Turkish Journal of Electrical Engineering & Computer Sciences

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

Research Article

Turk J Elec Eng & Comp Sci
(2023) 31: 180 – 192

doi:10.55730/1300-0632.3977

Binary text classification using genetic programming with crossover-based oversampling for imbalanced datasets

Mona Khalifa A. ALJERO[1], **Nazife DİMİLİLER**[2,*]

[1]Department of Computer Science, Faculty of Education, Misurata University, Misurata, Libya
[2]Department of Information Technology, School of Computing and Technology,
Eastern Mediterranean University, Famagusta, North Cyprus, via Mersin 10, Turkey

Received: 22.05.2022 • **Accepted/Published Online:** 01.12.2022 • **Final Version:** 19.01.2023

Abstract: It is well known that classifiers trained using imbalanced datasets usually have a bias toward the majority class. In this context, classification models can present a high classification performance overall and for the majority class, even when the performance for the minority class is significantly lower. This paper presents a genetic programming (GP) model with a crossover-based oversampling technique for oversampling the imbalanced dataset for binary text classification. The aim of this study is to apply an oversampling technique to solve the imbalanced issue and improve the performance of the GP model that employed the proposed technique. The proposed technique employs a crossover operator for generating new samples for the minority class in an imbalanced text dataset. By using a combination of this crossover-based oversampling technique with GP, the performance was improved. It is shown that the proposed combination outperforms all GP applications that use the original dataset without resampling. Moreover, the performance of the proposed system surpassed GP approaches using the synthetic minority oversampling technique (SMOTE) and random oversampling. Further comparison with the state-of-the-art on five imbalanced text datasets in terms of F1-score shows the superior performance of the proposed approach.

Key words: Imbalanced dataset, text classification, genetic programming, oversampling techniques, resampling

1. Introduction

Class imbalance is a frequently occurring problem in real life, and examples include spam emails, disease identification, credit card fraud, cyber-attack, prediction of manufacturing equipment and information retrieval [1]. Classifiers trained on imbalanced datasets tend to develop a bias toward the majority class that occurs in abundance compared to the minority classes [2]. However, in many cases, e.g., disease identification, the target or significant class is the minority class. The imbalanced datasets make it difficult to use machine learning in the current real-world applications. Increasingly more attention has been brought to this problem; it has been named among the ten difficult problems in data mining [3]. A variety of approaches have been proposed to address this problem. These approaches can be divided into two categories: (i) modify the dataset to balance the classes and (ii) propose a new classifier to deal with the imbalanced dataset. In this work, the first category is considered.

Resampling is one of the approaches used to deal with the imbalanced dataset by modifying the dataset itself. There are two types of resampling approaches, namely, undersampling and oversampling. Undersampling

*Correspondence: nazife.dimililer@emu.edu.tr

is defined as reducing the size of the majority class to that of the minority class by eliminating data in the majority class. On the other hand, oversampling is defined as increasing the size of the minority class to that of the majority class by either duplicating the data in the minority class or generating new data items using the existing ones in the minority class. Each of these resampling techniques has its own benefits and drawbacks. In the case of undersampling, valuable data could be eliminated, resulting in a loss of information [4–6]. Therefore, this study proposes an oversampling technique to balance a dataset. Even though many researchers have proposed various oversampling approaches, the most prevalent ones are random oversampling and SMOTE. The random oversampling approach adds a random selection from the minority class to the train set. SMOTE, first presented in 2002 by Chawla et al. [7], generates artificial samples for the minority class. SMOTE creates virtual training samples for the minority class using linear interpolation. For each sample in the minority class, some of the k-nearest neighbors are randomly chosen to generate these synthetic training samples. A large number of studies on extending and augmenting the SMOTE algorithms have been published in [8]–[15].

Algorithm 1 CROSS-OVER SAMPLE

Input T: Imbalanced Training set
  $T \leftarrow Set_{Minority}$: Set of Minority Samples
  $U \leftarrow Set_{Majority}$: Set of Majority Samples
Output T': Balanced Training Set
  $T' \leftarrow Set'_{Minority}$: Set of Minority Samples
  $U' \leftarrow Set'_{Majority}$: Set of Majority Samples
Initialize
  $Set'_{Minority} \leftarrow Set_{Minority}$
  $Set'_{Majority} \leftarrow Set_{Majority}$
while $Size(Set'_{Majority}) > Size(Set'_{Minority})$ **do**
  Randomly select $Text_1$ from $Set'_{Minority}$
  Randomly select $Text_2$ from $Set'_{Majority}$
  Generate random number t_1 between 0 and $length(Text_1)$
  Generate random number t_2 between 0 and $length(Text_2)$
  $Text_3 \leftarrow concatenate(Text_1[0..t1-1], Text_2[t2..length(Text_2)]$
  $Text_4 \leftarrow concatenate(Text_2[0..t2-1], Text_1[t1..length(Text_1)]$
  $Set'_{Minority} \leftarrow Set'_{Minority} U\ Text_3\ U\ Text_4$
End while

Tripathi et al. [16] proposed a novel technique to solve the imbalance issue of binary datasets. The authors firstly used SMOTE, then divided the obtained data by applying a Gaussian-Mixture framework based on the clustering technique. Finally, they selected artificial samples based on the weight assigned to the cluster. Their experiments used a support vector machine (SVM) as a classifier. The results from this technique were compared with the original SMOTE and the state-of-the-art. The proposed technique outperformed most of the other techniques. The authors indicated that this technique, which was applied for binary classification, could also be used for multiclassification by adopting some modifications. Azad et al. [17] proposed a novel method using SMOTE, a decision tree, and a genetic algorithm for classification. Their method contained four layers: preprocessing, feature selection by genetic algorithm, training the model, and evaluation.

The GP paradigm was proposed by Koza as a machine learning approach for automatically developing software programs in 1992 [14]. Since then, GP has been applied to various problems and has also been

employed successfully in creating reliable and efficient classifiers to address a variety of classification issues [14, 15]. However, if the data set is not balanced, similar to many other machine learning methods, the GP can evolve models biased toward the majority class [18].

The researchers in [19–22] addressed the imbalanced data issue by proposing alternative fitness functions in the GP model. The authors in [19] proposed a new fitness function based on the correlation ratio and the number of samples from the minority class. The authors evaluated their approach on five binary datasets. The authors in [20] proposed three fitness functions to improve the classification performance on binary imbalanced datasets. The first one extends the accuracy, weighted average accuracy, and geometric mean metrics, the second one extends the average mean squared error metric and the third one approximates the accuracy metric. Another fitness function namely tangent distance and weight-based (TDWB) fitness function is proposed in [21] to improve the performance. The proposed technique determines the difference between predicted and expected values to create an efficient classifier. Moreover, the TDWB fitness function solves the underfitting problem. Since the proposed fitness function treats both classes equally, the dominance of the majority class over the minority class during training is eliminated. Classifiers predicting values closer to the predicted values will be rewarded more based on a distance measure. Underfitting is also addressed by using this distance-considered approach. The authors show that their fitness function produces a well-fitted classifier due to the use of this distance measure. The authors in [22] introduced a new fitness function with the GP model to deal with the imbalanced dataset. They evaluated their approach on four imbalanced datasets. The proposed approach is compared to the SVM and the state-of-the-art. The experimental results demonstrated that the proposed approach outperformed or was at least as effective as the SVM and other state-of-the-art methods.

As another approach to dealing with imbalanced datasets, Mostafaei et al. [23] proposed a combination of undersampling and oversampling strategies to increase the accuracy of the classification. Mostafaei et al. show that the combination strategy could yield balanced datasets, which increase the accuracy. The authors in [23] presented a novel undersampling technique that chooses samples of the majority class with the highest distance and density. The authors combined this undersampling technique with SMOTE to increase the performance. Mostafaei et al. show that the combination strategy could yield balanced datasets, increasing accuracy.

Jiang et al. [24] proposed a new oversampling technique based on the contribution degree of the classification. The authors computed the ratio between the total number of positive samples, the minority set, and the number of samples in clusters formed using the k-means algorithm. For each potential sample, the classification contribution degree, based on safe neighborhoods, is used to calculate the number of synthetic samples generated by SMOTE. In another study, Pereira et al. [25] proposed a resampling technique to improve the performance in binary classification tasks. The authors proposed two oversampling and two undersampling techniques to address the imbalanced data issue. The evaluation results show a significant performance improvement.

In [26], a novel oversampling method was presented to deal with the imbalanced dataset. The presented method was based on clustering for imbalanced datasets. The experiments on five datasets and the comparison to other resampling techniques indicated the effectiveness of their approach.

The main contributions of this work can be listed as follows:

- We employ a crossover-based oversampling technique to deal with the imbalance issue, particularly in textual datasets.
- We employ the proposed crossover-based oversampling technique to improve the classification performance

of a novel GP-based system with hybrid mutation (GP-H) [27].

- The performance of the GP-H with crossover-based oversampling is compared with the prevalent oversampling techniques: random oversampling and SMOTE.
- The proposed system GP-H with crossover-based oversampling is compared with the standard GP with crossover-based oversampling to show the performance improvement of GP-H.
- For each dataset employed, we present the state-of-the-art result to show the performance of the proposed system.

This study focuses on binary classification since applications associated with imbalanced data frequently require binary classification. It should be noted that it is possible to solve a multiclass classification problem by breaking it down into many binary classification problems. Finally, it is also worth noting that binary classification is extremely challenging when the dataset is imbalanced. Therefore, the application domain of this study is of utmost importance.

The rest of this paper is organized as follows. Section 2 illustrates the methodology used in this study. Section 3 introduces the experimental setup. Section 4 presents the results on imbalanced datasets and discusses the findings. Finally, the conclusion of this work and a brief discussion of future work are presented in Section 5.

2. Methods

The proposed approach for dealing with the imbalanced dataset is described in this section. Unequal distribution of data, or imbalance, causes a bias during machine learning toward the majority class resulting in the misclassification of minority class samples. This study proposes an oversampling technique to solve the imbalanced data problem specifically for textual datasets. The experimental results show significant performance improvement for the GP-based classifiers.

Our proposed system employs the GP model with a hybrid mutation operator (GP-H) for binary classification, as described in [27]. A novel crossover-based oversampling technique for generating new data samples is employed to improve the performance.

2.1. Crossover-based oversampling of the imbalanced datasets

Figure 1 shows the application of the crossover-based oversampling on two randomly selected parents. By recombining Parent 1 and Parent 2, at random points, the proposed oversampling technique generates two new tweets, Child 1 and Child 2.

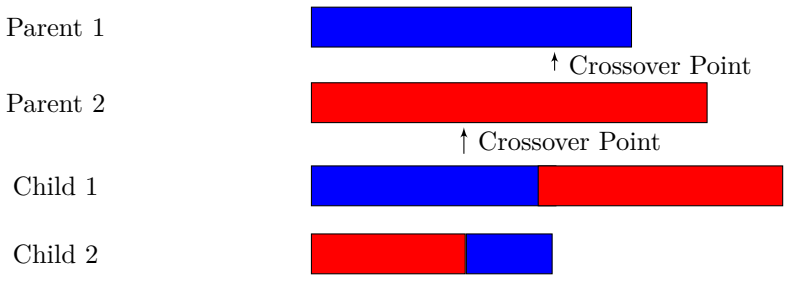

Figure 1. Example of the proposed oversampling technique.

2.2. Genetic programming

The GP framework employed in the current study was proposed in [27] as a modification of the GP technique employing a hybrid mutation operator, namely GP-H. Each individual in the GP-H framework is a tree that classifies texts. GP-H uses two genetic operators, crossover and mutation, with 80% and 20% probability, respectively. The standard one-point crossover is used as the crossover operator. The novel hybrid mutation method randomly selects one of two mutations: a unique feature mutation and a standard one-point mutation. Both mutation operators affect only feature values. The universal sentence encoder is used in the feature extraction stage to encode each text into a vector. Thus, the mutation operators are used to increase variation in terms of the offspring's features and in effect variation in the pool of possible solutions. The experimental findings in [27] demonstrate that GP-H has superior performance on numerous text datasets over the generic GP model.

2.3. Evaluation scheme

The most commonly used evaluation metrics, Recall, Precision and F1-score, are employed to present experimental results and allow comparison with the state-of-the-art. Recall, Precision, and F1-score are computed as follows:

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$F1 - score = \frac{2(Precision * Recall)}{Precision + Recall} \tag{4}$$

where TP corresponds to the true positive, TN corresponds to the true negative, FP is the false positive, and FN is the false negative.

2.4. Datasets

In order to verify the proposed technique's efficacy, we used five imbalanced textual datasets. The DATD-English dataset was available as train and test sets and is used in this form to allow for comparison with other work. The remaining datasets are split into train and test sets using the ratio of 80:20. Table 1 shows the five imbalanced text datasets employed in this study. The ling-Spam, DATD and SMS Spam datasets were retrieved from Kaggle[1] , DATD-English dataset was retrieved from GitHub[2] , and Re-dataset-two-labels dataset was retrieved from Bitbucket[3] repositories.

Ling-Spam: This dataset is a publicly available dataset which contains 2893 messages in the English language received via the Linguist List. The Linguist List is a regulated mailing collection about the study of

[1]https://www.kaggle.com/
[2]https://github.com/
[3]https://bitbucket.org/product/

linguistics. As shown in Table 2, 2412 messages, 83.3% of all messages, are labeled as Non-Spam. The remaining 481 messages, 16.6% of all messages, are labeled as Spam.

DATD: This is an abbreviation of Depression and Anxiety in the Twitter Dataset. This dataset was obtained from Indonesian Twitter and tags whether each tweet contains a sign of anxiety or depression. It consists of 2201 tweets, where 1468 tweets are annotated as containing anxiety or depression, and 733 tweets annotated as otherwise.

DATD-English: This dataset is a depression and anxiety dataset from English Twitter. Owen et al. retrieved tweets from May 2018 until August 2019 [28]. It comprises two training sets and one test set. DATD_training contains 900 tweets divided into 473 and 427 for Mental-Health and Other respectively. DATD+Rand_training contains 4500 tweets divided into 473 and 4027 for Mental-Health and Other, respectively. The DATD_and_DATD+Rand_test set contains 75 tweets labeled as Mental-Health and 75 labeled as other. The DATD_training set is balanced, while the DATD+Rand_training is imbalanced. Since this study focuses on imbalanced data, we will use the DATD+Rand_training set as the training set.

Table 1. The imbalanced datasets used in this study.

Dataset name	Source	Link
Ling-Spam	Kaggle	https://www.kaggle.com/mandygu/lingspam-dataset
DATD	Kaggle	https://www.kaggle.com/stevenhans/depression-and-anxiety-in-twitter-id
DATD-English	Bitbucket	https://bitbucket.org/nlpcardiff/preemptive-depression-anxiety-twitter
SMS Spam	Kaggle	https://www.kaggle.com/team-ai/spam-text-message-classification
Re-dataset-two-labels	GitHub	https://github.com/okkyibrohim/id-abusive-language-detection

Table 2. The statistics of the used imbalanced datasets.

Name	Majority class	Minority class	Total
Ling-Spam	Non-Spam (2412) 83.3%	Spam (481) 16.6%	2893
DATD	Otherwise (1468) 66.69%	Depression or anxiety sign (733) 33.3%	2201
DATD-English	Others (4027) 89.4%	Mental-Health (473) 10.5%	4500
SMS Spam	Ham (4827)86.5%	Spam (747) 13.4%	5574
Re_dataset_two_labels	Abusive (1685) 83.58%	Non-Abusive (331) 16.41%	2016

The **SMS Spam** dataset was created in 2011 by Almeida et al. [29]. This dataset is composed of 5574 SMS messages in the English language. This binary classified dataset classified messages as Ham or Spam. It is divided into 4827 SMS messages labeled as Ham messages and 747 labeled as Spam messages. This collection comprises SMS messages from different sources: 425 spam messages from Grumbletext[4] Website, 3375 Ham messages from the NUS[5] corpus, and 1324 messages (1002 Ham messages and 322 Spam messages) from the SMS Spam[6] corpus, and 450 Ham messages from a PhD thesis[7].

Re_dataset_two_labels This text dataset of Twitter consists of 2016 tweets from Indonesian Twitter. It is extracted from GitHub. The Re_dataset_two_labels dataset was created in 2018 by Ibrohim and Budi

[4]http://www.grumbletext.co.uk/
[5]http://www.comp.nus.edu.sg/ rpnlpir/downloads/corpora/smsCorpus/
[6]http://www.esp.uem.es/jmgomez/smsspamcorpus/
[7]http://etheses.bham.ac.uk/253/1/Tagg09PhD.pdf

[30] for abusive language detection. This dataset comprised 331 tweets classified as Non-Abusive, and 1685 tweets as Abusive.

Figure 2 shows the flow diagram of the proposed approach, starting with the splitting of the dataset into training and testing sets. The crossover-based technique is applied on the training set until the sizes of the majority set and minority set are equal. The training set thus balanced may be used for training classiers. In this work, we used both standard GP and the GP approach described by [29] to evolve optimal classifiers.

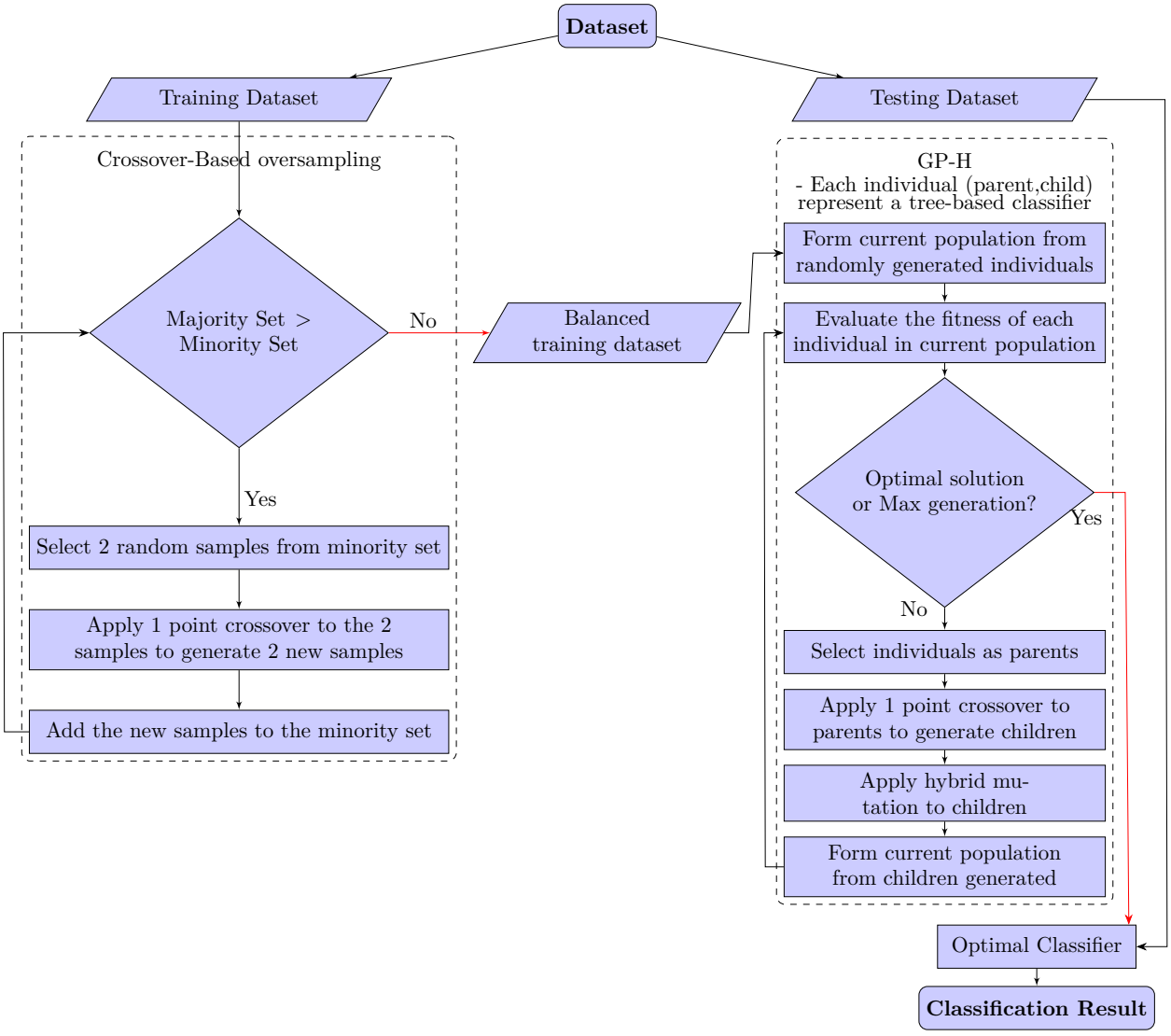

Figure 2. Flow diagram of the proposed approach.

2.5. Results and discussion

All the experiments in this study were implemented in Python 3.7. Table 3 shows the results on the five imbalanced datasets.

We conducted experiments on each dataset using eight different settings. The experiments were repeated ten times for each setting and dataset. The results presented here are the averages of 10 experiments. The

Table 3. The performance of the proposed approach versus other approaches on test sets. For each dataset, the best result is highlighted in bold.

Dataset name	Approach	F1-score	Recall	Precision	Accuracy
Ling-Spam	GP-H + crossover-based oversampling	**98.70**	**98.70**	**98.70**	**98.70**
	GP-H + SMOTE	95.71	95.71	95.73	95.71
	GP-H + Random oversampling	95.48	95.48	95.49	95.48
	GP-H + Without resampling	89.01	89.00	89.04	89.00
	Standard GP + crossover-based oversampling	94.86	94.86	94.88	94.86
	Standard GP + SMOTE	95.56	95.56	95.58	95.56
	Standard GP + Random oversampling	94.08	94.08	94.10	94.08
	Standard GP + Without resampling	94.93	94.99	94.88	94.93
DATD	GP-H + crossover-based oversampling	**79.65**	**80.00**	**79.31**	**79.54**
	GP-H + SMOTE	69.95	69.95	69.97	69.95
	GP-H + Random oversampling	69.15	69.15	69.14	69.15
	GP-H + Without resampling	64.13	64.13	64.14	64.13
	Standard GP + crossover-based oversampling	68.58	68.60	68.57	68.55
	Standard GP + SMOTE	68.60	68.80	68.41	68.60
	Standard GP + Random oversampling	68.80	68.82	68.79	68.80
	Standard GP + Without resampling	70.98	70.99	70.98	70.98
DATD-English	GP-H + crossover-based oversampling	**80.44**	**80.50**	**80.39**	**80.44**
	GP-H + SMOTE	76.89	76.89	76.89	76.89
	GP-H + Random oversampling	75.70	75.93	75.49	75.80
	GP-H + Without resampling	70.06	70.08	70.06	70.06
	Standard GP + crossover-based oversampling	78.25	78.31	78.21	78.25
	Standard GP + SMOTE	71.60	71.71	71.50	71.60
	Standard GP + Random oversampling	72.47	72.95	72.00	73.57
	Standard GP + Without resampling	75.33	75.49	75.19	75.33
SMS Spam	GP-H + crossover-based oversampling	**95.83**	**95.83**	**95.83**	**95.83**
	GP-H + SMOTE	86.14	86.20	86.10	85.26
	GP-H + Random oversampling	85.60	86.80	84.44	85.23
	GP-H + Without resampling	77.54	78.00	77.10	77.53
	Standard GP + crossover-based oversampling	87.80	88.39	87.23	87.72
	Standard GP + SMOTE	91.84	92.01	91.69	91.86
	Standard GP + Random oversampling	91.26	91.45	91.09	91.26
	Standard GP + Without resampling	86.76	86.90	86.63	86.35
Re_dataset_two_labels	GP-H + crossover-based oversampling	**88.54**	**88.52**	**88.58**	**88.52**
	GP-H + SMOTE	72.42	72.44	72.42	72.42
	GP-H + Random oversampling	72.60	73.00	72.21	72.58
	GP-H + Without resampling	70.97	71.90	70.08	71.03
	Standard GP + crossover-based oversampling	66.41	66.82	66.01	66.41
	Standard GP + SMOTE	70.34	70.60	70.10	70.34
	Standard GP + Random oversampling	69.48	69.48	70.49	69.48
	Standard GP + Without resampling	66.37	66.47	66.29	66.37

performance of two frequently used oversampling techniques, namely SMOTE and random oversampling, are presented to demonstrate the superiority of the proposed technique. The results of using the imbalanced data without any oversampling are also presented to provide a baseline.

The proposed approach achieved a significant performance improvement on all datasets compared with GP using SMOTE oversampling, GP using random oversampling, and GP without any oversampling techniques.

In general, we observed that the worst results are associated with the GP-H when the original dataset is used (GP-H + Without resampling in four out of five datasets, and the other dataset was with Standard GP + Random oversampling). The imbalanced dataset causes the model to misclassify the text as it biases toward the majority class. Moreover, when a classifier is trained on an imbalanced dataset, it frequently creates models that favor the majority class and perform poorly on the minority class [31].

The proposed oversampling technique worked better with the hybrid mutation technique, as seen in Table 3. Moreover, the proposed approach (crossover-based oversampling and GP-H) always achieved the best results. When employed with the standard GP framework, the crossover-based oversampling technique failed to improve the performance; however, it consistently scored the highest F1-score when combined with GP-H.

Table 4 presents the state-of-the-art F1-score and classification model for each dataset to the best of our knowledge. Only one of the state-of-the-art systems on the datasets mentioned above tackled the imbalance problem. The best-published result for DATD-English and SMS Spam are produced by SVM models in [28, 34], respectively. Even though the SVM model has been used successfully with any new test set that may come from various distributions, the SVM's goal is to reduce the training error rate, which results in a model biased to the majority class [35]. Naïve Bayes (NB) model was used in [30], whereas for DATD [33], an LSTM model was implemented. In [32], a GP approach was used with SVM as a classifier.

As shown in Table 4, the proposed approach achieved higher results than the state-of-the-art on all datasets proving the proposed approach's effectiveness in the context of an imbalanced dataset. The maximum improvement was 56.28% on the Ling-spam dataset, while the minimum improvement was 1.83% on SMS Spam.

As Table 5 shows the average running time of GP-H + crossover-based oversampling is faster than other methods on all used datasets. The proposed method achieves the best results on all datasets in terms of F1-score, Recall, Precision, and Accuracy; moreover, experimental results show that it is more efficient in terms of execution time.

Table 4. The state-of-the-art of each dataset.

Dataset name	Model	F1-score of state-of-the-art	F1-score of proposed approach
Ling-Spam	GA-SVM	33.42 [32]	**98.70**
DATD	LSTM	76.60 [33]	**79.65**
DATD-English	SVM	75.00 [28]	**80.44**
SMS Spam	SVM	94.00 [34]	**95.83**
Re_dataset_two_labels	LSTM	83.68 [30]	**88.54**

Table 5. Average running time of all methods. For each dataset, the shortest time is highlighted in bold.

Dataset name	Approach	Running Time (Seconds)
Ling-Spam	GP-H + crossover-based oversampling	**211.45**
	GP-H + SMOTE	212.23
	GP-H + Random oversampling	221.00
	GP-H + Without resampling	253.64
	Standard GP + crossover-based oversampling	267.38
	Standard GP + SMOTE	262.65
	Standard GP + Random oversampling	291.10
	Standard GP + Without resampling	256.52
DATD	GP-H + crossover-based oversampling	**164.37**
	GP-H + SMOTE	172.40
	GP-H + Random oversampling	251.70
	GP-H + Without resampling	243.38
	Standard GP + crossover-based oversampling	194.12
	Standard GP + SMOTE	201.00
	Standard GP + Random oversampling	217.96
	Standard GP + Without resampling	172.15
DATD-English	GP-H + crossover-based oversampling	**453.38**
	GP-H + SMOTE	471.51
	GP-H + Random oversampling	496.81
	GP-H + Without resampling	497.59
	Standard GP + crossover-based oversampling	490.73
	Standard GP + SMOTE	480.27
	Standard GP + Random oversampling	502.19
	Standard GP + Without resampling	510.91
SMS Spam	GP-H + crossover-based oversampling	**527.34**
	GP-H + SMOTE	539.45
	GP-H + Random oversampling	564.84
	GP-H + Without resampling	601.88
	Standard GP + crossover-based oversampling	571.14
	Standard GP + SMOTE	584.21
	Standard GP + Random oversampling	604.49
	Standard GP + Without resampling	591.34
Re_dataset_two_labels	GP-H + crossover-based oversampling	**204.03**
	GP-H + SMOTE	210.86
	GP-H + Random oversampling	232.95
	GP-H + Without resampling	235.35
	Standard GP + crossover-based oversampling	261.54
	Standard GP + SMOTE	225.58
	Standard GP + Random oversampling	229.32
	Standard GP + Without resampling	238.27

3. Conclusion

In machine learning, an imbalanced dataset's classification is a challenging task and an interesting research topic. In this study, the issue of the imbalanced dataset was addressed by proposing a complete framework combining GP and crossover-based oversampling for the classification task at hand. The crossover-based oversampling approach is easily implemented, does not result in information loss, and successfully reduces the bias of classifiers for the imbalanced datasets.

In order to illustrate the improved performance of the proposed system employing GP-H with cross-over based oversampling, it was compared to the GP model in conjunction with other oversampling techniques. Moreover, it was compared to the state-of-the-art. The experimental results proved that using the crossover-based oversampling technique in conjunction with the hybrid mutation of the GP improved the performance on imbalanced datasets.

The proposed system provides a complete solution that proved effective in the experiments presented; compared with the state-of-the-art, it provided the highest F1-scores in all five imbalanced datasets. Furthermore, in comparison to GP approaches that use different methods, experimental results demonstrated that the proposed approach greatly reduces running time and, more notably, improves classification performance.

In future work, we will determine an optimum ratio of the minority to the majority class to improve performance. Moreover, the effectiveness of the proposed approach on an imbalanced dataset for multiclassification tasks will be explored.

Author contributions

M.A. and N.D: Methodology, Conceptualization, Writing—original draft preparation, Writing—review and editing, validation, and visualization, M.A.: Software, Data curation, Supervision: N.D.

References

[1] Lee W, Seo K. Downsampling for Binary Classification with a Highly Imbalanced Dataset Using Active Learning. Big Data Research. 2022;28:100314. https://doi.org/10.1016/j.bdr.2022.100314

[2] Kumar P, Bhatnagar R, Gaur K, Bhatnagar A. Classification of imbalanced data: review of methods and applications. In: IOP Conference Series: Materials Science and Engineering; United Kingdom; 2021; 1099 (1): 012077.

[3] Hassan H, Ahmad NB, Anuar S. Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining. In: Journal of Physics: Conference Series; Xi'an, China; 2020; 1529 (5): 052041.

[4] Batuwita R, Palade V. Efficient resampling methods for training support vector machines with imbalanced datasets. In: IEEE 2010 International Joint Conference on Neural Networks (IJCNN);Barcelona, Spain; 2010; pp. 1-8.

[5] Bansal A, Jain A. Analysis of Focussed Under-Sampling Techniques with Machine Learning Classifiers. In: 2021 IEEE/ACIS 19th International Conference on Software Engineering Research, Management and Applications (SERA) 2021; Kanazawa, Japan; pp. 91-96.doi: 10.1109/SERA51205.2021.9509270

[6] Tao X, Li Q, Guo W, Ren C, He Q, Liu R, Zou J. Adaptive weighted over-sampling for imbalanced datasets based on density peaks clustering with heuristic filtering. Information Sciences 2020; 519:43-73. https://doi.org/10.1016/j.ins.2020.01.032

[7] Chawla N V, Bowyer K W, Hall L O, Kegelmeyer W P. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 2002; 16: 321-57. doi:10.1613/jair.953

[8] Liu J. Importance-SMOTE: a synthetic minority oversampling method for noisy imbalanced data. Soft Computing 2022; 26 (3): 1141-63. https://doi.org/10.1007/s00500-021-06532-4

[9] Chao X, Zhang L. Few-shot imbalanced classification based on data augmentation. Multimedia Systems 2021; 1-9. doi: https://doi.org/10.1007/s00530-021-00827-0

[10] Pradipta GA, Wardoyo R, Musdholifah A, Sanjaya IN. Radius-SMOTE: a new oversampling technique of minority samples based on radius distance for learning from imbalanced data. IEEE Access 2021; 9:74763-77. doi: 10.1109/ACCESS.2021.3080316

[11] Maulidevi NU, Surendro K. SMOTE-LOF for noise identification in imbalanced data classification. Journal of King Saud University-Computer and Information Sciences 2022; 34 (6): 3413-23. https://doi.org/10.1016/j.jksuci.2021.01.014

[12] Maldonado S, Vairetti C, Fernandez A, Herrera F. FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification. Pattern Recognition 2022; 124: 108511. https://doi.org/10.1016/j.patcog.2021.108511

[13] Li J, Zhu Q, Wu Q, Fan Z. A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. Information Sciences 2021; 565: 438-55. https://doi.org/10.1016/j.ins.2021.03.041

[14] Koza J R. Genetic programming: on the programming of computers by means of natural selection. MIT press, 1992.

[15] Poli R, Langdon W B, McPhee N F, Koza J R. A Field Guide to Genetic Programming. lulu. com. With contributions by JR Koza. 2008.

[16] Tripathi A, Chakraborty R, Kopparapu, S. K. A Novel Adaptive Minority Oversampling Technique for Improved Classification in Data Imbalanced Scenarios. In: 2020 25th International Conference on Pattern Recognition (ICPR); Milano, Italy; 2020; pp. 10650-10657.

[17] Azad C, Bhushan B, Sharma R, Shankar A, Singh KK, Khamparia A. Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus. Multimedia Systems 2022; 28 (4):1289-307. https://doi.org/10.1007/s00530-021-00817-2

[18] Kumar A, Sinha N, Bhardwaj A, Goel S. Clinical risk assessment of chronic kidney disease patients using genetic programming. Computer Methods in Biomechanics and Biomedical Engineering. 2022;25 (8): 887-95.

[19] Pei W, Xue B, Shang L, Zhang M. Reuse of program trees in genetic programming with a new fitness function in high-dimensional unbalanced classification. In: the Genetic and Evolutionary Computation Conference Companion; Prague Czech Republic; 2019; pp. 187-188.

[20] Pei W, Xue B, Shang L, Zhang M. New fitness functions in genetic programming for classification with high-dimensional unbalanced data. In: 2019 IEEE Congress on Evolutionary Computation; Wellington, New Zealand; 2019; pp. 2779-2786.

[21] Kumar A. A new fitness function in genetic programming for classification of imbalanced data. Journal of Experimental and Theoretical Artificial Intelligence 2022,1-3.doi: 10.1080/0952813X.2022.2120087

[22] Kumar A, Sinha N, Bhardwaj A. A novel fitness function in genetic programming for medical data classification. Journal of Biomedical Informatics 2020, 112:103623.https://doi.org/10.1016/j.jbi.2020.103623

[23] Mostafaei SH, Tanha J, Samadi N, Imanzadeh S, Razzaghi-Asl N. A boosting based approach to handle imbalanced data. In 2022 30th International Conference on Electrical Engineering;South Korea; 2022, pp. 295-299.

[24] Jiang Z, Pan T, Zhang C, Yang J. A new oversampling method based on the classification contribution degree. Symmetry. 2021;13 (2):194. https://doi.org/10.3390/sym13020194

[25] Pereira RM, Costa YM, Silla Jr CN. Toward hierarchical classification of imbalanced data using random resampling algorithms. Information Sciences. 2021;578:344-63.https://doi.org/10.1016/j.ins.2021.07.033

[26] Liu P, Liu X, Liu B, Chen X. A new over-sampling ensemble approach for imbalanced data. In: 2021 International Conference on Big Data Analysis and Computer Science; Nanjing, China; 2021, pp. 92-96.

[27] Aljero MKA, Dimililer N. Genetic Programming Approach to Detect Hate Speech in Social Media. IEEE Access 2021, 9: 115115-115125.doi:10.1109/ACCESS.2021.3104535

[28] Owen D, Collados J C, Espinosa-Anke L. Towards Preemptive Detection of Depression and Anxiety in Twitter. arXiv preprint arXiv:2011.05249

[29] Almeida T A, Hidalgo J M G, Yamakami A. Contributions to the study of SMS spam filtering: new collection and results. In: the 11th ACM symposium on Document engineering; California, USA; 2011, pp. 259-262.

[30] Ibrohim MO, Sazany E, Budi I. Identify abusive and offensive language in Indonesian twitter using deep learning approach. In Journal of Physics: Conference Series, IOP Publishing 2019; 1 (1196): 012041.

[31] Batuwita R, Palade V. Class imbalance learning methods for support vector machines. Imbalanced learning: Foundations, algorithms, and applications 2013, 83-99.doi:10.1002/9781118646106.ch5.

[32] Kumaresan K. Certain investigations on optimization techniques to enhance E mail spam classification, Anna University 2016. ch09, 73-91. doi: http://hdl.handle.net/10603/181292

[33] https://www.kaggle.com/stevenhans/depression-and-anxiety-in-twitter-id.

[34] Nagwani NK, Sharaff A. SMS spam filtering and thread identification using bi-level text classification and clustering techniques. Journal of Information Science 2017; 43 (1): 75-87. doi:10.1177/0165551512439173

[35] Mohd Pozi MS, Sulaiman MN, Mustapha N, Perumal T. A new classification model for a class imbalanced data set using genetic programming and support vector machines: case study for wilt disease classification. Remote Sensing Letters 2015; 6 (7): 568-577.doi:10.1080/2150704X.2015.1062159