Genetic Algorithm (GA) to generate data points from the existing minority ones rather than to use random methods. This work proposes a data-level strategy for addressing the class imbalance problem. The main objective of this work is to investigate the suitability of the proposed method in achieving optimal performance results and facilitating the learning process by the learners from imbalance datasets. A thorough empirical study was carried out which proves the significant performance gains by the proposed method compared to other popular oversampling algorithms.

The rest of the paper is organized as follows: Section "Literature Review" reviews Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling Method (ADASYN) oversampling methods. Section "Proposed Method" presents the proposed method. Section "Experiment Design" describes the imbalance customer churn datasets used to examine the proposed method, while Sect. "Results and Discussion" provides the experiment design used in this work. Section "Conclusion and Future Work" presents the results and discussion of this research. The final section concludes the paper along with future work.

## Literature Review

Research on synthesizing minority samples has been widely studied to address the problem of class imbalance distribution at data level. The random sampling method is the simplest way. Its main goal is to improve data quality in the preprocessing phase before training classification algorithms. Random sampling can be divided into two categories: random undersampling and random oversampling. In the undersampling technique, the same samples belonging to the same majority samples are removed from the dataset. For example, 30% undersampling means that 30% of the available majority instances are randomly removed from the dataset. However, by removing significant instances, this method may potentially lose valuable information. The second category attempts to create a superset of the original dataset. This can be achieved by replicating the minority instances from the existing dataset. The replication can be done either randomly or using an intelligent method. For example, 100% oversampling means that the minority instances are replicated once in average. However, a drawback with this method is that creating additional instances could have significant impacts on computational cost and overfitting.

SMOTE is an advanced method of oversampling, and it was developed by Chawla et al. [13]. This approach randomly picks one data point from the k neighbors of a minority class sample and inserts a new synthetic minority class sample on the line that connects the randomly chosen minority class sample and one of its k minority nearest neighbors, belonging to minority class sample as illustrated in Fig. 1.

He et al. [14] proposed ADASYN to overcome the problem of class imbalance. It is an oversampling method that was basically developed to reduce generating noise data and the ambiguity along the decision boundaries produced by SMOTE. The major difference between SMOTE and ADASYN is in the generation of synthetic sample points for minority data points. In ADASYN, the data points that are harder to learn are more frequently presented by this method, as shown in Fig. 2.

Recent developments of SMOTE and ADASYN, Borderline-SMOTE [15], Safe-Level-SMOTE [16], and Local Neighbourhood SMOTE [17] are some other extensions to reduce generating noise data and the ambiguity along the decision boundaries that are produced by SMOTE. These extensions attempt to create data points from the minority class that are close to the borderline between the two classes;

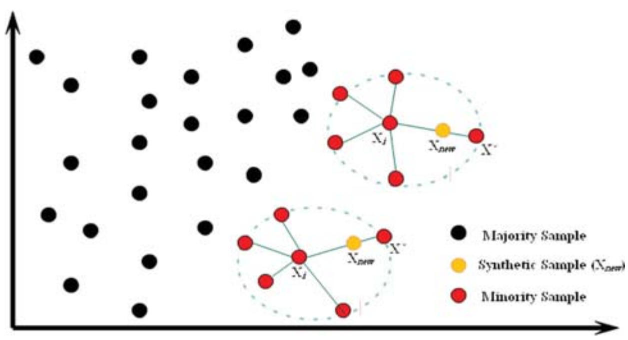


**Fig. 1** Generation of synthetic samples using SMOTE, a randomly selected minority class sample and of its $k = 5$ nearest neighbors





Journal : Large 42979 | Article No : 850 | Pages : 12 | MS Code : 850 | Dispatch : 14-9-2021

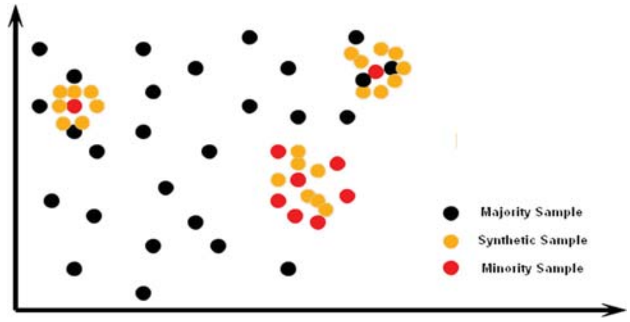


**Fig. 2** Generation of synthetic samples using ADASYN



for example, ADASYN aimed at generating minority data samples based on their distribution. Barua et al. [18] recently proposed another recent technique for imbalanced data problem; named, Majority Weighted Minority Oversampling Technique (MWMOTE). This method has several functions, which include: (a) generate a useful synthetic class sample, (b) add weights to the selected sample based on their importance, and (c) use clustering approach to produce suitable synthetic minority class samples.

Zhu et al. [19] assessed the suitability of ADASYN, Borderline-SMOTE, Random oversampling, and SMOTE strategies for class imbalance in churn prediction using 11 datasets. The results recommended that suitable sampling strategies needed to be selected, and setting of class ratio had an impact on the model performance. In another work [20], the authors investigated six sampling techniques and their accounts on four customer churn datasets. These methods include Mega-trend Diffusion Function (MTDF), Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling approach (ADASYN), Couples Top-N Reverse k-Nearest Neighbor (TRkNN), Majority Weighted Minority Oversampling Technique (MWMOTE), and Immune centroids oversampling technique (ICOTE). Their empirical results demonstrated that MTDF performed better than the other oversampling methods they used in the study. Salunkhe et al. [21] proposed a hybrid data-level approach for handling class imbalance problems. The authors combined SMOTE and undersampling techniques to achieve better results. Their aim was to focus on the majority class's necessary data and avoid removing valuable information when using the undersampling technique before the model training stage. They achieved results better than the other techniques for class imbalance.

During the last decade, a worldwide range of studies has applied Genetic Algorithm (GA) for class imbalance problems [22–24]. In the approach of [25], GA with SMOTE was combined to perform oversampling and they used different sampling rates for different minority examples until reaching the desired oversampling rate. The results showed that the proposed method achieved better performance compared to SMOTE. In another work, GenSample was proposed by [26]. They used the a GA method for oversampling minority class by taking into account the difficulty in the learning of an example and the improved performance caused by oversampling it. Their final results showed that better performance was achieved by the GenSample method compared to the traditional methods.

Distance-based algorithms are widely used for class imbalance problems to provide a numerical description of the similarity between two objects [27]. Several studies confirmed that improving the performances of distance metrics makes ML algorithms more accurate [28–30]. The aim of the research done by [31] is to improve the categorization process of the minority class by incorporating an idea of using dataset-specific distance function and choose the appropriate distance metric and k nearest-neighbor value among the five used distance metrics for five datasets. They concluded that there is no optimal distance metric for all the datasets.

Modifications can be made at the algorithm level by incorporating the cost of misclassifying minority samples or integrating one class learning algorithm. Bagging and boosting ensemble techniques can be used as cost-sensitive methods, where the classification outcome is some combination of multiple classifiers built on the dataset. Guo et al. [32] applied data boosting to improve the performance on hard samples that are difficult to classify. The algorithm-level method tries to adapt existing learning algorithms to strengthenen their learning capability regarding the majority class. However, this approach requires a deep level of understanding related to the application domain and corresponding classifiers.

Hybrid methods are also used to conquer the problem of class imbalance recently. An ensemble of classifiers can be used at the algorithm level and different sampling methods and cost-sensitive learning methods can be hybridized at the data level. The authors in [33] incorporated oversampling and undersampling with an ensemble Support Vector Machine (SVM) to improve its prediction performance. Experimental results showed that better performance was achieved by SVM when the problem of class imbalance was contained by the use of oversampling and undersampling methods compared to other classifiers and SVM alone. Based on the conducted review, the first observation indicates that solving class imbalance at the data level seems to be the most viable and widely used option in practice to provide the learner with more robust training data.

## Proposed Method

### HEOM

There are a number of distance metrics that are designed and used for measuring similarity and dissimilarity among





| Journal : Large 42979 | Article No : 850 | Pages : 12 | MS Code : 850 | Dispatch : 14-9-2021 |
|---|---|---|---|---|

samples within a given dataset. The use of these metrics depends on the nature of a dataset's attributes, whether they are numerical or only contain categorical attributes. For example, Euclidean distance is the most widely used when all the attributes are numerical. Another example, Hamming Distance can be used when only have categorical attributes. However, some other metrics were designed to handle nominal and categorical attributes, i.e., mixed or heterogeneous data such as HEOM.

HEOM becomes more popular due to its simplicity and efficiency in handling continuous and discrete attributes independently [34–37].

Considering two input vectors, x and y, the HEOM distance can be calculated by

$$d(x,y) = \sqrt{\sum_{i=1}^{n} d_i\left(x_i, y_i\right)^2}; \tag{1}$$

$d(x, y)$ is the distance between the two cases on its $i$th attribute, where

$$d(x,y) = \begin{cases} 1, & \text{if } x \text{ or } y \text{ is missing} \\ d_o(x,y), & \text{if } x \text{ and } y \text{ are discrete variables} \\ d_n(x,y), & \text{if } x \text{ and } y \text{ are continuous variables} \end{cases}. \tag{2}$$

HEOM uses the overlap metric, $d_o$, for categorical attributes

$$d_o(x,y) = \begin{cases} 0, & \text{if } x = y \\ 1, & \text{otherwise} \end{cases}. \tag{3}$$

The normalized Euclidean distance, $d_n(x, y)$, for continuous attributes

$$d_n(x,y) = \frac{\left|(x_i - y_i)\right|}{\max_a - \min_a}. \tag{4}$$

## GA

A GA searches for the global solution through an iterative process; a new population is produced at each iteration, which contains evolutions of individuals selected from the previous iteration. The initial population is generally composed of random solutions. The individuals are codified by a data structure named chromosome. In the basic or standard GA, the chromosomes are represented by a bit of string. Each bit is also named, a gene that represents the presence (value 1) or absence (value 0) of a specific characteristic in the individual.

At each generation, the individuals have evaluated their fitness to solve the problem. This evaluation is performed by a fitness function, which decodes the information contained in each individual chromosome into a measure of its quality. The evaluation of a chromosome is done to test its "fitness" as a solution. The fitness function plays a vital role of the environment in natural evolution by rating individuals in terms of their fitness. Selecting and formulating an appropriate fitness function are crucial to the efficient solution of any given GA problem. In our case, selecting the optimal samples (data points) in the initial population, which are the minority class, is set to HEOM. After evaluation, some individuals in the population are selected for reproduction, producing descendants, which will form a new population. This selection must privilege the fittest individuals, according to the natural selection principles.

In the reproduction of the selected individuals, their characteristics or genes are combined to obtain two descendants. This combination process is performed with the application of the crossover operator, which is a binary operator applied to two individuals. These individuals are named parents, and their chromosomes are combined to produce two new individuals, named offspring. For the bit-string representation, a common crossover operator is a one-point crossover. A second genetic operator usually applied is the mutation, which enforces a genetic variability in the new solutions. The boundary mutation alters genes from the individuals generated in the crossover step.

The procedures of population generation, evaluation of its individuals, selection, and application of the genetic operators are iterated, forming the basis of the GAs. Depending on the initial population, the GA may produce distinct solutions to the same problem. Therefore, the GA is usually run several times with different initial populations, and to stop the GA, other criteria can be used. For example, the GA may be when a maximum number of generations are reached.

## HEOMGA Method

HEOM measures the distance of a minority data point to all other minority data points in a population, which is the square root of their summation to produce the fitness scores for those data points in the population. HEOM acts as a fitness function for measuring similarity (distance) between the individuals (data points) in a population which contains all minority class samples in the training dataset to decide to use which data points. The two data points with smallest fitness scores produced from HEOM are selected as parents for mating, and then, the GA variants (crossover and mutation) are applied to produce offspring within the same iteration. Based on the three genetic operators and the evaluations, the better new populations of a candidate after the specified number of generations (e.g., number of generations = 5), the best solution (a newly generated data point) is formed and appended to the initial population. To start the next iteration, two data points with the smaller fitness scores in the updated





| Journal : Large 42979 | Article No : 850 | Pages : 12 | MS Code : 850 | Dispatch : 14-9-2021 |
|---|---|---|---|---|

population are selected as parents by returning the corresponding distance to each data point in the initial population in addition to the appended data points from the distances list produced in the previous iteration. The role of crossover and mutation operators then begins. This procedure will be repeated until the minority data points in the current population are equal to the number of majority data points in the original data set. Finally, to avoid the generation of newly duplicated data points, the algorithm will check and delete any duplicated ones. Figure 3 depicts the proposed method process.

SMOTE and ADASYN generate noise samples that have penetrated in the majority class region, resulting in an increase in overlapping. These noise samples are less useful, because they do not add any new information to the imbalance datasets, and they may lead to overfitting. It was confirmed that using the Euclidean distance metric that SMOTE and ADASYN use to measure the distance between two objects introduces some issues regarding imbalanced data and performance problems regarding computation or approximation of the square root [33]. Most datasets have both nominal and categorical attributes, and the major weakness of the Euclidean distance is that when some attributes have a large range of values as opposed to the remaining attributes, they may influence a bigger impact on the computed distance, while attributes

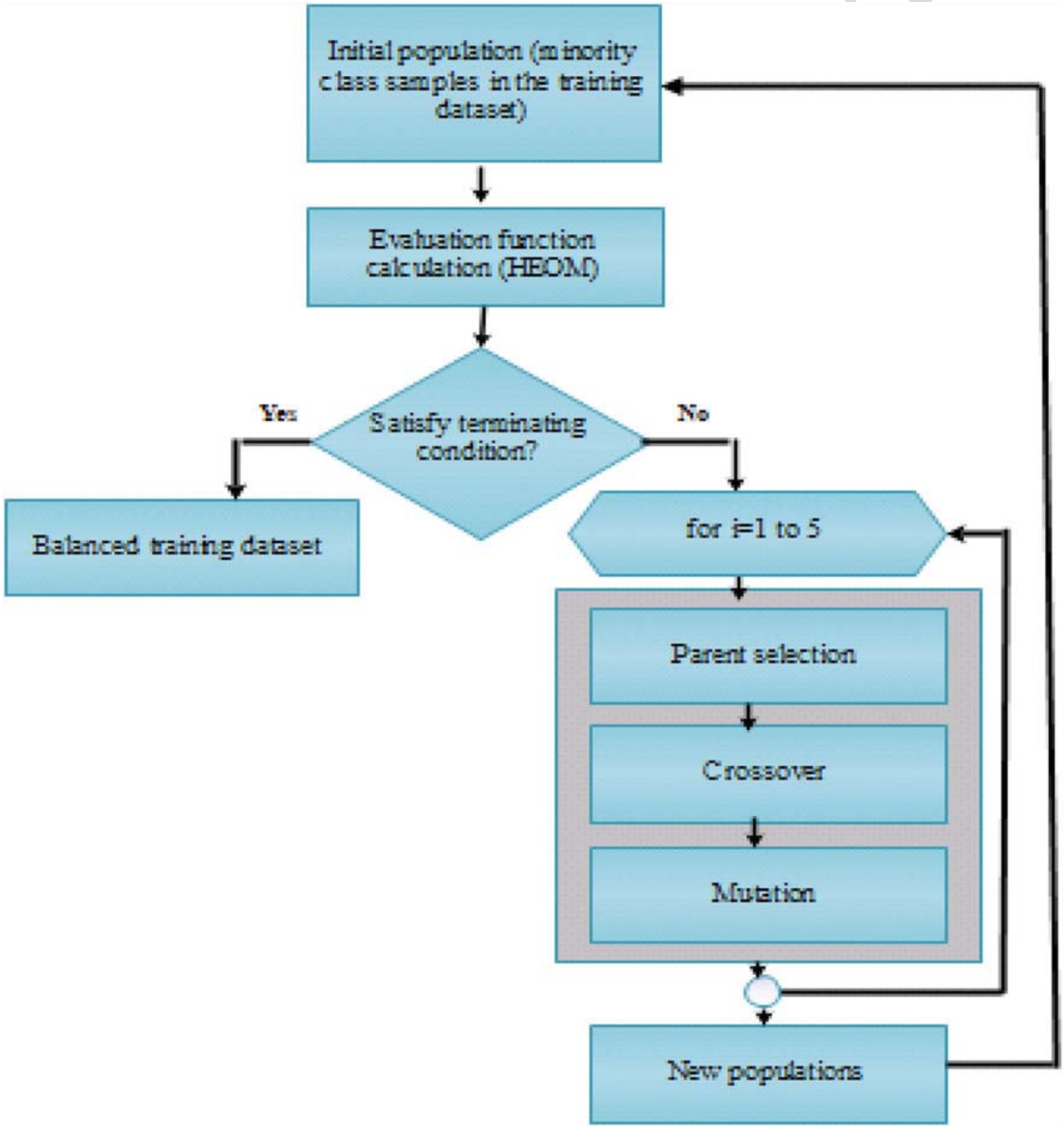


**Fig. 3** Basic structure of the HEOMGA method



| Journal : Large 42979 | Article No : 850 | Pages : 12 | MS Code : 850 | Dispatch : 14-9-2021 |
|---|---|---|---|---|

with a lower range of values will have a lesser impact on the results.

In the proposed method, all the minority data points are selected as the initial population, and the HEOM finds the distance between them by calculating the square root of their summation to produce the final fitness scores. In HEOM, normalized Euclidean distance is used for numeric features, and the overlap distance for categorical features is employed to find the distance between two instances $x_1$ and $x_2$ as provided in Eqs. (3 and 4). Applying the HEOM distance metric allows better handling of nominal and categorical attributes in accordance with the dataset nature. In addition, HEOM will help obtain better representation capability for minority data points and will enable us to appropriately select the data points that will be used as input for mating in the GA.

Crossover and mutation operators in the GA realize on the search exploration and exploitation, respectively. Exploration is the ability to create diversity in the population by exploring the search space, while exploitation is the reduction of diversity by focusing on individuals with higher fitness scores. Therefore, the newly generated synthetic data points will be produced in a safe region within the boundaries of the minority data points that are selected by the HEOM. As shown in Fig. 4, overlapping and overfitting problems will be somehow alleviated by causing the distance ($d$) between the generation area (the pink dotted oval) and the decision boundary to be larger and spread the newly generated data points far from the majority space (Table 1).

The use of crossover and mutation operators assists in improving the learning process by providing rich information about the newly generated data points, since they are inherited from the original data points, as shown in Fig. 5. This will make the learning process by a given learner easier. Finally, the HEOMGA will check and delete any duplicated data points during the generation process to avoid the generation of newly repeated samples.

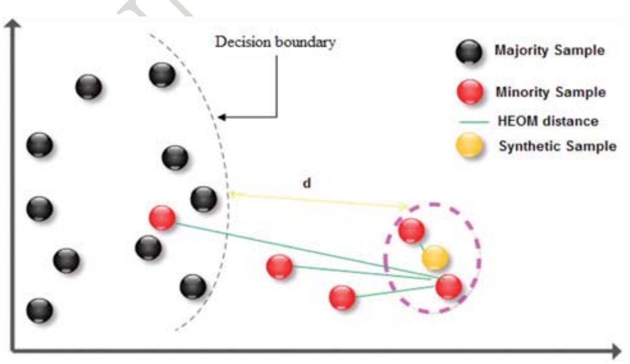


**Fig. 4** An example of how can HEOMGA avoid overlapping

## Experiment Design

### Datasets

A set of publicly available datasets for customer churn prediction are used in this work. Table 2 gives the details for each dataset. Evaluation of data mining and ML methods on publically available datasets offers different advantages [38]

- In terms of comparability of results, ranking methods, and evaluation of existing methods with new techniques
- Study the impact of the data and their characteristics on the performance of a technique
- Using available datasets provide insight into the effect of each phase of the followed methodology.

### Baseline Approaches and Learners

To examine the capability of the methods, three different learners are used: Decision Trees (DTs i.e., C4.5 algorithm), Bagging, and SVM with radial basis function kernel ($SVM_{rbf}$), due to their popularity with classification problems and their sensitivity to imbalance datasets. The DTs rely on greedy-search heuristics that checks one variable at a time [39], and therefore, it can attain a high level of accuracy by predicting the majority class, particularly if the majority class constitutes most of the dataset.

An SVM learner tries to find the hyper plane splitting instances of two classes based on the largest distance between them. It is useful mainly due to its capability to work in high feature space, since the learner can map complex nonlinear relationships between input and output with relatively high accuracy [40]. SVM with a radial basis function ($SVM_{rbf}$) kernel is used. Bagging is an ensemble learning learner, which has proved the ability to handle class imbalance problems effectively. The number of the nearest neighbors ($K$) parameter in both SMOTE and ADASYN was set to 5 [41].

Tenfold cross-validation is used to avoid picking particular parts that are for training and testing. The number of k was adjusted to 10; the data were split into ten parts; the procedure starts by splitting the dataset into 90% for training and 10% for testing. To finalize the process, the procedure was repeated ten times to allow each part of data being as testing data, and finally, the average results are considered for the used datasets on the ten partitions. Min–Max method is applied to transform training datasets into the range of 0 to $-$ 1, which means that all the





Journal : Large 42979 | Article No : 850 | Pages : 12 | MS Code : 850 | Dispatch : 14-9-2021

**Table 1** Description of imbalanced datasets characteristics

| Dataset source | Number of samples | Number of attributes | Minority (%) | Majority (%) | Imbalanced ratio |
|---|---|---|---|---|---|
| [a]Real world dataset[a] | 3333 | 21 | 14.49 | 85.51 | 5.90 |
| [b]Real world dataset[b] | 7043 | 21 | 26.54 | 73.46 | 2.77 |
| [c]Real world dataset[c] | 100,000 | 50 | 49.56 | 50.43 | 1.02 |

[a]http://www.sgi.com/tech/mlc/db/
[b]https://www.ibm.com/analytics/us/en/
[c]https://www.kaggle.com/abhinav89/telecom-customer/data

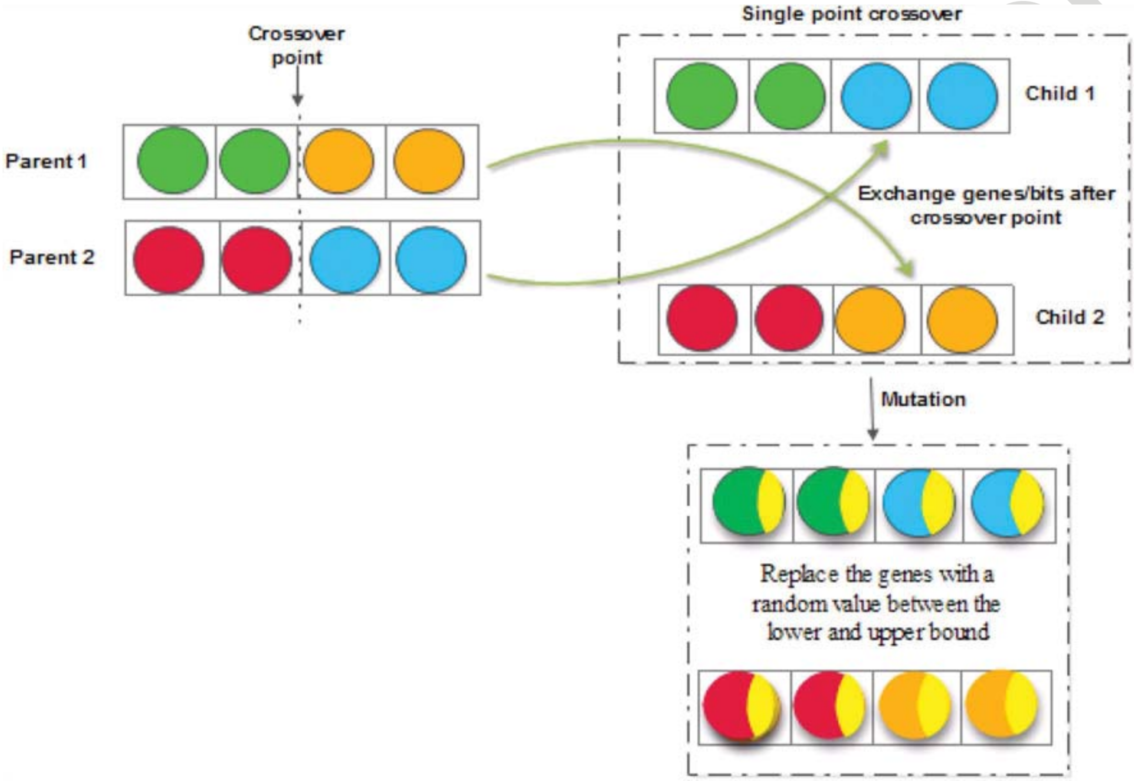


**Fig. 5** GA operators' processes



**Table 2** Confusion matrix for two-class problem

| | Actual | |
|---|---|---|
| | Churn customers | Non-churn customers |
| Predicted churn customers | TP | FP |
| Predicted non-churn customers | FN | TN |

values of numeric range of a feature are reduced to a scale between 0 and − 1 range. All the experiments are implemented using Python scikit-learn and the DTs $SVM_{rbf}$ and bagging learners are constructed based on the use of default parameters on Windows 7 with 2 Duo CPU running on 3.13 GHz PC with 44.25 GB RAM.

## Evaluation Metrics

To assess learners' results, a confusion matrix was employed to count: True Positive (TP) and True Negative (TN) denote the number of positive and negative examples that are classified correctly, while False Negative (FN) and False Positive (FP) represent a number of misclassified positive and negative examples, respectively. Table 2 shows a confusion matrix of a two-class problem. The first column of the table is the actual class label of the examples, and the first row presents their predicted class label.



Journal : Large 42979 | Article No : 850 | Pages : 12 | MS Code : 850 | Dispatch : 14-9-2021

The Recall is the True-Positive rate, which refers to the percentage of positive instances correctly predicted as positive class instances

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{5}$$

### Geometric Mean (G mean)

Gmean is a good indicator that can be used to assess the overall performance for a given learner, because it combines the learner's accuracy on the positive class and negative class samples. Therefore, a large value of this measure indicates that the learner performs well on both classes' samples

$$\text{Gmean} = \sqrt{\frac{\text{TP}}{\text{TP} + \text{FN}} \times \frac{\text{TN}}{\text{TN} + \text{FP}}}. \tag{6}$$

### Area Under Curve (AUC)

Receiver-Operating Curve (ROC) is usually known as AUC. The ROC graph plots true-positive rates versus false-positive rates. Learners can be selected based on their trade-off between true positives and false positives. Rather than visually comparing curves, the ROC metric aggregates the performance of classification methods into a single number, which makes it easier to compare the overall performance of different learners. This metric can also be applied to evaluate learning from imbalanced data. The bigger the AUC indicates, the better the generalization of the methods. The AUC can be determined as follows:

$$\text{AUC} = \frac{\left(1 + \frac{\text{TP}}{\text{TP} + \text{FN}} - \frac{\text{FP}}{\text{FP} + \text{TN}}\right)}{2}. \tag{7}$$

The above evaluation metrics can reasonably evaluate the learning process from imbalanced datasets, since their formulae are relative to the rare class, which is in our case the churn class. These measurements are used to evaluate the proposed method and its effectiveness to overcome class imbalance.

## Results and Discussion

The performance of three learners without using any balancing method (i.e., 0% balancing) and the results of the proposed method against SMOTE, ADASYN, G-SMOTE [42], and Gaussian method [43] were applied over three customer churn datasets to study the impact of different balancing technique on the evaluation measures used in this work. The results are summarized in Tables 3, 4, and 5.

**Table 3** DTs results based on the evaluation metrics for all the datasets

| Dataset | Method | Recall | G mean | AUC |
|---|---|---|---|---|
| Dataset 1 | 0% balancing | 0.780 | 0.852 | 0.856 |
| | SMOTE | 0.741 | 0.798 | 0.833 |
| | ADASYN | 0.725 | 0.802 | 0.812 |
| | Proposed method | **0.926** | **0.944** | **0.944** |
| | G-SMOTE | 0.758 | 0.841 | 0.847 |
| | Gaussian method | 0. 852 | 0.921 | 0.924 |
| Dataset 2 | 0% balancing | 0.481 | 0.626 | 0.648 |
| | SMOTE | 0.559 | 0.659 | 0.684 |
| | ADASYN | 0.500 | 0.635 | 0.673 |
| | Proposed method | **0.816** | **0.816** | **0.818** |
| | G-SMOTE | 0.609 | 0.716 | 0.789 |
| | Gaussian method | 0.734 | 0.804 | 0.808 |
| Dataset 3 | 0% balancing | 0.522 | 0.523 | 0.523 |
| | SMOTE | 0.466 | 0.544 | 0.551 |
| | ADASYN | 0.464 | 0.540 | 0.546 |
| | Proposed method | **0.523** | **0.552** | **0.554** |
| | G-SMOTE | 0.473 | 0.536 | 0.541 |
| | Gaussian method | 0.478 | 0.539 | 0.543 |

The best result of each dataset is emphasized in bold

**Table 4** SVM results based on the evaluation metrics for all the datasets

| Dataset | Method | Recall | G mean | AUC |
|---|---|---|---|---|
| Dataset 1 | 0% balancing | 0.219 | 0.466 | 0.724 |
| | SMOTE | 0.814 | 0.816 | 0.817 |
| | ADASYN | 0.676 | 0.739 | 0.739 |
| | Proposed method | **0.845** | **0.919** | **0.919** |
| | G-SMOTE | **0.845** | **0.919** | 0.918 |
| | Gaussian method | 0.837 | 0.913 | 0.914 |
| Dataset 2 | 0% balancing | 0.484 | 0.662 | 0.734 |
| | SMOTE | 0.674 | 0.740 | 0.752 |
| | ADASYN | 0.636 | 0.725 | 0.738 |
| | Proposed method | **0.815** | **0.846** | **0.847** |
| | G SMOTE | 0.681 | 0.748 | 0.748 |
| | Gaussian method | 0.633 | 0.792 | 0.792 |
| Dataset 3 | 0% balancing | 0.443 | 0.537 | 0.547 |
| | SMOTE | 0.395 | 0.524 | 0.545 |
| | ADASYN | 0.402 | 0.535 | 0.544 |
| | Proposed method | **0.502** | **0.557** | **0.560** |
| | G SMOTE | 0.500 | 0.529 | 0.530 |
| | Gaussian method | 0.498 | 0.551 | 0.553 |

The best result of each dataset is emphasized in bold

Tables 3, 4, 5 show that HEOMGA performs better than0% balancing, SMOTE, ANDSYN, G SMOTE, and Gaussian method in term of Recall for all the used datasets.





| Journal : Large 42979 | Article No : 850 | Pages : 12 | MS Code : 850 | Dispatch : 14-9-2021 |
|---|---|---|---|---|

**Table 5** Bagging results based on the evaluation metrics for all the datasets

| Dataset | Method | Recall | G mean | AUC |
|---|---|---|---|---|
| Dataset 1 | 0% balancing | 0.137 | 0.371 | 0.591 |
| | SMOTE | 0.131 | 0.362 | 0.581 |
| | ADASYN | 0.098 | 0.313 | 0.545 |
| | Proposed method | **0.875** | **0.934** | **0.934** |
| | G SMOTE | 0.762 | 0.862 | 0.876 |
| | Gaussian method | 0.867 | 0.927 | 0.928 |
| Dataset 2 | 0% balancing | 0.455 | 0.646 | 0.794 |
| | SMOTE | 0.724 | 0.745 | 0.786 |
| | ADASYN | 0.534 | 0.680 | 0.733 |
| | Proposed method | **0.776** | **0.849** | **0.853** |
| | G SMOTE | 0.554 | 0.678 | 0.796 |
| | Gaussian method | 0.651 | 0.801 | 0.802 |
| Dataset 3 | 0% balancing | 0.418 | 0.521 | 0.533 |
| | SMOTE | 0.416 | 0.513 | 0.524 |
| | ADASYN | 0.413 | 0.512 | 0.523 |
| | Proposed method | **0.480** | **0.537** | **0.540** |
| | G SMOTE | 0.428 | 0.529 | 0.538 |
| | Gaussian method | 0.437 | 0.523 | 0.539 |

The best result of each dataset is emphasized in bold

Therefore, an improvement in the churn rate is achieved by the proposed methods among the other used oversampling methods.

The bigger the AUC and G mean indicate the better the generalization of the methods. Empirical experiment results indicated that the proposed method outperforms the tested oversampling methods in terms of G mean and AUC for the datasets. The proposed method for the three datasets obtained the best G mean and AUC values compared to other methods. This can be explained by the fact that the use of the proposed method provides rich information to the learners, which in turn improve prediction results and the learning process.

The receiver-operating characteristic (ROC) graph calculates the learner performance by changing the DTs' confidence level, $SVM_{rbf}$, and Bagging scores to get distinct values of $TP_{rate}$ and $FP_{rate}$, as shown in Figs. 6, 7, and 8.

The lines of the proposed method in Figs. 6, 7, and 8 is closer to the left-hand border and the top border compared to 0% balancing, SMOTE, ANDSYN, G SMOTE, and Gaussian method. This indicates that the proposed method offers the finest results among the other methods for class imbalance problems in the application of customer churn prediction.

For further check the statistical significance of the proposed method and whether it significantly outperforming the other used oversampling algorithms in terms of Recall, G mean, and AUC, Wilcoxon signed-rank test [44] is performed. The results of the test are provided in Tables 6, 7 and 8. The test's confidence level is set 0.05, given the null hypothesis that the learners' performance varies significantly across the various algorithms and evaluation metrics with the proposed method as a control algorithm.

The test results in terms of Recall, G mean, and AUC are given in Tables 6, 7 and 8 to validate the proposed method significantly outperforms 0% balancing, SMOTE, ADASYN, G SMOTE, and Gaussian method.

## Conclusion and Future Work

This work proposes an effective preprocessing approach, called HEOMGA, to overcome class imbalance issues and assist the learners in improving their generalization capacity and performance. This work has conducted a set of experiments on publicly available customer churn prediction datasets to assess the performance of the proposed method. Experimental results showed the efficiency of the proposed method as compared to the other tested oversampling methods. Moreover, the proposed HEOMGA method significantly outperformed the other oversampling



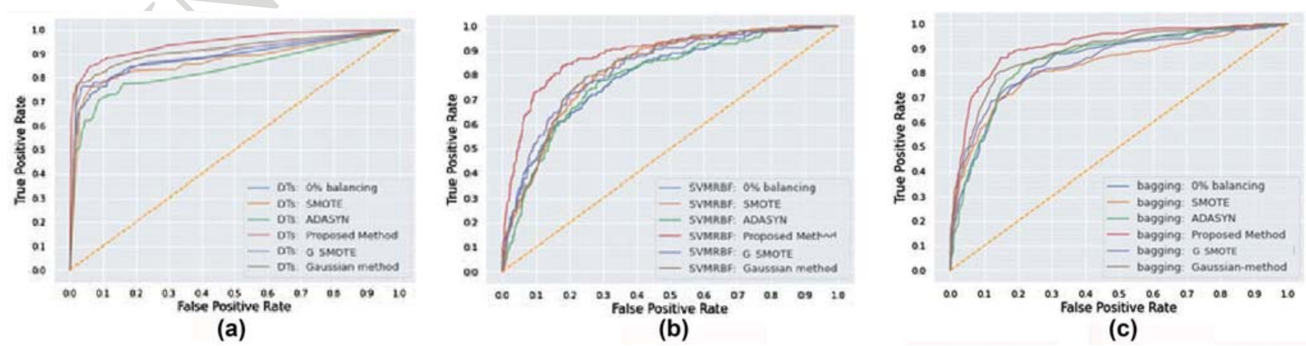


**Fig. 6** ROC curve comparison among 0% balancing, SMOTE, ANDSYN, proposed method, G SMOTE, and Gaussian method for dataset 1 using **a** DTs, **b** $SVM_{rbf}$, and **c** Bagging



Journal : Large 42979 | Article No : 850 | Pages : 12 | MS Code : 850 | Dispatch : 14-9-2021

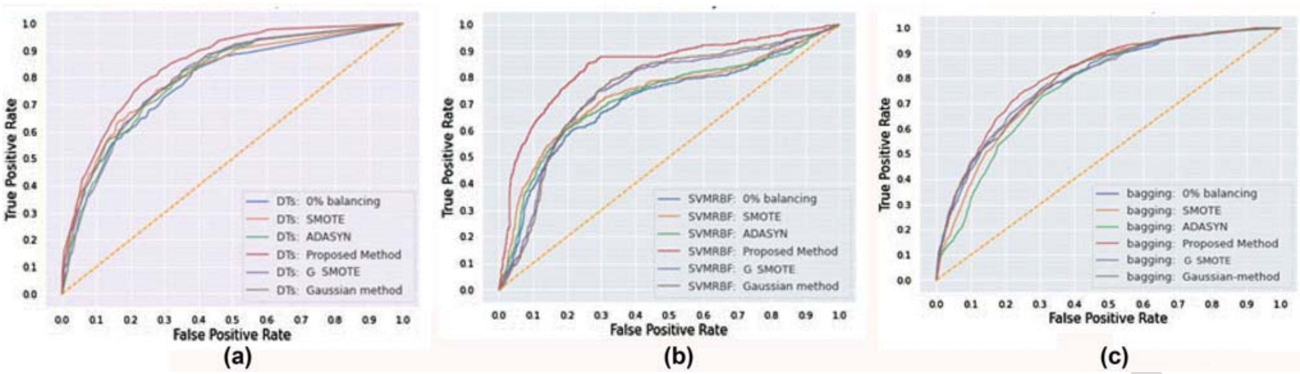


**Fig. 7** ROC curve comparison among 0% balancing, SMOTE, ANDSYN, proposed method, G SMOTE, and Gaussian method for dataset 2 using **a** DTs, **b** $SVM_{rbf}$, and **c** Bagging

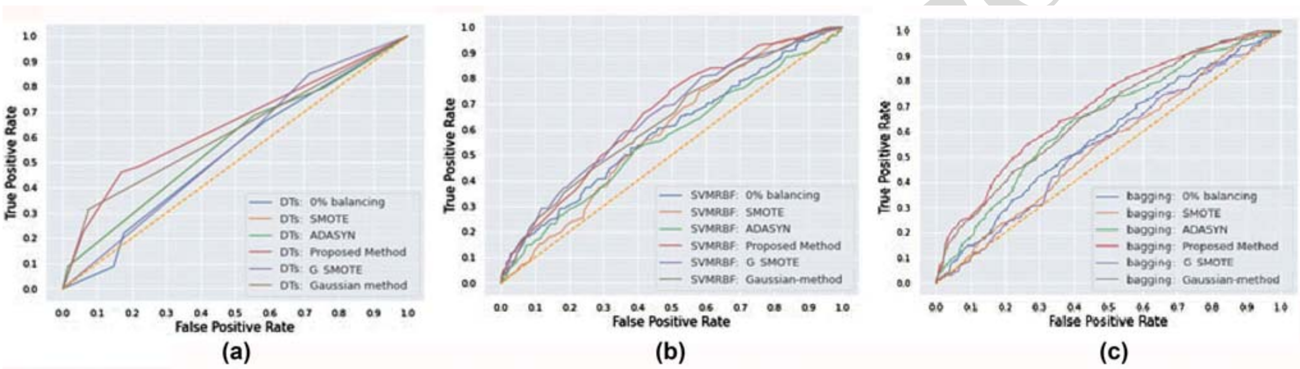


**Fig. 8** ROC curve comparison among 0% balancing, SMOTE, ANDSYN, proposed method, G SMOTE, and Gaussian method for dataset 3 using **a** DTs, **b** $SVM_{rbf}$, and **c** Bagging



**Table 6** Wilcoxon signed-rank test evaluation results based on Recall

| Comparison | *p* value | *W* value | Mean difference | $R^+$ | $R^-$ | *Z*-value | Mean (*W*) | Std (*W*) | Significance |
|---|---|---|---|---|---|---|---|---|---|
| Proposed method vs. 0% balancing | 0.05 | 10 | 0.50 | 455 | 10 | − 4.5765 | 232.5 | 48.62 | + |
| Proposed method vs. SMOTE | 0.05 | 1 | − 0.10 | 464 | 1 | 4.7616 | 232.5 | 48.62 | + |
| Proposed method vs. ADASYN | 0.05 | 0 | 0.04 | 465 | 0 | − 4.7821 | 232.5 | 48.62 | + |
| Proposed method vs. G-SMOTE | 0.05 | 0 | − 0.13 | 435 | 0 | − 4.703 | 217.5 | 46.25 | + |
| Proposed method vs. Gaussian Method | 0.05 | 0 | − 0.13 | 406 | 0 | − 4.6226 | 203 | 43.91 | + |

$R^+$ is the sum of ranks for the datasets in which the first method outperforms the second and $R^-$ is the sum of ranks of the opposite, Std is standard deviation (*W*), and + refers to significance at 0.05 level

methods in terms of Recall, G mean, and AUC based on the Wilcoxon signed-rank test analysis. In the future, it would be interesting to see the results of the proposed HEOMGA in conjunction with applying feature selection methods. Another research direction can be to test other distance metrics to tackle the class imbalance, and finally, another line of future research would be to try to tackle class overlap situations.



| Journal : Large 42979 | Article No : 850 | Pages : 12 | MS Code : 850 | Dispatch : 14-9-2021 |
|---|---|---|---|---|

**Table 7** Wilcoxon signed-rank test evaluation results based on G mean

| Comparison | *p* value | *W* value | Mean difference | $R^+$ | $R^-$ | *Z*-value | Mean (*W*) | Std (*W*) | Significance |
|---|---|---|---|---|---|---|---|---|---|
| Proposed method vs. 0% balancing | 0.05 | 0 | 0.3 | 465 | 0 | − 4.7821 | 232.5 | 48.62 | + |
| Proposed method vs. SMOTE | 0.05 | 1 | − 0.05 | 464 | 1 | − 4.7616 | 232.5 | 48.62 | + |
| Proposed method vs. ADASYN | 0.05 | 1 | 0.03 | 464 | 1 | − 4.7616 | 232.5 | 48.62 | + |
| Proposed method vs. G-SMOTE | 0.05 | 8.5 | − 0.15 | 426.5 | 8.5 | − 4.5192 | 217.5 | 46.25 | + |
| Proposed method vs. Gaussian Method | 0.05 | 20 | − 0.14 | 445 | 20 | − 4.3708 | 232.5 | 48.62 | + |

$R^+$ is the sum of ranks for the datasets in which the first method outperforms the second and $R^-$ is the sum of ranks of the opposite, Std is standard deviation (*W*), and + refers to significance at 0.05 level

**Table 8** Wilcoxon signed-rank test evaluation results based on AUC

| Comparison | *p* value | *W* value | Mean difference | $R^+$ | $R^-$ | *Z*-value | Mean (*W*) | Std (*W*) | Significance |
|---|---|---|---|---|---|---|---|---|---|
| Proposed method vs. 0% balancing | 0.05 | 0 | 0.05 | 465 | 0 | − 4.7821 | 232.5 | 48.62 | + |
| Proposed method vs. SMOTE | 0.05 | 0 | − 0.04 | 465 | 0 | − 4.7821 | 232.5 | 48.62 | + |
| Proposed method vs. ADASYN | 0.05 | 0 | 0.04 | 465 | 0 | − 4.7821 | 232.5 | 48.62 | + |
| Proposed method vs. G-SMOTE | 0.05 | 0 | − 0.15 | 435 | 0 | − 4.703 | 217.5 | 46.25 | + |
| Proposed method vs. Gaussian Method | 0.05 | 0 | − 0.14 | 435 | 0 | − 4.703 | 217.5 | 46.25 | + |

$R^+$ is the sum of ranks for the datasets in which the first method outperforms the second and $R^-$ is the sum of ranks of the opposite, Std is standard deviation (*W*), and + refers to significance at 0.05 level

**Funding** No funding sources.

## Declarations

**Conflict of interest** Authors have declared that no conflict of interest exists.

## References


1. Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: a review. Int J Pattern Recogn Artif Intell. 2009;23(04):687–719.
2. Chen Z, Yan Q, Han H, Wang S, Peng L, Wang L, Yang B. Machine learning based mobile malware detection using highly imbalanced network traffic. Inf Sci. 2018;433:346–64.
3. Jain A, Ratnoo S, Kumar D (2017) Addressing class imbalance problem in medical diagnosis: a genetic algorithm approach. In: 2017 international conference on information, communication, instrumentation and control (ICICIC) (pp. 1–8), IEEE
4. Ramli NA, Ismail MT, Wooi HC. Measuring the accuracy of currency crisis prediction with combined classifiers in designing early warning system. Mach Learn. 2015;101(1–3):85–103.
5. Dwiyanti E, Ardiyanti A (2016) Handling imbalanced data in churn prediction using rusboost and feature selection (case study: Pt. telekomunikasiindonesia regional 7). In: International conference on soft computing and data mining (pp 376–385). Springer, Cham
6. He B, Shi Y, Wan Q, Zhao X. Prediction of customer attrition of commercial banks based on SVM model. Procedia Comput Sci. 2014;31:423–30.
7. Huang PJ (2015) Classication of imbalanced data using synthetic over-sampling techniques, Doctoral dissertation, University of California
8. Chawla NV (2009) Data mining for imbalanced datasets: an overview. In: Data mining and knowledge discovery handbook (pp 875–886). Springer, Boston
9. Burez J, Van den Poel D. Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009;36(3):4626–36.
10. Amin A, Al-Obeidat F, Shah B, Adnan A, Loo J, Anwar S. Customer churn prediction in telecommunication industry using data certainty. J Bus Res. 2019;94:290–301.
11. Chawla NV, Japkowicz N, Kotcz A. Special issue on learning from imbalanced data sets. ACM SIGKDD Explor Newsl. 2004;6(1):1–6.
12. Liu XY, Wu J, Zhou ZH (2009) Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern Part B Cybern 39(2):539–550
13. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
14. He H, Bai Y, Garcia EA, Li S (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: Neural networks, 2008. IJCNN 2008 (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on (pp 1322–1328), IEEE
15. Han H, Wang WY, Mao BH (2005) Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing (pp 878–887). Springer, Berlin, Heidelberg
16. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Adv Knowl Discov Data Min. 2009;2009:475–82.






Journal : Large 42979 | Article No : 850 | Pages : 12 | MS Code : 850 | Dispatch : 14-9-2021

17. Maciejewski T, Stefanowski J (2011) Local neighbourhood extension of SMOTE for mining imbalanced data. In: Computational intelligence and data mining (CIDM), 2011 IEEE symposium on (pp 104–111), IEEE
18. Barua S, Islam MM, Yao X, Murase K. MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans Knowl Data Eng. 2014;26(2):405–25.
19. Zhu B, Broucke S, Baesens B, Maldonado S (2017) improving resampling-based ensemble in churn prediction. In: First international workshop on learning with imbalanced domains: theory and applications, pp 79–91
20. Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hussain A, et al. Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. IEEE Access. 2016;4:7940–57.
21. Salunkhe UR, Mali SN. A hybrid approach for class imbalance problem in customer churn prediction: a novel extension to undersampling. Int J Intell Syst Appl. 2018;10(5):71.
22. Zou S, Huang Y, Wang Y, Wang J, Zhou C (2008) SVM learning from imbalanced data by GA sampling for protein domain prediction. In: 2008 the 9th international conference for young computer scientists (pp 982–987), IEEE
23. Haque MN, Noman N, Berretta R, Moscato P. Heterogeneous ensemble combination search using genetic algorithm for class imbalanced data classification. PLoS ONE. 2016;11:1.
24. Cervantes J, Li X, Yu W (2013) Using genetic algorithm to improve classification accuracy on imbalanced data. In: 2013 IEEE international conference on systems, man, and cybernetics (pp 2659–2664), IEEE
25. Jiang K, Lu J, Xia K. A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE. Arab J Sci Eng. 2016;41(8):3255–66.
26. Karia V, Zhang W, Naeim A, Ramezani R (2019) GenSample: a genetic algorithm for oversampling in imbalanced datasets. arXiv: 1910.10806
27. Mahin M, Islam MJ, Khatun A, Debnath BC (2018) A comparative study of distance metric learning to find sub-categories of minority class from imbalance data. In: 2018 international conference on innovation in engineering and technology (ICIET) (pp 1–6), IEEE
28. El Hindi K. Specific-class distance measures for nominal attributes. AI Commun. 2013;26(3):261–79.
29. Li C, Li H. A survey of distance metrics for nominal attributes. J Softw. 2010;5(11):1262–9.
30. Wilson DR, Martinez TR. Improved heterogeneous distance functions. J Artif Intell Res. 1997;6:1–34.
31. Mahin M, Islam MJ, Debnath BC, Khatun A (2019) Tuning distance metrics and K to find sub-categories of minority class from imbalance data using K nearest neighbours. In: 2019 international conference on electrical, computer and communication engineering (ECCE) (pp 1–6), IEEE
32. Guo H, Viktor HL. Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. ACM SIGKDD Explor Newsl. 2004;6(1):30–9.
33. Liu Y, Yu X, Huang JX, An A. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Inf Process Manage. 2011;47(4):617–31.
34. Santos MS, Abreu PH, García-Laencina PJ, Simão A, Carvalho A. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J Biomed Inform. 2015;58:49–59.
35. Kagie M, van Wezel M, Groenen PJ (2009) An empirical comparison of dissimilarity measures for recommender systems
36. Tsymbal A, Pechenizkiy M, Cunningham P (2006) Dynamic integration with random forests. In: European conference on machine learning, pp 801–808. Springer, Berlin, Heidelberg
37. El-Sappagh S, Elmogy M, Ali F, Abuhmed T, Islam SM, Kwak KS. A comprehensive medical decision-support framework based on a heterogeneous ensemble classifier for diabetes prediction. Electronics. 2019;8(6):635.
38. Vandecruys O, Martens D, Baesens B, Mues C, De Backer M, Haesen R. Mining software repositories for comprehensible software fault prediction models. J Syst Softw. 2008;81(5):823–39.
39. Rokach L, Maimon OZ (2008) Data mining with decision trees: theory and applications (vol 69). World scientific
40. Das B, Krishnan NC, Cook DJ (2013) Handling class overlap and imbalance to detect prompt situations in smart homes. In: 2013 IEEE 13th international conference on data mining workshops, pp 266–273, IEEE
41. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2008;9:1263–84.
42. Douzas G, Bacao F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. Inf Sci. 2019;501:118–35.
43. Zhang H, Wang Z (2011) A normal distribution-based over-sampling approach to imbalanced data classification. In: International conference on advanced data mining and applications, pp 83–96. Springer, Berlin, Heidelberg
44. García S, Molina D, Lozano M, Herrera F. A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: a case study on the CEC'2005 special session on real parameter optimization. J Heuristics. 2009;15(6):617.


**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.





| Journal : Large 42979 | Article No : 850 | Pages : 12 | MS Code : 850 | Dispatch : 14-9-2021 |
|---|---|---|---|---|

Author Proof

| | |
|---|---|
| Journal: | **42979** |
| Article: | **850** |

# Author Query Form

**Please ensure you fill out your response to the queries raised below and return this form along with your corrections**

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

| Query | Details Required | Author's Response |
|---|---|---|
| AQ1 | Please check and confirm the inserted citation of Table [1] is correct. If not, please suggest an alternative citation. Please note that figures and tables should be cited in sequential order in the text. | |