AIMS Mathematics, 8(4): 9400–9415.
DOI: 10.3934/math.2023473
Received: 10 October 2022
Revised: 30 January 2023
Accepted: 30 January 2023
Published: 15 February 2023

http://www.aimspress.com/journal/Math

Research article

A new hybrid approach based on genetic algorithm and support vector machine methods for hyperparameter optimization in synthetic minority over-sampling technique (SMOTE)

Pelin Akın*

Faculty of Science, Department of Statistics, Cankiri Karatekin University, Cankiri, Turkey

* **Correspondence:** Email: pelinakin@karatekin.edu.tr.

Abstract: The crucial problem when applying classification algorithms is unequal classes. An imbalanced dataset problem means, particularly in a two-class dataset, that the group variable of one class is comparatively more dominant than the group variable of the other class. The issue stems from the fact that the majority class dominates the minority class. The synthetic minority over-sampling technique (SMOTE) has been developed to deal with the classification of imbalanced datasets. SMOTE algorithm increases the number of samples by interpolating between the clustered minority samples. The SMOTE algorithm has three critical parameters, "k", "perc.over", and "perc.under". "perc.over" and "perc.under" hyperparameters allow determining the minority and majority class ratios. The "k" parameter is the number of nearest neighbors used to create new minority class instances. Finding the best parameter value in the SMOTE algorithm is complicated. A hybridized version of genetic algorithm (GA) and support vector machine (SVM) approaches was suggested to address this issue for selecting SMOTE algorithm parameters. Three scenarios were created. Scenario 1 shows the evaluation of support vector machine SVM) results without using the SMOTE algorithm. Scenario 2 shows that the SVM was used after applying SMOTE algorithm without the GA algorithm. In the third scenario, the results were analyzed using the SVM algorithm after selecting the SMOTE algorithm's optimization method. This study used two imbalanced datasets, drug use and simulation data. After, the results were compared with model performance metrics. When the model performance metrics results are examined, the results of the third scenario reach the highest performance. As a result of this study, it has been shown that a genetic algorithm can optimize class ratios and k hyperparameters to improve the performance of the SMOTE algorithm.

Keywords: SMOTE; genetic algorithm; support vector machine
Mathematics Subject Classification: 68Txx, 68w01, 68w40, 68wxx, 82-08

1. Introduction

Imbalanced datasets are the most prevalent problem in classification algorithms. An imbalanced dataset problem means, particularly in a two-class dataset, that the group variable of one class is comparatively more dominant than the group variable of the other class. In such a case, the group with fewer observations is referred to as the "minority class", whereas the other group with significantly more observations is referred to as the "majority class". When the model is inadequately trained with samples that have limited label data, it produces inaccurate estimations during the estimation process. The issue stems from the majority class dominating the minority class. When the applied model is biased toward the class with the majority of the observations, it results in the inaccurate classification of the minority class. The problem of unbalanced data is growing in importance in real-world domains. Examples are high-resolution aerial images, remote sensing, facial recognition, and medical diagnosis [1].

The problem encountered in these datasets is the calculation of the performance metrics of the algorithms. Accuracy, one of the performance metrics, is the main principle the classification algorithms function under, and they primarily aim to keep general errors, over which minority class has little influence, to the minimum. Classification algorithms assume that the data distribution is uniform across all classes and that the errors stemming from different classes are the same [2]. As a result of that, they perform poorly with imbalanced datasets. In other words, since they assume that the dataset is balanced, most data mining algorithms produce degenerate models that take no account of minority classes. Different imbalanced techniques have been developed to eliminate this problem. Applying the synthetic minority over-sampling technique (SMOTE) algorithm is one way to overcome this problem. The SMOTE-based sampling method is predicated on an over-sampling method proposed by Chawla et al. SMOTE algorithm increases the number of samples by interpolating between the clustered minority samples [3]. The SMOTE algorithm has three parameters, the first of which, "k" is the number of the closest neighbors used to generate new samples of the minority class. The second parameter, "perc.over", is the number that determines the number of additional cases to be generated from the minority classes. The last parameter, "perc.under" decides how many extra cases are from the majority classes for each case generated from the minority class. Accurate selection of the parameters mentioned above is of vital importance for the SMOTE algorithm to function correctly. If the SMOTE algorithm parameters are selected inaccurately, the ensuing problems one may face include over-training and underperformance of the algorithm. The optimization method was used to estimate these parameters.

Optimization is finding the solution that gives the best result in the solution space of a problem. Mathematical and heuristic techniques are used in solving optimization problems. Most metaheuristic algorithms are adapted from biological evolutionary processes, swarm behavior, and the law of physics and fall into two categories: single-solution and population-based [4,5]. Single-solution-based algorithms use a single candidate solution and develop this solution using local search [6]. Some recently developed single-solution-based metaheuristic methods are simulated annealing, tabu search (TS), microcanonical annealing (MA), and guided local search (GLS). Population-based metaheuristics utilize multiple candidate solutions during the search process [4]. Some population-

based metaheuristic algorithms are genetic algorithm (GA), particle swarm optimization (PSO), and ant colony optimization (ACO).

GA mimics the Darwinian theory of survival of the fittest in nature. J.H. Holland proposed GA in 1992. The essential elements of GA are chromosome representation, fitness selection, and biological-inspired operators [7]. The genetic algorithm from population-based models is used because all data were considered while optimizing SMOTE parameters. The purpose of choosing the genetic algorithm can be summed up into four items. Firstly, it is the most widely used optimization strategy [8]. Second, it is relatively easy to implement, and there is much flexibility in setting up the algorithm so that it can be applied to various problems [9]. Thirdly, genetic algorithms are often used for search-based optimization problems that are difficult and time-intensive to solve by other general algorithms. Lastly, Genetic Algorithms are faster and more efficient when compared to traditional brute-force search methods. Genetic Algorithms have been proven to have many parallel capabilities [10]. The studies in recent years have been examined in the following paragraphs.

The imbalanced dataset issues have become more prevalent in recent years. According to Nimankar and D. Vora [11], the proposed system solved the class imbalance problem by balancing the unbalanced class with SLS (Safe-Level Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic Sampling Approach. KNN (K-Nearest Neighbors) and SVM (Support Vector Machine) algorithms whose parameters were selected by Cuckoo Search optimization were applied to the balanced data set. According to the results of the analysis, it has been seen that this optimization technique improves performance. Jiang et al. [12] suggest GASMOTE, a new genetic-based SMOTE algorithm that takes advantage of different sampling rates for each minority class sample. Risk prediction in rockburst instance data was used for analysis. They compared the results after creating decision tree algorithms. The results show that the prediction accuracy of analysis is significantly improved. Obiedat et al. [13] proposed a new hybrid model to analyze the impact on sentiment toward various restaurants in Jordan. They solved the imbalanced dataset problem with Smote techniques. Afterward, the SVM algorithm with particle swarm optimization (PSO) was combined, and classification methods were compared. According to this study, their proposed combined PSO-SVM approach produces the best results in terms of accuracy, f-measure, G-mean, and AUC for different versions of the datasets when compared to other classification methods. Wang [14] proposes an ensemble model for estimating imbalanced credit risk and developed algorithms for the attributes of financial datasets in their study. In principle, the ensemble model incorporates SMOTE and Multi-Kernel Fuzzy C-optimized by PSO. The k parameter, the neighborhood value in SMOTE, is essential when choosing a sample. They claim that specifying this parameter with PSO will eliminate the overfitting problem in the algorithm. According to the results produced by the Matlab software, the proposed ensemble model produces the best performance in estimating imbalanced credit risk. Generally, the proposed ensemble model outperforms the competition in predicting business credit risk. Demidova and Klyueva [15], when addressing the SVM classification of imbalanced data clusters, employed PSO that was oriented toward selecting optimum parameter values for the SMOTE algorithm. They found optimal values by optimizing two SMOTE parameters, k (number of neighbors) and m (nearest neighbors, which is used to determine if the minority object is in danger). They analyzed the classification results in comparison on the base of the SVM classifier without SMOTE algorithm and with the proposed algorithm for two data. According to the results of the study, the hybrid SMOTE algorithm shows that it improves the classification quality. Sreejith et al. [16] proposed a framework

for classifying unbalanced datasets by combining Chaotic Multi-Verse Optimization and SMOTE. They set up the hybrid model to better select the nearest neighbor (k). As a result of this study, the proposed method was superior. Wang and Cheng [17] suggested multiple combined methods to rebalance medical data featuring class imbalances. The methods proposed in the study created various combinations of SMOTE, PSO-based attribute selection, and Meta cost. As a result of their proposed method, it has been shown that it can effectively improve class imbalance performance. Zorić et al. [18] proposed a model to select the SMOTE hyperparameters as the number of synthetic samples (N) and the number of neighbors (k). They used the artificial bee colony, bat algorithm and particle swarm algorithm, differential evolution, and Nelder-Mead algorithm. They mentioned that when optimization techniques are compared, they may vary according to the dataset. However, they found that PSO and ABC algorithms have a higher convergence rate. Sara et al. [19] are to develop models for predicting bug prone using SMOTE for balancing datasets and grid search to tune the hyperparameters of techniques. They applied five machine-learning techniques. Grid search performance turned out to be better performance accuracy than default settings. Ren et al. [20] proposed a combined model (GA, AdaBoost, Random forest) to predict oil temperature in tunnel boring machines. With this model, they optimized the number and depth of trees. When the study results are examined, it has a better prediction performance than traditional machine learning algorithms. Yuan et al. [21] improved the efficiency and accuracy of solving engineering optimization problems and proposed EOBLS and CKGS improved the GWO algorithm, called EOCSGWO. The performance of EOCSGWO is compared with meta-heuristic optimization algorithms. The results show that EOCSGWO ranks first among other optimization algorithms in accuracy and robustness. Shi et al. [22] proposed a new multi-fidelity model based on support vector regression (coSVR). The kernel function was utilized to map the discrepancy between HF and LF responses into the high-dimensional (or infinite-dimensional) feature space. Co_SVR also performs better than the other models for both numerical and engineering cases.

This study utilizes two imbalanced datasets, drug use and simulation data. SMOTE algorithm was used to ensure the contribution of the minority class and to increase the classification algorithm's success rate, which is one of the most common problems with imbalanced datasets. Problems arise when choosing SMOTE hyperparameters. If optimal hyperparameters are used, the best result is achieved. For this reason, a hybridized version of GA and SVM approaches was developed for selecting the “perc.over”, “perc.under”, and “k” hyperparameters in the SMOTE algorithm.

The flow of the hybrid model is calculated by the genetic algorithm and the values of these three hyperparameters. Different hyperparameters are obtained. The Smote algorithm is applied with these hyperparameters, and the data becomes balanced. The SVM algorithm is applied to this data set, and the gmean value is calculated. These values are compared, and optimal coefficients are reached with the highest gmean value. To compare the hybrid model, three scenarios were set up. The first scenario's purpose is to see the results obtained when the SMOTE algorithm is not applied to unbalanced data. Installing the second scenario is to see the results of the SMOTE algorithm without optimization. The third scenario includes the results of the hybrid model.

In previous studies, the minority class was balanced by producing synthetic data according to the majority class count. This issue causes overfitting problems in machine learning. This study aims to balance the majority and minority by determining a collective and optimal sample number. The article is divided as follows: In Section 2, GA, SMOTE, SVM, and hybrid models are defined. Section 3 explains the application of the analysis results. Finally, a brief discussion is given in Section 4.

2. Materials and methods

2.1. Genetic algorithm optimization

The genetic algorithm (GA) is a method that works similarly to the evolutionary process observed in nature. The genetic algorithm was first applied to optimization problems by John Holland [23]. The cycle of the genetic algorithm called generations is as follows [24]:

- The coefficients (genes) forming for each member (chromosome) of the population take on their new values through various evolutionary processes.
- The population is reproduced in a series of iterations. One or more parents are stochastically selected, but sequences with higher fitness values are more likely to contribute to an offspring.
- Genetic operators such as crossover and mutation are applied to the parents to produce offspring.
- Offspring are added to the population, and the process is repeated.

The various advantages of GA that have made it popular lately are that it is applicable, faster, and more efficient. It also optimizes both continuous and discrete functions and multi-purpose problems. It offers a list of "good" solutions, not just a single solution. It always gets an answer to the problem that gets better with time. Useful when the search space is very large, and many hyperparameters are involved [25].

The disadvantages of the genetic algorithm are that the fitness value is calculated repeatedly, which can be computationally long for some problems. Since it is stochastic, there are no guarantees as to the optimality or quality of the solution. If not appropriately implemented, GA may not converge to the optimal solution [26].

2.2. Support vector machine

The support vector machine (SVM) method is used in many areas, including recognizing handwritten digits, objects, and speaker identification. It is used for classification and regression learning methods. SVM are supervised learning techniques based on statistical learning theory and the principle of structural risk minimization [27]. Vapnik and Alexei Chervonenkis developed SVMs in 1960. In contrast, however, the first paper on SVMs was published by Vladimir Vapnik and his colleagues Bernhard Boser and Isabelle Guyon in 1992 [28,29]. The mathematical function of the SVM algorithm is as follows.

$$\mathrm{f(x)} = \mathrm{sgn}\big(\mathrm{(}\ \mathrm{w}x_i) + \mathrm{b}\big) = \mathrm{sgn}(\textstyle\sum_{i=1}^{l} y_i \alpha_i (x_i x_j)), \tag{1}$$

$$Subject\ to \begin{cases} min\ \frac{\|w\|^2}{2} \\ \mathrm{y_i(wx + b)} \geq 1 \end{cases}. \tag{2}$$

Let each x_i be defined as an input whose instance has D attributes and each y_i as an output representing the class to which the instances can take two values, +1 or -1. Given an n-volume training set S consisting of (x_i,y_i) pairs, it helps to find the linear hyperplane that can best divide into different classes. The w weight vector and b constant are defined.

The model can be used for linear and nonlinear problems. SVM is used for both linearly or non-

linearly separable data. If the data are not separated linearly, kernel functions are used. The kernel function is used for nonlinear problem solving to find an optimal hyperplane to separate the reference point [30]. The SVM classification equation is followed as follows [31].

$$f(x) = \sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)K\langle x_i, x_j\rangle + b^*. \quad (3)$$

This method is called the kernel trick. The inner product is shown in the equation:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{l} y_i\alpha_i\, \varphi(x_i)\varphi(x_j) + b)\varphi(x_j). \quad (4)$$

In this study, the radial kernel function was used. The radial basis function kernel function is given below:

$$K(X_i, X_j) = \exp\left(-\frac{\|X_i, X_j\|^2}{2\sigma^2}\right). \quad (5)$$

2.3. Synthetic minority over-sampling technique (SMOTE)

Chawla et al. developed the Synthetic Minority Over Sampling Technique (SMOTE), an oversampling method, in 2002 [32]. In SMOTE, For a given minority class observation, synthetic observations are generated in a random range between the observation and its k-nearest minority class neighbors [33]. These examples sit along with the line segments of all the nearest neighbors soft the k minority class. Table 1 shows the Pseudo-code of SMOTE [34].

Table 1. SMOTE algorithm.

Algorithm: SMOTE
Input: P number of minority class sample; S% amount of synthetic to be generated; k Number of nearest neighbors Output: Ns=(S/100)*P synthetic samples 1) Create function Comput KNN (i→1 to P, Pi, Pj) {For i→1 to P Compute k nearest neighbors of each minority instance Pi and other minority instances Pj. Save the indices in the nnarray. Populate (Ns, i, nnarray) to generate new instance. En for} 2) Ns=(S/100)*P While Ns≠ 0 3) Create function GenerateS (Pi, Pj) {Choose a random number between 1 and k, call it nn For attr →1 to numattrs dif= Pi[nnarray[nn]][attr]- Pj[i][attr] gap random number between 0 and 1 Synthetic [newindex][attr]= Pj[i][attr] +gap*dif End for

Continued on next page

```
                newindex=newindex+1}
            End while
    4) Return (*End of Popurate. *)
    5) End of Pseudo-Code
```

When we examine the SMOTE algorithm given in Figure 1, firstly, the algorithm computes the distance between the feature vectors and their nearest neighbors. After that, we multiply the difference by a random number between (0, 1) and add it back into the feature. Following that, we under-sampling the majority class by randomly extracting samples from the majority class before the minority class reaches a certain percentage of the majority class [35].

2.4. SMOTE hyperparameter optimization with genetic algorithm

The selected "k", "perc.over", and "perc.under" hyperparameters in the SMOTE algorithm are shown in Table 1. Furthermore, k is the number of nearest neighbors utilized to generate new minority class samples, whereas "perc.over" (P) is the number that specifies how many more cases are to be generated from the minority class. Finally, "perc.under" (S) is the value that specifies how many more examples from the majority classes should be chosen for each case created from the minority class. Over-training and under-performance of the model occur because these hyperparameters are chosen incorrectly.

Figure 1 shows a flowchart of the improved methodology that systematically implements hyperparameter optimization.

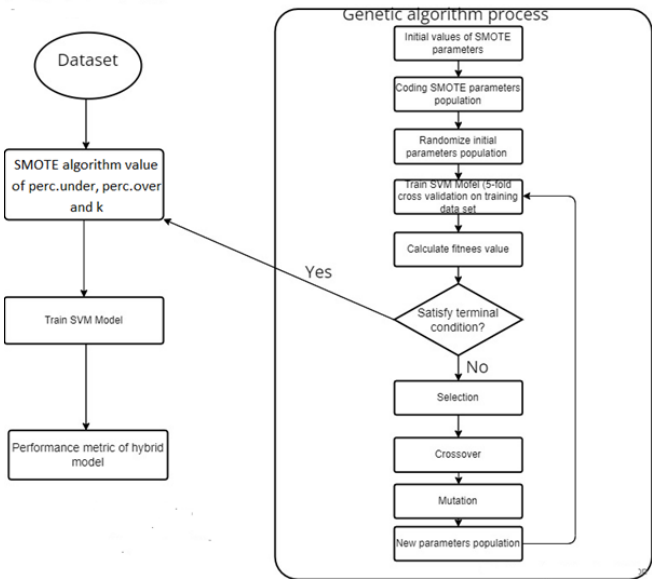

Figure 1. The flowchart of the hybrid model.

The fitness function is a function used to evaluate whether or not the population is fit. A genetic algorithm is a bridge that combines optimization with the algorithm [12]. G-mean classification results were utilized to evaluate this function. The excellence of the representative individual depends on how

high the fitness value is. The SVM classification algorithm was employed to classify the sample by computing the fitness function to determine the SMOTE hyperparameters with a genetic algorithm.

2.5. Model performance metrics for classification

The data set is divided into training and test sets to measure the model's success. The data in the training set is used to train the model. It uses data from the test set to measure the model's performance [36]. In this study, the 5-fold cross-validation method was used. In this method, the data set is divided into k equal parts. While the k-1 number of clusters is used in training, the remaining piece is used as test data. This process is repeated k times, and The accuracy value of the model is found by taking the average of the calculated accuracy values. So it creates a confusion matrix for each model. Classification accuracy alone can be misleading for imbalanced data. To better understand the classification models, performance measures were calculated from a confusion matrix (Table 2) [27].

Table 2. Confusion matrix.

		Actual	
		Yes	No
Predict	Yes	True positives (TP)	False Negatives (FN)
	No	False Positives (FP)	True Negatives (TN)

Some performance criteria are given in below to determine the classification performance.

$$\begin{cases} \text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FN+FP}}, \\ \text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}, \\ \text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}, \\ \text{Precision} = \frac{\text{TP}}{\text{TP+FP}}, \\ Gmean = \sqrt{\text{Sensitivity} * \text{Specificity}}. \end{cases} \tag{6}$$

3. Results

In this part of the study, simulation, and actual data application were made by setting up three different scenarios.

3.1. The analysis of simulation data

Logistic regression for modeling binary data as a function of other variables. Therefore, imbalanced simulation data suitable for logistic regression were produced. The general logistic regression model [37],

$$P\{Y = 1\backslash X\} = \frac{1}{1+exp(-X\beta)}. \tag{7}$$

The X represents our predictors. The β represents weights or coefficients for our predictors. The following equations show the generation process:

$$X\beta = 2X_1 + 3X_2 + X_3. \tag{8}$$

The data were derived by taking sample sizes of 1200. In order to understand the success of our proposed model, the ratio of minority class was chosen as 2.5%. Details about majority and minority classes are given in Table 3.

Table 3. An overview of the datasets.

Minority Class	Yes	2.5%	30
Majority Class	No	97.5%	1170

Three different scenarios were created for the application. The first scenario is that the support vector machine was applied without using SMOTE algorithm. The SVM was used in the second scenario after applying the SMOTE algorithm without selecting hyperparameters. The third scenario, SVM, was applied after selecting the SMOTE algorithm's optimization methods. In Table 4, optimal values were found by optimizing the hyperparameters in SMOTE algorithm with GA for scenario 3.

Table 4. Results of genetic algorithm for Simulation data.

GA settings:			
Type	=	real-valued	
Population size	=	50	
Number of generations	=	10	
Elitism	=	2	
Crossover probability	=	0.8	
Mutation probability	=	0.1	
Search domain =			
	perc_over	perc_under	k
lower	100	100	1
upper	2000	2000	10
GA results:			
Iterations	= 10		
Fitness function value	= 0.7480852		
Solution =			
	perc_over	perc_under	k
[1,]	1477.119	126.0694	5.545501

After that, "perc.over", "perc.under", and "k" values of, respectively, 343.3326, 313.0131, and 5.1290 were selected, which produced the optimal fitness function value of 0.7528. In Table 5, performance measurements of the three scenarios for the simulation data are given.

Table 5. Performance measurements of scenarios for Simulation data.

	Gmean	Accuracy	Sensitivity	Specificity
Scenario 1	0	0.9748	0	1
Scenerio 2	0.7350	0.8275	0.6607	0.8308
Scenerio 3	0.7938	0.8799	0.7131	0.8854

The accuracy rate is the percentage of correct predictions of the "yes" and "no" classes. The accuracy rate came out as 0.9748. Sensitivity gives the percentage of prediction of the "yes" class, that is, the minority class. This ratio was found to be 0%. Specificity gives the percentage of prediction of the "no" class, that is, the majority class. This ratio was found to be 1%. In scenario 1, problems are occurred in predicting the minority class. In the second scenario, these problems are eliminated by applying the SMOTE algorithm. Even though the accuracy rate dropped to 0.7350, the success rate in predicting those minority classes increased to 66%. In order to increase the prediction success of this minority class, the third scenario was established. While the accuracy rate came out as comparatively lower than the rate in the other two scenarios with 0.7938, the success rate at predicting those that did minority class was the highest with 71%. The accuracy rate does not yield the best results in the imbalanced dataset method. Hence, the G-mean value was generated by calculating the mean of sensitivity and specificity values. The third scenario produced the highest G-mean value with 0.7938. After examining the simulation results, examine the results of the proposed model with actual data.

3.2. Analysis of drug users' data

This study surveyed 1200 students currently enrolled at Samsun Ondokuz Mays University to assess drug use patterns. The questions sought to assess whether there is an essential link between students' drug use habits and their social lives and internet usage regarding the sex and age of the students [38].

Before carrying out an analysis, the Multiple Imputation with Chained Equations (MICE) methods, one of the multiple imputation methods, accounting for missing data. A statistical distribution is obtained through the dataset in this multiple imputation method. Then, this distribution uses a link to fill in the missing data. This process is repeated more than once, and each data set is stored to be used later. In addition, the error rate of each dataset is calculated. [39] The experimental results were gained by the 5-fold cross-validation way. Figure 3 shows an overview of the dataset used for three scenarios repeated 1000 times, and the results were taken as mean. The minority class rate in the original data increased from 3% to approximately 47% after the methods mentioned above were applied.

In Figure 2, the first and second column shows the number of "yes" (Actual_yes) and "no" (Actual_no) in the training data according to the different k values of the first scenario. In the second scenario, the output values of the training set after applying the SMOTE algorithm are given in the third (No_GA_Yes) and fourth (No_GA_No) columns. The last two columns ((With_GA_Yes) and (With_GA_No)) are the output values of the training set of the hybrid model in the third scenario. In the third scenario, SMOTE method was utilized to address the imbalance in the dataset, and then optimal values were found by optimizing the hyperparameters in SMOTE algorithm with GA (Table 6).

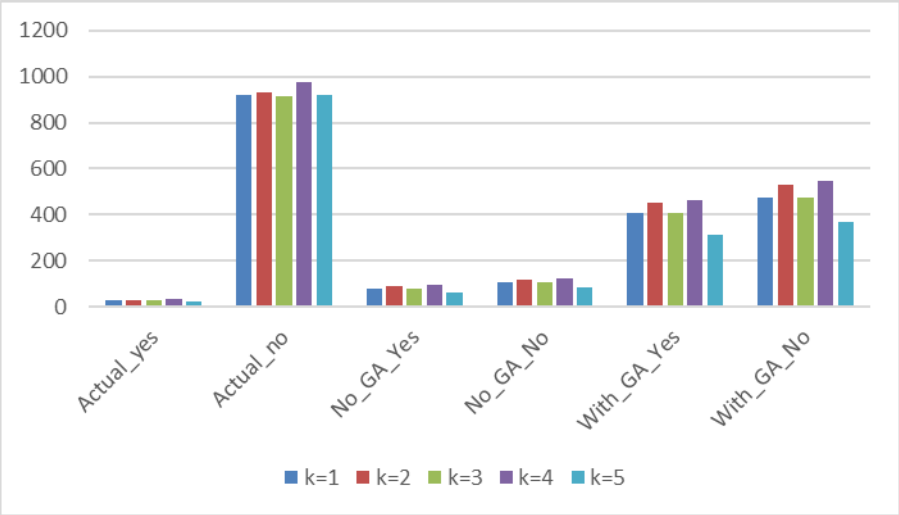

Figure 2. An overview of the datasets of the three scenarios for each 5-fold.

Table 6. Results of genetic algorithm for drug users' data.

GA settings:			
Type	=	real-valued	
Population size	=	50	
Number of generations	=	10	
Elitism	=	2	
Crossover probability	=	0.8	
Mutation probability	=	0.1	
Search domain =			
	perc_over	perc_under	k
lower	100	100	1
upper	2000	2000	10
GA results:			
Iterations	= 10		
Fitness function value	= 0.7480852		
Solution =			
	perc_over	perc_under	k
[1,]	1477.119	126.0694	5.545501

After that, "perc.over", "perc.under", and "k" values of, respectively, 1477.119, 126.0694, and 5.5455 were selected, which produced the optimal fitness function value of 0.7481. Table 7 shows performance measurements of the three scenarios for the drug users' data.

Table 7. Performance measurements of scenarios for drug users' data.

	Gmean	Accuracy	Sensitivity	Specificity
Scenario 1	0	0.9725	0	1
Scenario 2	0.6368	0.7924	0.5504	0.8013
Scenario 3	0.6978	0.7763	0.6532	0.7826

The support vector machine was applied in the first scenario without utilizing the imbalanced data method. While the accuracy rate came out as 0.9725, it is seen that only the majority class is good at estimating. When we look at the sensitivity value of the minority class in estimating, it is 0 percent. In the imbalanced data, we want to predict the minority class rather than the majority class. In this study, we want to predict the minority class of drug users by looking at their habits and social lives. So, there is a problem in estimating drug users. In scenario 2, support vector machines were used after applying the SMOTE algorithm without selecting hyperparameters. Even though the accuracy rate dropped to 0.7924, the success rate in predicting those using drugs increased up to 55%. We use a genetic algorithm to increase this rate. Lastly, the results were analyzed using the SVM algorithm after selecting the SMOTE algorithm's optimization method in the third scenario. While the accuracy rate was comparatively lower than the rate in the other two scenarios with 0.7763, the success rate at predicting those using drugs was the highest at 65%. The third scenario produced the highest G-mean value with 0.6978. As a result, when the model performances were examined for three different scenarios, the best model was obtained with scenario 3. When knowing internet use, gender, social lives, habit, and the independent variables in the model, drug users are predicted by 65% with the hybrid model.

4. Discussion

An imbalanced dataset means that one of the classes consists of more samples than the other class, which results in inequality among the classes. The most prevalent issue classifying encounters right now is the imbalance in datasets. Most data mining algorithms yield degenerate models that fail to take the minority class into account because they assume that the dataset consists of equal data clusters. SMOTE method was developed in order to address this issue. The SMOTE algorithm consists of three critical hyperparameters: “perc.over”, “perc.under”, and “k”. Failure to select the proper hyperparameters impacts the performance of the algorithm.

This study applied simulation and real data applications to the three scenarios, and their performances were compared. In the first scenario, only the SVM method was applied to the original version of the dataset. In the second scenario, SMOTE algorithm was applied. However, the hyperparameters were set to automatically, after which the support vector machines were applied to the dataset. In the third scenario, the hyperparameters of the SMOTE algorithm were determined as hyperparameter values through the use of the genetic algorithm. The result was then evaluated with the SVM algorithm.

Firstly, 1200 samples of data were simulated using logistic regression. In this simulation dataset, the ratio of the minority class was chosen as 2.5%. The SMOTE algorithm’s hyperparameters were determined as “perc.over”, “perc.under” and “k” values, respectively 343.3326, 313.0131, 5.129047, and the optimal fitness function value was found to be 0.7528. The result was then evaluated with the SVM algorithm. When all three scenarios are considered, the accuracy rate decreased from 0.9748

to 0.8799; however, the higher accuracy rate was achieved without taking the minority class into account in the imbalanced dataset. The success rate in predicting the minority class increased from 0% to 66% with the method implemented in the second scenario, whereas in the third scenario, the success rate increased to 71%. The G-mean value was generated by calculating the mean of sensitivity and specificity values to compare the performance of imbalanced dataset models. The third scenario yielded the highest G-mean value of 0.7938.

Secondly, it used a survey dataset to determine student drug user patterns. In the third scenario, the hyperparameters of the SMOTE algorithm were determined as "perc.over", "perc.under" and "k" values, respectively 1477.119, 126.0694, 5.545501, and the optimal fitness function value was found to be 0.7481. The result was then evaluated with the SVM algorithm. When all three scenarios are considered, the accuracy rate decreased from 0.9725 to 0.7763; however, the higher accuracy rate was achieved without taking the minority class into account in the imbalanced dataset. The success rate in predicting the minority class, those that used drugs, increased from 0% to 55% with the method implemented in the second scenario, whereas in the third scenario, the success rate increased up to 65%. The third scenario yielded the highest G-mean value of 0.6978.

5. Conclusions

Machine learning algorithms are divided into classification and regression. When applying classification algorithms, the biggest problem is that the classes are not evenly distributed. In such cases, it becomes impossible to predict the minority class. When predicting the class of the data, it assigns it to the majority class. However, in some studies, it is more critical to predict minority candidates. For example, in studies of drug users or not. SMOTE algorithm is used to eliminate this problem. There are three critical hyperparameters in the SMOTE algorithm. These hyperparameters add synthetic data to the minority class, and data from the majority class are also selected. A genetic algorithm was used to optimize these hyperparameters. Firstly, simulation data with a minority ratio of 2.5 percent and a majority ratio of 97.5 percent are generated. Actual data on drug use were used as the second data set. Three different scenarios were set up for each dataset. Scenario 1 shows the results without using the SMOTE algorithm. Scenario 2 shows that the SVM was used after applying SMOTE algorithm without the GA algorithm. In the third scenario, the results were analyzed using the SVM algorithm after selecting the SMOTE algorithms with the GA algorithm. Scenario 1 and scenario 2 are compared, and there is a problem estimating the minority in the first scenario. The problem was resolved in the second scenario. To further increase the success of this method, scenario three was installed. SMOTE hyperparameters were selected using a genetic algorithm in the scenario. The third scenario has the highest performance of the two datasets.

In summary, this study indicates that the utilization of a genetic algorithm in order to determine the optimum hyperparameter values for the SMOTE algorithm saves time while simultaneously improving its performance [15]. The critical part that distinguishes this study from other studies is that it does not focus only on the minority class when applying the SMOTE algorithm.

In this study, both the majority class sample was selected, and the minority synthetic data were produced. In other words, when set by looking at the majority class, we may encounter an overfitting problem. The overfitting problem is a situation that occurs when it gives predictions for training sets but incorrect predictions for a new dataset. With the hybrid model, a joint optimal sample of the majority and minority classes was obtained. In this way, the overfitting problem is eliminated. Future

research can test the model's generalizability by conducting experiments on more clinical datasets. In future studies, different optimization techniques can be examined to estimate the hyperparameters of the SMOTE algorithm.

Acknowledge

We would like to thank the Reviewers for taking the time and effort necessary to review the manuscript. We sincerely appreciate all valuable comments and suggestions, which helped us improve the manuscript's quality.

Conflict of interest

The authors declare no conflict of interest.

References

1. A. Fernández, S. García, F. Herrera, Addressing the classification with imbalanced data: open problems and new challenges on class distribution, In: *Lecture Notes in Computer Science*, Heidelberg: Springer, **6678** (2011). https://doi.org/10.1007/978-3-642-21219-2_1
2. M. Liuzzi, P. A. Pelizari, C. Geiß, A. Masi, V. Tramutoli, H. Taubenböck, A transferable remote sensing approach to classify building structural types for seismic risk analyses: the case of Val d'Agri area (Italy), *Bull. Earthq. Eng.*, **17** (2019), 4825–4853.
3. D. Devarriya, C. Gulati, V. Mansharamani, A. Sakalle, A. Bhardwaj, Unbalanced breast cancer data classification using novel fitness functions in genetic programming, *Expert Syst. Appl.*, **140** (2020), 112866. https://doi.org/10.1016/j.eswa.2019.112866
4. S. Katoch, S. S. Chauhan, V. Kumar, A review on genetic algorithm: past, present, and future, *Multimed. Tools Appl.*, **80** (2021), 8091–8126. https://doi.org/10.1007/s11042-020-10139-6
5. Y. L. Yuan, J. J. Ren, S. Wang, Z. X. Wang, X. K. Mu, W. Zhao, Alpine skiing optimization: A new bio-inspired optimization algorithm, *Adv. Eng. Softw.*, **170** (2022), 103158 https://doi.org/10.1016/j.advengsoft.2022.103158
6. J. F. Goycoolea, M. Inostroza-Ponta, M. Villalobos-Cid, M. Marín, Single-solution based metaheuristic approach to a novel restricted clustering problem, 2021. https://doi.org/10.1109/SCCC54552.2021.9650429
7. J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, Bradford Books, 1992.
8. S. N. Sivanandam, S. N. Deepa, *Introduction to Genetic Algorithms*, Heidelberg: Springer Berlin, 2010.
9. F. Ortiz, J. R. Simpson, J. Pignatiello, A. Heredia-Langner, A genetic algorithm approach to multiple-response optimization, *J. Qual. Technol.*, **36** (2004), 432–450. https://doi.org/10.1080/00224065.2004.11980289
10. H. I. Calvete, C. Gale, P. M. Mateo, A new approach for solving linear bilevel problems using genetic algorithms, *European J. Oper. Res.*, **188** (2008), 14–28 https://doi.org/10.1016/j.ejor.2007.03.034

11. S. S. Nimankar, D. Vora, Designing a model to handle imbalance data classification using SMOTE and optimized classifier, In: *Data Management, Analytics and Innovation*, Singapore: Springer, 2020, 323–334.
12. K. Jiang, J. Lu, K. L. Xia, A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE, *Arab. J. Sci. Eng.*, **41** (2016), 3255–3266. http://doi.org/10.1007/s13369-016-2179-2
13. R. Obiedat, R. Qaddoura, A. M. Al-Zoubi, L. Al-Qaisi, O. Harfoushi, M. Alrefai, et al., Sentiment analysis of customers' reviews using a hybrid evolutionary SVM based approach in an imbalanced data distribution, *IEEE Access*, **10** (2022), 22260–22273. https://doi.org/10.1109/ACCESS.2022.3149482
14. L. Wang, Imbalanced credit risk prediction based on SMOTE and multi-kernel FCM improved by particle swarm optimization, *Appl. Soft Comput.*, **114** (2022), 108153. https://doi.org/10.1016/j.asoc.2021.108153
15. L. Demidova, I. Klyueva, SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem, 2017. https://doi.org/10.1109/MECO.2017.7977136
16. S. Sreejith, H. K. Nehemiah, A. Kannan, Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection, *Comput. Biol. Med.*, **126** (2020), 103991. https://doi.org/10.1016/j.compbiomed.2020.103991
17. Y.-C. Wang, C.-H. Cheng, A multiple combined method for rebalancing medical data with class imbalances, *Comput. Biol. Med.*, **134** (2021), 104527. https://doi.org/10.1016/j.compbiomed.2021.104527
18. B. Zorić, D. Bajer, G. Martinović, Employing different optimisation approaches for SMOTE parameter tuning, 2016. https://doi.org/10.1109/SST.2016.7765657
19. E. Sara, C. Laila, I. Ali, The impact of SMOTE and grid search on maintainability prediction models, 2019. https://doi.org/10.1109/AICCSA47632.2019.9035342
20. J. J. Ren, Z. X. Wang, Y. Pang, Y. L. Yuan, Genetic algorithm-assisted an improved AdaBoost double-layer for oil temperature prediction of TBM, *Adv. Eng. Inform.*, **52** (2022), 101563. https://doi.org/10.1016/j.aei.2022.101563
21. Y. L. Yuan, X. K. Mu, X. Y. Shao, J. J. Ren, Y. Zhao, Z. X. Zhao, Optimization of an auto drum fashioned brake using the elite opposition-based learning and chaotic k-best gravitational search strategy based grey wolf optimizer algorithm, *Appl. Soft Comput.*, **123** (2022), 108947. https://doi.org/10.1016/j.asoc.2022.108947
22. M. L. Shi, S. Wang, W. Sun, L. Y. Lv, X. G. Song, A support vector regression-based multi-fidelity surrogate model, 2019.
23. D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Professional, 1989.
24. S. Panda, N. P. Padhy, Comparison of particle swarm optimization and genetic algorithm for FACTS-based controller design, *Appl. Soft Comput.*, **8** (2008), 1418–1427. https://doi.org/10.1016/j.asoc.2007.10.009
25. D. Orvosh, L. Davis, Using a genetic algorithm to optimize problems with feasibility constraints, *IEEE World Congress on Computational Intelligence*, 1994. https://doi.org/10.1109/ICEC.1994.350001

26. E. C. Gonçalves, A. Plastino, A. A. Freitas, A genetic algorithm for optimizing the label ordering in multi-label classifier chains, In: *2013 IEEE 25th International Conference on Tools with Artificial Intelligence*, 2013. https://doi.org/10.1109/ICTAI.2013.76
27. J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, 2011.
28. V. Vapnik, Principles of risk minimization for learning theory, In: *Proceedings of the 4th International Conference on Neural Information Processing Systems*, 1991, 831–838.
29. T. Koc, P. Akın, Estimation of high school entrance examination success rates using machine learning and beta regression models, *J. Intell. Syst. Theory Appl.*, **5** (2022), 9–15. http://doi.org/10.38016/jista.922663
30. D. Guleryuz, Estimation of soil temperatures with machine learning algorithms-Giresun and Bayburt stations in Turkey, *Theor. Appl. Climatol.*, **147** (2022), 109–125.
31. Q. Quan, Z. Hao, X. F. Huang, J. C. Lei, Research on water temperature prediction based on improved support vector regression, *Neural Comput. Appl.*, 2020, 1–10. https://doi.org/10.1007/S00521-020-04836-4
32. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, **16** (2002), 321–357. https://doi.org/10.5555/1622407.1622416
33. J. Brandt, E. Lanzén, A comparative review of SMOTE and ADASYN in imbalanced data classification, In: *Bachelor's Thesis*, Uppsala: Uppsala University, 2021.
34. H. Al Majzoub, I. Elgedawy, Ö. Akaydın, M. K. Ulukök, HCAB-SMOTE: A hybrid clustered affinitive borderline SMOTE approach for imbalanced data binary classification, *Arab. J. Sci. Eng.*, **45** (2020), 3205–3222.
35. P. Akin, Y. Terzi, Comparison of unbalanced data methods for support vector machines, *Turkiye Klinikleri J. Biostat.*, **13** (2021), 138–146. http://doi.org/10.5336/biostatic.2020-80268
36. S. Uğuz, Makine öğrenmesi teorik yönleri ve Python uygulamaları ile bir yapay zeka ekolü, *Nobel Yayıncılık Ankara*, 2019.
37. R. E. Wright, Logistic regression, In: *Reading and Understanding Multivariate Statistics*, 1995, 217–244.
38. T. Koc, H. Koc, E. Ulas, Üniversite öğrencilerinin kötü alışkanlıklarının bayesci ağ yöntemi ile belirlenmesi, *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, **26** (2017), 230–240.
39. S. V. Buuren, K. Groothuis-Oudshoorn, Mice: Multivariate imputation by chained equations in R, *J. Statist. Softw.*, **45** (2011), 1–68.